

DISC: Dimension-Informed Clustering with Entropy Metric Warping and Attractor Basins

Qin Xu

Xgen AI

Chenggang Xu

Xgen AI

Hao Helen Zhang

Xgen AI

Tian Xu

`tianxu@xgen-ai.net`

Xgen AI <https://orcid.org/0009-0004-6859-9491>

Research Article

Keywords:

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11435776/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

DISC: Dimension-Informed Clustering with Entropy Metric Warping and Attractor Basins

Qin Xu
Xgen AI LLC, Delaware, USA
qinxu@xgen-ai.net

Chenggang Xu
Xgen AI LLC, Delaware, USA
chenggangxu@xgen-ai.net

Hao Helen Zhang
Xgen AI LLC, Delaware, USA
haozhang@xgen-ai.net

Tian Xu
(Corresponding author)
Xgen AI LLC, Delaware, USA
tianxu@xgen-ai.net

October 19, 2026

Abstract

Graph-based clustering builds a neighborhood graph and partitions it, treating every neighborhood alike regardless of its local geometry. We propose DISC (Dimension-Informed Spectral-entropy Clustering), which lets the local geometry decide which edges are trusted. For every point, the Shannon entropy of the normalized eigenvalues of its neighborhood covariance measures how many directions the data locally spread over. This entropy $H(\mathbf{x})$ defines a conformal metric warping, $g_{\mathbf{x}} = H(\mathbf{x})^{-2} \mathbf{I}$, under which directional neighbor affinities are computed. A reciprocity gate then keeps only mutual neighbor pairs whose weaker affinity exceeds the median. The resulting graph is clean (half the cross-class edge weight of a k -NN graph on MNIST) but fragmented. DISC therefore partitions it locally: every point follows its strongest gated edge toward a deeper neighbor, where depth is gated degree divided by entropy, and the resulting attractor basins, 72–99% class-pure on 16 of 17 datasets, are merged by average linkage. We evaluate DISC on 9 benchmark and 8 held-out datasets covering images, deep image features, text, speech, sensor and tabular data. Each subset is clustered with three algorithm seeds, and every method receives the true number of clusters or a label-free search for it. DISC outperforms K-Means, Ward linkage, spectral clustering, DBSCAN and HDBSCAN in adjusted Rand index and accuracy on 13–17 of 17 datasets (Wilcoxon $p \leq 10^{-10}$). These gains are as large on the eight held-out datasets, which played no role in its development, as on the benchmark datasets. DISC is statistically on par with Leiden community detection, and it runs in 3.5s on 70,000 MNIST points.

1 Introduction

The manifold hypothesis holds that high-dimensional data concentrate near lower-dimensional structures [8]. Graph-based clustering exploits this through a neighborhood graph: spectral clustering [18, 31], diffusion maps [5], Laplacian eigenmaps [1], UMAP [17] and community detection [27] all start from one. The graph determines what can be found. Edges between classes merge clusters, and missing edges within a class split them.

Most constructions treat all neighborhoods alike. Local scaling [35, 17] adapts edge weights to the local density, but not to the local shape of the data, although the intrinsic dimension of real data varies from place to place [15, 7]. Some neighborhoods are close to low-dimensional curves, while others spread over many directions.

DISC uses this local geometry at two levels. At the level of edges, the spectral entropy $H(\mathbf{x})$ of a point’s neighborhood covariance warps the metric. Distances out of low-entropy, nearly low-dimensional neighborhoods are lengthened, and a reciprocity gate keeps only edges that both endpoints strongly endorse.

At the level of clusters, the gated graph is too sparse for global partitioning, but it supports a local one. Following each point’s strongest edge to a deeper neighbor splits the data into small attractor basins, which are almost always class-pure. These are then merged into clusters.

Contributions.

- A local spectral entropy whose exponential is the effective number of directions of the local data (Section 3.1). The entropy defines a conformal metric warping and directional affinities (Section 3.2).
- A reciprocity gate, which produces a graph with half the cross-class edge weight of a k -NN graph (Section 4.3).
- An attractor-basin partition of the gated graph that uses neither eigenvectors nor modularity (Section 3.4).
- An evaluation on 17 datasets, 8 of which were never used during development, with four metrics, three algorithm seeds per subset, label-free cluster-count searches for methods that need them, paired significance tests, sensitivity and scalability to 70,000 points (Section 4).

Scope. DISC clusters a given representation of the data, and we compare it with methods that do the same. Deep clustering methods, such as DEC [34], SpectralNet [25] and SCAN [29], solve a different problem: they learn the representation and the clustering jointly. They train neural networks, often with data augmentation, and tune architectures per dataset. A comparison with them would mostly measure representation learning rather than the clustering rule, so we do not consider it a fair comparison. The two approaches are complementary: DISC can be applied to any learned embedding, as we do with ResNet-50 features for CIFAR-10.

2 Related Work

Neighborhood graphs and local scaling. Spectral clustering [18, 31] partitions a graph through the leading eigenvectors of its normalized affinity matrix. Self-tuning spectral clustering [35] scales each affinity by the distances of both endpoints to their k -th neighbors, and UMAP [17] and t-SNE [28] calibrate a per-point bandwidth. These constructions adapt to density but not to the local shape of the data. Leiden [27] optimizes a modularity-type objective on a graph and is a strong general-purpose partitioner of k -NN graphs.

Intrinsic dimension. Local intrinsic dimension is classically estimated by maximum likelihood on neighbor distances [15] or from ratios of neighbor distances [7]. Spectral estimates summarize the eigenvalues of a local covariance; the effective rank [23] is the exponential of their Shannon entropy [26]. DISC uses this entropy to warp the metric that defines the graph.

Graph sparsification. Mutual nearest neighbor graphs [9] keep an edge only if each endpoint is among the other’s neighbors. Shared-neighbor similarity [13] down-weights edges with few common neighbors. DISC’s reciprocity gate keeps mutual pairs whose weaker directional affinity, computed in the warped metric, exceeds the median.

Mode seeking. Mean shift [4] and quick shift [30] move points toward modes of a density. Density peaks [22] picks cluster centers as dense points far from denser ones. Persistence-based clustering [3] merges modes by topological persistence. DISC’s attractor basins follow the same hill-climbing principle on its gated graph, with a depth given by gated degree over entropy, and merge the basins by average linkage on the warped affinities.

Deep clustering. Methods that learn representations for clustering [34, 25, 29] are outside our scope, as explained in Section 1.

3 Method

Let $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\} \subset \mathbb{R}^D$ be the data and c the number of clusters to be found. Throughout, $\|\cdot\|$ is the Euclidean norm, $\mathbf{I}_D$ the $D \times D$ identity matrix, $\chi(\cdot)$ the indicator function (1 if its argument is true, 0 otherwise), and $|S|$ the number of elements of a set S . For each point, $\mathcal{N}(i) \subset \{1, \dots, N\} \setminus \{i\}$ denotes the indices of its k nearest neighbors in Euclidean distance. We use $k = 15$ throughout.

3.1 Local spectral entropy

For each point $\mathbf{x}_i$ with neighbors $\mathcal{N}(i) = \{j_1, \dots, j_k\}$, we form the displacement matrix $\mathbf{D}_i = [\mathbf{x}_{j_1} - \mathbf{x}_i, \dots, \mathbf{x}_{j_k} - \mathbf{x}_i]^\top \in \mathbb{R}^{k \times D}$ and the local covariance $\mathbf{C}_i = \frac{1}{k-1} \mathbf{D}_i^\top \mathbf{D}_i \in \mathbb{R}^{D \times D}$. Let $\lambda_{i,1} \geq \dots \geq \lambda_{i,D} \geq 0$ be the eigenvalues of $\mathbf{C}_i$, and let $q_{i,t} = \lambda_{i,t} / \sum_{u=1}^D \lambda_{i,u}$ be the share of local variance along the t -th principal direction, so that $(q_{i,1}, \dots, q_{i,D})$ is a probability vector. The local spectral entropy is

$$H_i = - \sum_{t=1}^D q_{i,t} \ln q_{i,t}, \quad (1)$$

with the convention $0 \ln 0 = 0$; we also write $H(\mathbf{x}_i) = H_i$. In the implementation, eigenvalues are floored at 10^{-12} , so that $H_i > 0$ and the warping below is always defined.

Proposition 3.1. *Let $r = \text{rank}(\mathbf{C}_i) \geq 1$. Then (i) $0 \leq H_i \leq \ln r \leq \ln \min(k, D)$; (ii) $H_i = \ln r$ if and only if the r non-zero eigenvalues are equal, and $H_i = 0$ if and only if $r = 1$; (iii) H_i is invariant to rotations, translations and uniform rescaling of the data.*

Thus $\exp(H_i)$, the effective rank of $\mathbf{C}_i$ [23], counts the directions that carry the local variance, weighted continuously by their share. It is close to 1 on a locally linear curve and close to d on an evenly filled d -dimensional patch. H_i is the order-1 member of the Rényi family $H_i^{(\alpha)} = \frac{1}{1-\alpha} \ln \sum_t q_{i,t}^\alpha$ for $\alpha \geq 0$, $\alpha \neq 1$, extended to $\alpha = 1$ by its limit [21], and $\exp(H_i^{(\alpha)})$ is the Hill number of order α [11]. Order 0 gives the rank and order 2 the participation ratio. Proofs are in Appendix A. The estimate is computed at the single scale set by k . It is related to, but distinct from, the information dimension of a measure, which is defined by a limit over scales [21].

3.2 Entropy metric warping and directional affinities

DISC equips each point $\mathbf{x}_i$ with the conformal metric $g_i = H_i^{-2} \mathbf{I}_D$, under which the length of a displacement from $\mathbf{x}_i$ to a neighbor is

$$d_{ij} = \sqrt{(\mathbf{x}_j - \mathbf{x}_i)^\top g_i (\mathbf{x}_j - \mathbf{x}_i)} = \frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{H_i}, \quad j \in \mathcal{N}(i). \quad (2)$$

Distances measured from low-entropy, nearly low-dimensional neighborhoods are thus lengthened, and distances from diffuse neighborhoods are shortened. The warped distance is directional: d_{ij} uses the metric at $\mathbf{x}_i$, so in general $d_{ij} \neq d_{ji}$. Directional affinities follow from a Gaussian kernel normalized over the neighbors of each point,

$$p_{j|i} = \begin{cases} \frac{\exp(-d_{ij}^2/2\sigma_i^2)}{\sum_{l \in \mathcal{N}(i)} \exp(-d_{il}^2/2\sigma_i^2)}, & j \in \mathcal{N}(i), \\ 0, & \text{otherwise,} \end{cases} \quad \sigma_i = \frac{\ln k}{2H_i} \max_{l \in \mathcal{N}(i)} d_{il}^2, \quad (3)$$

so that $\sum_j p_{j|i} = 1$ for every i . We collect them in the matrix $P \in \mathbb{R}^{N \times N}$ with $P_{ij} = p_{j|i}$. The bandwidth σ_i has units of squared distance, so the affinities depend on the overall scale of the data. We apply DISC to unwhitened principal components, so this scale is fixed by the data.

3.3 Reciprocity gate

An edge is kept only if both endpoints endorse it:

$$W_{ij} = p_{j|i} p_{i|j} \cdot \chi(\min(p_{j|i}, p_{i|j}) > \tau), \quad i \neq j, \quad (4)$$

where τ is the median of $\min(p_{j|i}, p_{i|j})$ over all mutual neighbor pairs $\{i, j\}$, that is, pairs with $j \in \mathcal{N}(i)$ and $i \in \mathcal{N}(j)$ (exactly the pairs for which both directional affinities are positive). $W \in \mathbb{R}^{N \times N}$ is symmetric with zero diagonal, it is non-zero only on mutual neighbor pairs, and by construction it removes half of them. Section 4.3 shows that the result is a graph with little cross-class weight but many connected pieces.

3.4 Attractor basins and basin merging

The gated graph W is sparse and splits into many connected pieces (Section 4.3). DISC therefore partitions it in two stages: a local stage, which groups points into small attractor basins by following the strongest gated edges, and a global stage, which merges the basins into c clusters.

Depth. The depth of point i is its gated degree divided by its local spectral entropy,

$$s_i = \frac{1}{H_i} \sum_{j=1}^N W_{ij}. \quad (5)$$

A point is deep when it has strong reciprocal edges and lies in a low-entropy neighborhood.

Attractor flow. Let

$$\Gamma(i) = \{j \in \mathcal{N}(i) : W_{ij} > 0 \text{ and } s_j > s_i\} \quad (6)$$

be the set of deeper gated neighbors of i . Each point is linked to its strongest deeper gated neighbor:

$$\pi(i) = \begin{cases} \arg \max_{j \in \Gamma(i)} W_{ij}, & \Gamma(i) \neq \emptyset, \\ i, & \Gamma(i) = \emptyset, \end{cases} \quad (7)$$

with ties broken by the smaller index. A point with $\pi(i) = i$ is an *attractor*. Write π^ℓ for the ℓ -fold composition of π , with $\pi^0(i) = i$. Because every step $i \rightarrow \pi(i) \neq i$ strictly increases the depth, $s_{\pi(i)} > s_i$, the sequence $i, \pi(i), \pi^2(i), \dots$ cannot revisit a point and therefore reaches an attractor

$$r(i) = \pi^{L_i}(i), \quad L_i = \min\{\ell \geq 0 : \pi^{\ell+1}(i) = \pi^\ell(i)\} \leq N - 1, \quad (8)$$

after finitely many steps. The edges $i \rightarrow \pi(i)$ thus form a forest, and the *attractor basins*

$$B_r = \{i : r(i) = r\}, \quad r \in \mathcal{R} = \{i : \pi(i) = i\}, \quad (9)$$

are its trees. They partition the data, and each basin contains exactly one attractor.

Isolated points. A point without any gated edge, $i \in \mathcal{I} = \{i : \sum_j W_{ij} = 0\}$, is its own attractor. Since $W_{i\pi(i)} > 0$ whenever $\pi(i) \neq i$, no other point flows into it, so it would form a singleton basin. Such points instead join the basin of their nearest connected point,

$$r(i) \leftarrow r\left(\arg \min_{j \notin \mathcal{I}} \|\mathbf{x}_i - \mathbf{x}_j\|\right), \quad i \in \mathcal{I}. \quad (10)$$

We write $B_1, \dots, B_m$ for the resulting m basins, indexed by $a, b \in \{1, \dots, m\}$.

Basin graph. Basins are compared through the symmetrized directional affinities of Section 3.2, $\bar{P} = \frac{1}{2}(P + P^\top)$, that is, $\bar{P}_{ij} = \frac{1}{2}(p_{j|i} + p_{i|j})$. $\bar{P}_{ij}$ is positive whenever $j \in \mathcal{N}(i)$ or $i \in \mathcal{N}(j)$, including the pairs removed by the reciprocity gate, so basins that the gate separated can still be connected. The affinity between two basins is

$$A_{ab} = \sum_{i \in B_a} \sum_{j \in B_b} \bar{P}_{ij}, \quad a \neq b. \quad (11)$$

Average-linkage merging. A cluster C is a set of point indices, and $\mathcal{C}$ denotes the current partition into clusters. Starting from $\mathcal{C} = \{B_1, \dots, B_m\}$, DISC repeatedly merges the two clusters with the largest average affinity,

$$S(C, C') = \frac{A(C, C')}{|C||C'|}, \quad A(C, C') = \sum_{i \in C} \sum_{j \in C'} \bar{P}_{ij}, \quad (C^*, C'^*) = \arg \max_{\substack{C \neq C' \in \mathcal{C} \\ S(C, C') > 0}} S(C, C'), \quad (12)$$

and replaces them by $C^* \cup C'^*$, until $|\mathcal{C}| = c$. Because A is additive over the points of a cluster, and $A(B_a, B_b) = A_{ab}$ for the initial basins, the affinity of a merged cluster to any other cluster C'' is

$$A(C^* \cup C'^*, C'') = A(C^*, C'') + A(C'^*, C''), \quad |C^* \cup C'^*| = |C^*| + |C'^*|, \quad (13)$$

so S is updated in constant time per affected pair. $S(C, C')$ is always the mean affinity over all point pairs of the two clusters, the average-linkage criterion applied to basins rather than points. If no remaining pair has $S(C, C') > 0$, the clusters left are disconnected in $\bar{P}$, and the pair with the closest centroids $\mu_C = |C|^{-1} \sum_{i \in C} \mathbf{x}_i$ is merged instead. When $|\mathcal{C}| = c$, the clustering is $\mathcal{C}$: point i is assigned to the cluster that contains its basin. Since $m \ll N$ (a few hundred basins for 5,000 points in our experiments), the merging operates on a much smaller graph than the original data, and it is implemented with a priority queue over the non-zero entries of A .

3.5 Complexity

k -NN search costs $O(N \log N D)$ with a space-partitioning tree. The local covariances and their eigendecompositions cost $O(N(kD^2 + D^3))$. Affinities, the gate and the depth cost $O(Nk)$, and the attractor flow costs $O(Nk + N \log L)$ with pointer jumping, where $L = \max_i L_i$. Merging costs $O(E \log E)$ for the $E \leq Nk$ non-zero entries of the basin affinity A , plus the rare centroid merges. Memory is $O(N(D + k))$. Section 4.4 reports measured runtimes.

4 Experiments

4.1 Protocol

Datasets. We use nine benchmark datasets and eight held-out datasets. The held-out datasets were never used to develop or tune DISC. The benchmark datasets are MNIST and Fashion-MNIST (28×28 images), USPS (16×16 digits), CIFAR-10 (OpenML `CIFAR_10_small`, mapped to 2,048-dimensional features of an ImageNet-pretrained ResNet-50 [10]), Olivetti faces (400 images of 40 people), 20 Newsgroups (TF-IDF and latent semantic analysis), the tabular datasets Letter and Pendigits, and an imbalanced MNIST with class sizes from 1,000 to 50. The held-out datasets are Optdigits and Semeion (digit images), Isolet (speech), HAR (smartphone sensors), Satimage (multispectral), Segment (image regions), Mfeat-factors (digit features) and Gas-drift (gas sensors). For every subset seed we draw a class-stratified subset of up to 5,000 points without replacement. Tabular and sensor features are standardized, and all data are projected onto 20 principal components, or kept as they are if they have fewer features. All methods see the same data.

Baselines and cluster counts. We compare against K-Means [16], Ward linkage [32] with a 30-NN connectivity constraint, spectral clustering [18] on a k -NN affinity, DBSCAN [6] with `min_samples` = $2D$, where $D = 20$ is the input dimension [24], HDBSCAN [2], and Leiden [27] on the symmetrized 15-NN graph. K-Means, Ward linkage, spectral clustering and DISC receive the true number of clusters c . For DBSCAN, HDBSCAN and Leiden, we search without labels for the setting whose number of clusters is closest to c , breaking ties by the smaller noise fraction; the settings searched are `ε`, `min_cluster_size` and the resolution, respectively. No method is tuned with labels.

Seeds and metrics. Each dataset is subsampled with five seeds (600–604) that were not used during development. Each subset is clustered with three algorithm seeds and the scores are averaged. We report three external metrics, which compare a clustering with the true classes: the adjusted Rand index (ARI) [12],

Table 1: Results on the nine benchmark datasets: mean $\pm$ standard deviation over 5 subsets, each averaged over 3 algorithm seeds. Best value per dataset and metric in bold; the mean rank is taken over datasets (undefined silhouettes rank last).

Dataset	K-Means	Ward	Spectral	DBSCAN	HDBSCAN	Leiden	DISC
<i>ARI $\uparrow$</i>							
MNIST	0.355 $\pm$ 0.010	0.451 $\pm$ 0.028	0.537 $\pm$ 0.032	0.059 $\pm$ 0.021	0.057 $\pm$ 0.006	0.677$\pm$0.015	0.667 $\pm$ 0.033
F-MNIST	0.360 $\pm$ 0.010	0.359 $\pm$ 0.017	0.423 $\pm$ 0.036	0.108 $\pm$ 0.008	0.032 $\pm$ 0.029	0.456 $\pm$ 0.011	0.463$\pm$0.033
USPS	0.529 $\pm$ 0.006	0.621 $\pm$ 0.017	0.568 $\pm$ 0.003	0.137 $\pm$ 0.006	0.087 $\pm$ 0.005	0.704 $\pm$ 0.053	0.721$\pm$0.059
CIFAR-10	0.446 $\pm$ 0.015	0.388 $\pm$ 0.011	0.418 $\pm$ 0.054	0.128 $\pm$ 0.036	0.042 $\pm$ 0.023	0.543$\pm$0.013	0.397 $\pm$ 0.027
Olivetti	0.444 $\pm$ 0.008	0.451 $\pm$ 0.000	0.472 $\pm$ 0.008	0.000 $\pm$ 0.000	0.055 $\pm$ 0.001	0.409 $\pm$ 0.007	0.546$\pm$0.002
20NG	0.208 $\pm$ 0.007	0.188 $\pm$ 0.008	0.103 $\pm$ 0.016	0.027 $\pm$ 0.008	0.011 $\pm$ 0.004	0.222$\pm$0.010	0.183 $\pm$ 0.009
Letter	0.157 $\pm$ 0.004	0.170 $\pm$ 0.013	0.049 $\pm$ 0.008	0.025 $\pm$ 0.007	0.014 $\pm$ 0.006	0.215 $\pm$ 0.007	0.229$\pm$0.007
Pendigits	0.549 $\pm$ 0.008	0.598 $\pm$ 0.024	0.618 $\pm$ 0.056	0.497 $\pm$ 0.032	0.324 $\pm$ 0.015	0.758$\pm$0.027	0.710 $\pm$ 0.021
MNIST-imb	0.339 $\pm$ 0.027	0.376 $\pm$ 0.032	0.511 $\pm$ 0.060	0.137 $\pm$ 0.038	0.129 $\pm$ 0.065	0.558 $\pm$ 0.062	0.573$\pm$0.076
<i>Mean rank</i>	<i>4.00</i>	<i>3.78</i>	<i>3.44</i>	<i>6.11</i>	<i>6.89</i>	<i>1.89</i>	<i>1.89</i>
<i>ACC $\uparrow$</i>							
MNIST	0.534 $\pm$ 0.009	0.589 $\pm$ 0.032	0.636 $\pm$ 0.013	0.139 $\pm$ 0.016	0.180 $\pm$ 0.020	0.749$\pm$0.014	0.749$\pm$0.025
F-MNIST	0.506 $\pm$ 0.015	0.495 $\pm$ 0.026	0.572 $\pm$ 0.047	0.205 $\pm$ 0.009	0.132 $\pm$ 0.040	0.571 $\pm$ 0.007	0.596$\pm$0.043
USPS	0.666 $\pm$ 0.005	0.698 $\pm$ 0.023	0.587 $\pm$ 0.003	0.223 $\pm$ 0.009	0.269 $\pm$ 0.012	0.728 $\pm$ 0.068	0.758$\pm$0.067
CIFAR-10	0.623 $\pm$ 0.038	0.550 $\pm$ 0.014	0.631 $\pm$ 0.030	0.234 $\pm$ 0.027	0.168 $\pm$ 0.014	0.746$\pm$0.010	0.587 $\pm$ 0.056
Olivetti	0.585 $\pm$ 0.015	0.588 $\pm$ 0.000	0.615 $\pm$ 0.006	0.025 $\pm$ 0.000	0.376 $\pm$ 0.002	0.510 $\pm$ 0.007	0.642$\pm$0.000
20NG	0.373 $\pm$ 0.010	0.352 $\pm$ 0.015	0.317 $\pm$ 0.009	0.139 $\pm$ 0.018	0.112 $\pm$ 0.026	0.385$\pm$0.005	0.342 $\pm$ 0.009
Letter	0.296 $\pm$ 0.009	0.296 $\pm$ 0.009	0.267 $\pm$ 0.008	0.119 $\pm$ 0.018	0.110 $\pm$ 0.044	0.382 $\pm$ 0.006	0.394$\pm$0.012
Pendigits	0.683 $\pm$ 0.011	0.708 $\pm$ 0.017	0.753 $\pm$ 0.033	0.603 $\pm$ 0.027	0.522 $\pm$ 0.007	0.863$\pm$0.023	0.815 $\pm$ 0.017
MNIST-imb	0.437 $\pm$ 0.021	0.488 $\pm$ 0.026	0.590 $\pm$ 0.049	0.222 $\pm$ 0.049	0.246 $\pm$ 0.027	0.624 $\pm$ 0.051	0.646$\pm$0.072
<i>Mean rank</i>	<i>3.89</i>	<i>3.89</i>	<i>3.33</i>	<i>6.44</i>	<i>6.56</i>	<i>2.00</i>	<i>1.89</i>
<i>VI $\downarrow$</i>							
MNIST	2.363 $\pm$ 0.051	1.733 $\pm$ 0.096	1.376 $\pm$ 0.075	2.463 $\pm$ 0.147	2.109 $\pm$ 0.016	1.032$\pm$0.045	1.088 $\pm$ 0.087
F-MNIST	2.173 $\pm$ 0.033	2.039 $\pm$ 0.086	1.794 $\pm$ 0.088	2.245 $\pm$ 0.082	2.599 $\pm$ 0.191	1.717 $\pm$ 0.032	1.703$\pm$0.061
USPS	1.775 $\pm$ 0.016	1.195 $\pm$ 0.061	1.015 $\pm$ 0.015	2.186 $\pm$ 0.023	1.976 $\pm$ 0.012	0.854 $\pm$ 0.050	0.837$\pm$0.077
CIFAR-10	2.013 $\pm$ 0.071	2.155 $\pm$ 0.053	1.778 $\pm$ 0.050	2.213 $\pm$ 0.033	2.326 $\pm$ 0.016	1.721$\pm$0.030	2.203 $\pm$ 0.077
Olivetti	1.641 $\pm$ 0.039	1.517 $\pm$ 0.000	1.539 $\pm$ 0.019	3.689 $\pm$ 0.000	2.438 $\pm$ 0.010	1.873 $\pm$ 0.024	1.168$\pm$0.001
20NG	3.811 $\pm$ 0.037	3.895 $\pm$ 0.066	3.545 $\pm$ 0.097	3.115$\pm$0.180	3.203 $\pm$ 0.136	3.681 $\pm$ 0.071	3.995 $\pm$ 0.057
Letter	3.943 $\pm$ 0.014	3.688 $\pm$ 0.049	2.951 $\pm$ 0.097	3.447 $\pm$ 0.023	3.522 $\pm$ 0.162	3.057 $\pm$ 0.028	2.924$\pm$0.041
Pendigits	1.384 $\pm$ 0.017	1.098 $\pm$ 0.081	0.870 $\pm$ 0.073	1.258 $\pm$ 0.036	1.620 $\pm$ 0.029	0.724$\pm$0.034	0.800 $\pm$ 0.067
MNIST-imb	2.142 $\pm$ 0.086	1.785 $\pm$ 0.098	1.321 $\pm$ 0.163	2.128 $\pm$ 0.113	1.772 $\pm$ 0.111	1.255$\pm$0.187	1.255$\pm$0.162
<i>Mean rank</i>	<i>5.33</i>	<i>4.33</i>	<i>2.78</i>	<i>5.44</i>	<i>5.44</i>	<i>2.22</i>	<i>2.44</i>
<i>Silhouette $\uparrow$</i>							
MNIST	0.122 $\pm$ 0.002	0.085 $\pm$ 0.006	0.085 $\pm$ 0.002	0.133 $\pm$ 0.096	0.227$\pm$0.024	0.094 $\pm$ 0.003	0.079 $\pm$ 0.003
F-MNIST	0.240 $\pm$ 0.005	0.207 $\pm$ 0.013	0.224 $\pm$ 0.013	0.256$\pm$0.034	0.066 $\pm$ 0.126	0.215 $\pm$ 0.006	0.171 $\pm$ 0.022
USPS	0.191 $\pm$ 0.002	0.163 $\pm$ 0.007	0.104 $\pm$ 0.002	0.154 $\pm$ 0.023	0.391$\pm$0.014	0.131 $\pm$ 0.007	0.123 $\pm$ 0.027
CIFAR-10	0.154 $\pm$ 0.009	0.118 $\pm$ 0.004	0.105 $\pm$ 0.013	0.274 $\pm$ 0.039	0.336$\pm$0.047	0.123 $\pm$ 0.002	0.095 $\pm$ 0.014
Olivetti	0.239 $\pm$ 0.002	0.253 $\pm$ 0.000	0.236 $\pm$ 0.002	–	0.393$\pm$0.004	0.130 $\pm$ 0.006	0.208 $\pm$ 0.002
20NG	0.175 $\pm$ 0.005	0.128 $\pm$ 0.005	0.125 $\pm$ 0.009	0.308 $\pm$ 0.118	0.418$\pm$0.122	0.156 $\pm$ 0.003	0.092 $\pm$ 0.008
Letter	0.144$\pm$0.002	0.112 $\pm$ 0.004	0.015 $\pm$ 0.013	0.072 $\pm$ 0.064	0.140 $\pm$ 0.064	0.070 $\pm$ 0.010	0.055 $\pm$ 0.020
Pendigits	0.283 $\pm$ 0.005	0.271 $\pm$ 0.006	0.222 $\pm$ 0.006	0.347 $\pm$ 0.036	0.401$\pm$0.006	0.235 $\pm$ 0.016	0.222 $\pm$ 0.007
MNIST-imb	0.135 $\pm$ 0.011	0.097 $\pm$ 0.011	0.093 $\pm$ 0.007	0.119 $\pm$ 0.065	0.197$\pm$0.038	0.111 $\pm$ 0.012	0.094 $\pm$ 0.010
<i>Mean rank</i>	<i>2.44</i>	<i>4.11</i>	<i>5.78</i>	<i>3.00</i>	<i>1.78</i>	<i>4.56</i>	<i>6.33</i>

clustering accuracy (ACC) under the optimal one-to-one matching of clusters to classes [14], and the variation of information (VI) [19]. We also report the silhouette coefficient [20], an internal metric computed in the 20-dimensional input space without labels. Points labeled as noise count as errors for ACC, form one extra group for ARI and VI, and are excluded from the silhouette. All metrics are given as mean $\pm$ standard deviation over subsets, and paired comparisons use the Wilcoxon signed-rank test [33] over subsets. Definitions are in Appendix B.

Table 2: Results on the eight held-out datasets, which were never used during development (same protocol as Table 1).

Dataset	K-Means	Ward	Spectral	DBSCAN	HDBSCAN	Leiden	DISC
<i>ARI</i> $\uparrow$							
Optdigits	0.671 $\pm$ 0.005	0.763 $\pm$ 0.022	0.847 $\pm$ 0.011	0.493 $\pm$ 0.041	0.274 $\pm$ 0.023	0.909$\pm$0.015	0.908 $\pm$ 0.025
Semeion	0.398 $\pm$ 0.004	0.490 $\pm$ 0.008	0.398 $\pm$ 0.004	0.221 $\pm$ 0.015	0.089 $\pm$ 0.011	0.587 $\pm$ 0.012	0.653$\pm$0.047
Isolet	0.446 $\pm$ 0.010	0.456 $\pm$ 0.014	0.423 $\pm$ 0.013	0.084 $\pm$ 0.008	0.013 $\pm$ 0.002	0.520$\pm$0.016	0.484 $\pm$ 0.017
HAR	0.399 $\pm$ 0.026	0.439 $\pm$ 0.046	0.462 $\pm$ 0.011	0.282 $\pm$ 0.024	0.315 $\pm$ 0.011	0.517 $\pm$ 0.018	0.523$\pm$0.042
Satimage	0.530 $\pm$ 0.003	0.505 $\pm$ 0.029	0.479 $\pm$ 0.007	0.270 $\pm$ 0.018	0.100 $\pm$ 0.002	0.659$\pm$0.017	0.531 $\pm$ 0.062
Segment	0.463$\pm$0.003	0.437 $\pm$ 0.000	0.306 $\pm$ 0.000	0.277 $\pm$ 0.000	0.240 $\pm$ 0.000	0.380 $\pm$ 0.009	0.400 $\pm$ 0.000
Mfeat-fac	0.612 $\pm$ 0.005	0.497 $\pm$ 0.005	0.846$\pm$0.000	0.193 $\pm$ 0.001	0.303 $\pm$ 0.070	0.841 $\pm$ 0.003	0.802 $\pm$ 0.000
Gas-drift	0.082 $\pm$ 0.006	0.103 $\pm$ 0.031	0.101 $\pm$ 0.038	0.050 $\pm$ 0.034	0.071 $\pm$ 0.032	0.269 $\pm$ 0.063	0.297$\pm$0.065
<i>Mean rank</i>	<i>3.88</i>	<i>3.50</i>	<i>3.88</i>	<i>6.38</i>	<i>6.62</i>	<i>1.88</i>	<i>1.88</i>
<i>ACC</i> $\uparrow$							
Optdigits	0.793 $\pm$ 0.003	0.838 $\pm$ 0.015	0.912 $\pm$ 0.007	0.609 $\pm$ 0.042	0.487 $\pm$ 0.026	0.953 $\pm$ 0.011	0.955$\pm$0.014
Semeion	0.585 $\pm$ 0.010	0.670 $\pm$ 0.021	0.542 $\pm$ 0.003	0.391 $\pm$ 0.012	0.334 $\pm$ 0.014	0.769 $\pm$ 0.011	0.803$\pm$0.033
Isolet	0.541 $\pm$ 0.011	0.546 $\pm$ 0.019	0.513 $\pm$ 0.012	0.177 $\pm$ 0.011	0.080 $\pm$ 0.015	0.573$\pm$0.020	0.545 $\pm$ 0.018
HAR	0.534 $\pm$ 0.022	0.557 $\pm$ 0.015	0.589 $\pm$ 0.003	0.233 $\pm$ 0.047	0.318 $\pm$ 0.025	0.633$\pm$0.025	0.628 $\pm$ 0.061
Satimage	0.675 $\pm$ 0.003	0.643 $\pm$ 0.034	0.606 $\pm$ 0.010	0.365 $\pm$ 0.020	0.299 $\pm$ 0.009	0.793$\pm$0.019	0.671 $\pm$ 0.060
Segment	0.551$\pm$0.007	0.523 $\pm$ 0.000	0.452 $\pm$ 0.000	0.449 $\pm$ 0.000	0.405 $\pm$ 0.000	0.533 $\pm$ 0.008	0.539 $\pm$ 0.000
Mfeat-fac	0.722 $\pm$ 0.004	0.646 $\pm$ 0.006	0.926$\pm$0.000	0.407 $\pm$ 0.001	0.376 $\pm$ 0.016	0.922 $\pm$ 0.002	0.905 $\pm$ 0.000
Gas-drift	0.373 $\pm$ 0.009	0.353 $\pm$ 0.022	0.374 $\pm$ 0.028	0.186 $\pm$ 0.008	0.190 $\pm$ 0.032	0.506 $\pm$ 0.062	0.527$\pm$0.061
<i>Mean rank</i>	<i>3.62</i>	<i>3.88</i>	<i>3.75</i>	<i>6.25</i>	<i>6.75</i>	<i>1.75</i>	<i>2.00</i>
<i>VI</i> $\downarrow$							
Optdigits	1.120 $\pm$ 0.020	0.720 $\pm$ 0.074	0.483 $\pm$ 0.037	1.273 $\pm$ 0.071	1.626 $\pm$ 0.058	0.349$\pm$0.036	0.373 $\pm$ 0.070
Semeion	2.085 $\pm$ 0.007	1.584 $\pm$ 0.039	1.681 $\pm$ 0.013	2.052 $\pm$ 0.016	2.126 $\pm$ 0.007	1.325 $\pm$ 0.015	1.177$\pm$0.105
Isolet	1.975 $\pm$ 0.042	1.866 $\pm$ 0.020	1.714 $\pm$ 0.070	2.802 $\pm$ 0.068	3.296 $\pm$ 0.031	1.664$\pm$0.049	1.856 $\pm$ 0.043
HAR	1.570 $\pm$ 0.064	1.421 $\pm$ 0.116	1.348 $\pm$ 0.024	1.506 $\pm$ 0.027	1.453 $\pm$ 0.074	1.209 $\pm$ 0.053	1.172$\pm$0.080
Satimage	1.335 $\pm$ 0.011	1.411 $\pm$ 0.052	1.304 $\pm$ 0.009	1.676 $\pm$ 0.035	1.713 $\pm$ 0.046	1.060$\pm$0.014	1.362 $\pm$ 0.144
Segment	1.472 $\pm$ 0.008	1.457$\pm$0.000	1.662 $\pm$ 0.000	1.748 $\pm$ 0.000	1.942 $\pm$ 0.000	1.540 $\pm$ 0.034	1.576 $\pm$ 0.000
Mfeat-fac	1.302 $\pm$ 0.015	1.515 $\pm$ 0.006	0.631$\pm$0.000	1.818 $\pm$ 0.004	1.760 $\pm$ 0.147	0.634 $\pm$ 0.008	0.743 $\pm$ 0.000
Gas-drift	2.431 $\pm$ 0.018	2.514 $\pm$ 0.113	1.806 $\pm$ 0.146	2.249 $\pm$ 0.020	2.217 $\pm$ 0.081	1.495$\pm$0.074	1.541 $\pm$ 0.115
<i>Mean rank</i>	<i>4.75</i>	<i>4.12</i>	<i>2.88</i>	<i>5.88</i>	<i>6.25</i>	<i>1.62</i>	<i>2.50</i>
<i>Silhouette</i> $\uparrow$							
Optdigits	0.219 $\pm$ 0.001	0.204 $\pm$ 0.003	0.201 $\pm$ 0.001	0.241 $\pm$ 0.016	0.286$\pm$0.013	0.194 $\pm$ 0.002	0.193 $\pm$ 0.003
Semeion	0.126 $\pm$ 0.001	0.109 $\pm$ 0.004	0.103 $\pm$ 0.000	0.209 $\pm$ 0.005	0.243$\pm$0.014	0.121 $\pm$ 0.002	0.099 $\pm$ 0.002
Isolet	0.148 $\pm$ 0.002	0.127 $\pm$ 0.003	0.139 $\pm$ 0.004	0.154 $\pm$ 0.041	0.247$\pm$0.013	0.135 $\pm$ 0.002	0.114 $\pm$ 0.004
HAR	0.210 $\pm$ 0.003	0.199 $\pm$ 0.019	0.207 $\pm$ 0.007	0.465$\pm$0.038	0.446 $\pm$ 0.067	0.168 $\pm$ 0.021	0.132 $\pm$ 0.015
Satimage	0.360$\pm$0.002	0.318 $\pm$ 0.009	0.323 $\pm$ 0.005	0.132 $\pm$ 0.010	0.075 $\pm$ 0.124	0.279 $\pm$ 0.004	0.263 $\pm$ 0.027
Segment	0.349 $\pm$ 0.009	0.335 $\pm$ 0.000	0.239 $\pm$ 0.000	0.371$\pm$0.000	0.315 $\pm$ 0.001	0.216 $\pm$ 0.016	0.193 $\pm$ 0.000
Mfeat-fac	0.223 $\pm$ 0.002	0.194 $\pm$ 0.001	0.197 $\pm$ 0.000	0.344$\pm$0.001	0.252 $\pm$ 0.018	0.197 $\pm$ 0.001	0.186 $\pm$ 0.000
Gas-drift	0.325 $\pm$ 0.007	0.214 $\pm$ 0.015	-0.204 $\pm$ 0.064	0.296 $\pm$ 0.105	0.511$\pm$0.105	-0.069 $\pm$ 0.101	-0.067 $\pm$ 0.116
<i>Mean rank</i>	<i>2.50</i>	<i>4.50</i>	<i>4.62</i>	<i>2.25</i>	<i>2.38</i>	<i>5.25</i>	<i>6.50</i>

4.2 Results

Tables 1 and 2 report all four metrics. Against the classical baselines, DISC is consistently more accurate on the external metrics. Over all 17 datasets it has a higher ARI than K-Means on 14 datasets, than Ward linkage and spectral clustering on 15, and than DBSCAN and HDBSCAN on all 17 (paired Wilcoxon tests over the 85 subsets, $p \leq 1.1 \times 10^{-11}$). The mean ARI gains are +0.12, +0.11, +0.09, +0.36 and +0.41, respectively. Accuracy gives the same picture: DISC is more accurate than K-Means on 13 datasets, than Ward linkage and spectral clustering on 15, and than DBSCAN and HDBSCAN on all 17 ($p \leq 5 \times 10^{-11}$), with mean gains between +0.07 and +0.38. These gains generalize to data that played no role in the development of DISC. On the held-out datasets, DISC has a higher ARI than K-Means, Ward linkage and spectral clustering on seven of eight datasets, and its mean ARI gains over these three methods are as large as on the benchmark datasets: between +0.09 and +0.13 on the held-out datasets, against +0.09 to +0.12 on the benchmark datasets. The method was therefore not overfit to the datasets used in its development.

Against Leiden, DISC is statistically on par. It has a higher ARI on 9 of the 17 datasets (mean difference -0.008 , $p = 0.2$) and a higher accuracy on 9 (-0.008 , $p = 0.31$), and the two methods share the best mean ARI rank on both groups of datasets. Their strengths lie in different places, however. DISC is clearly better on Olivetti, whose 40 small classes suit its basins (+0.14 ARI), and on Semeion (+0.07), and it is ahead on Fashion-MNIST, USPS, Letter, imbalanced MNIST, HAR, Segment and Gas-drift. Leiden is clearly better on the deep CIFAR-10 features (-0.15) and on Satimage (-0.13), and slightly ahead on MNIST, 20 Newsgroups, Pendigits, Optdigits, Isolet and Mfeat-factors. VI follows the same pattern, with mean ranks of 2.44 for DISC and 2.22 for Leiden on the benchmark datasets.

The silhouette coefficient ranks the methods very differently. HDBSCAN and DBSCAN obtain the highest silhouettes on most datasets, although their external scores are the lowest, and both DISC and Leiden obtain low silhouettes. The silhouette rewards compact, convex groups in the input space, and density-based methods can raise it by declaring the hardest points noise, which the silhouette ignores. Clusters that follow curved or elongated class structure, as those of DISC do, receive lower silhouettes even when they agree better with the classes. We therefore read the silhouette as a description of cluster shape rather than of clustering quality.

4.3 Analysis: clean edges and pure basins

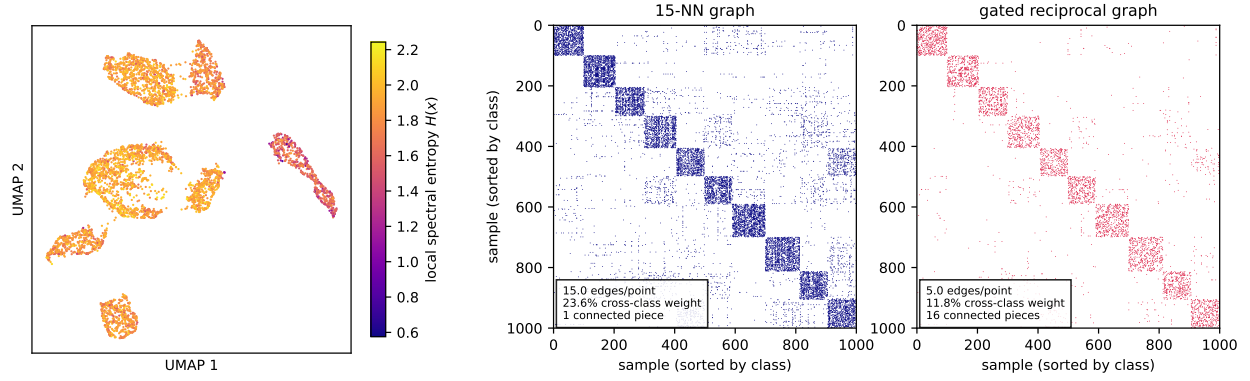
Figure 1 examines the three stages of DISC. The local spectral entropy varies mainly between classes rather than within them (Figure 1a). On MNIST, the digit 1, whose images lie close to a one-dimensional family of strokes, has a median entropy of 1.61, against 1.79–1.92 for the other digits. The warping therefore lengthens distances out of the low-entropy digits and shortens them in diffuse regions.

The warped affinities and the reciprocity gate change the graph substantially (Figure 1b). On 1,000 class-sorted MNIST points, 23.6% of the edge weight of the 15-NN graph connects different classes. After the gate, only 11.8% does. The price is connectivity: with 5.0 edges per point on average, the gated graph falls apart into 16 pieces. A global partition of this graph, such as spectral clustering, would spend its eigenvectors on the small pieces. This is why DISC partitions the graph locally.

The local partition is reliable. Following the strongest gated edges to deeper neighbors produces a few hundred attractor basins per dataset, and these basins are almost always class-pure (Figure 1c). Their purity is 0.95 on USPS, 0.98 on Optdigits and Gas-drift, and 0.99 on Pendigits, and it is at least 0.72 on every dataset except 20 Newsgroups. The final clusters are less pure than the basins, most clearly on Letter (0.82 for the basins against 0.42 for the clusters), Segment (0.93 against 0.57) and Gas-drift (0.98 against 0.60). Most of the remaining error therefore arises when basins are merged, not when they are formed, and better merging criteria are the most promising route to further improvement.

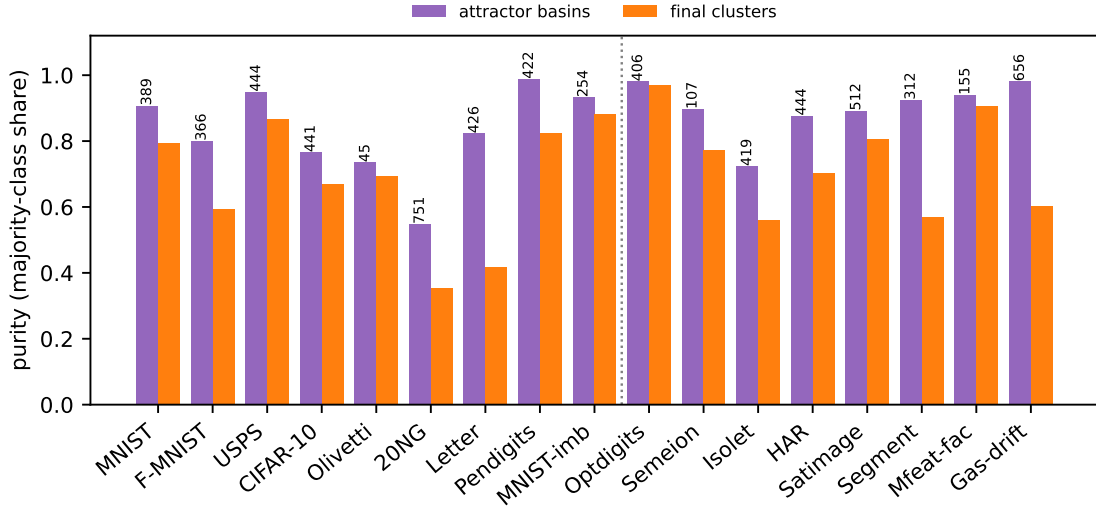
4.4 Sensitivity and scalability

The neighborhood size k is DISC’s only free parameter. Between $k = 10$ and $k = 50$, its ARI varies by at most 0.03 on MNIST, 20 Newsgroups and Letter, by 0.06 on Semeion and by 0.08 on USPS (Figure 2a). Very small neighborhoods are unreliable: with $k = 5$, DISC fails on Letter (ARI 0.03). Satimage shows a single dip, to 0.44 at $k = 30$.



(a) Local spectral entropy

(b) k -NN graph vs. gated reciprocal graph



(c) Purity of attractor basins and final clusters (numbers: basins per dataset)

Figure 1: Analysis of DISC (subset seed 600). (a) $H(x)$ on 5,000 MNIST points (UMAP layout); its median is 1.61 for the digit 1 and 1.79–1.92 for the other digits. (b) 1,000 class-sorted MNIST points: the reciprocity gate keeps 5.0 edges per point and halves the cross-class weight (23.6% $\rightarrow$ 11.8%), but splits the graph into 16 pieces. (c) The attractor basins are 72–99% pure on 16 of 17 datasets (20 Newsgroups: 55%). Dotted line: held-out datasets to the right.

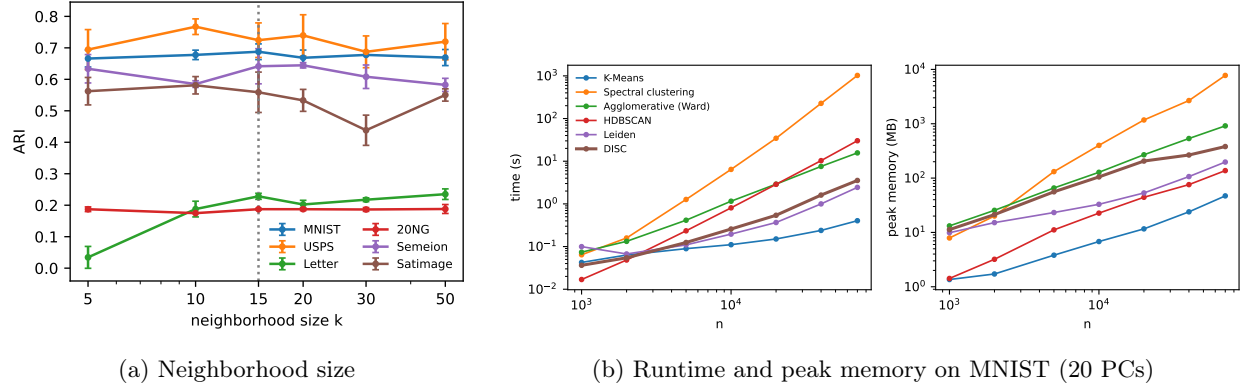


Figure 2: (a) ARI of DISC against k (mean $\pm$ std over subsets 600–602); dotted line: default $k = 15$. (b) Single fits for $N = 1,000$ –70,000, each in a fresh process.

DISC scales to large datasets (Figure 2b). On 70,000 MNIST points it takes 3.5s and 381 MB, compared with 2.4s for Leiden, 16s for Ward linkage, 30s for HDBSCAN and 1,024s and 7.7 GB for the spectral clustering of scikit-learn. Between 20,000 and 70,000 points, the measured scaling exponent, the slope $\Delta \ln t / \Delta \ln N$ of the logarithm of the runtime t against the logarithm of the number of points N , is 1.50 for DISC and 1.51 for Leiden, against 2.71 for spectral clustering. Most of DISC’s runtime is spent on the k -NN search.

4.5 Development protocol

All design decisions were made on development subsets drawn with seeds 100–102 from MNIST, USPS, Letter and 20 Newsgroups. Each development subset was clustered with three algorithm seeds, because algorithm-seed noise alone reaches 0.12 ARI on USPS. In total, 29 configurations were compared on these subsets. The reported numbers come from a single evaluation on subsets 600–604, which had not been used before. The first four held-out datasets (Optdigits, Semeion, Isolet and HAR) had been used in earlier evaluations of intermediate configurations, and one design choice, the graph on which basins are merged, was finalized after the held-out evaluation. Code, data loaders, all results and the full development record are available in the accompanying repository.

5 Discussion and Limitations

DISC replaces a uniform neighborhood graph with one shaped by local geometry. The entropy warping and the reciprocity gate keep edges that both endpoints strongly endorse, and attractor basins turn the resulting sparse, clean graph into clusters without eigenvectors or modularity. On fixed representations, DISC is far more accurate than the classical baselines and on par with Leiden, with clear gains on small-class data such as Olivetti and on several held-out datasets. Its advantage over the classical baselines is as large on datasets never used during its development as on the benchmark datasets, so it generalizes beyond the data on which it was designed.

Several limitations remain. Although the basins are pure, average linkage merges some of them wrongly on Letter, Segment and Gas-drift, and DISC trails Leiden on deep CIFAR-10 features and on Satimage. The bandwidth σ_i makes the affinities depend on the scale of the data, and a scale-free bandwidth may improve robustness. The neighborhood size is fixed, and very small values fail on Letter. Finally, we evaluate on fixed representations with up to 5,000 points per subset, complemented by a 70,000-point timing study; combining DISC with learned representations is a natural next step.

A Proofs

Proof of Proposition 3.1. $\mathbf{C}_i$ is symmetric positive semi-definite with $r \geq 1$ positive eigenvalues, so the non-zero entries of $(q_{i,1}, \dots, q_{i,D})$ form a probability vector on r outcomes (zero eigenvalues contribute $0 \ln 0 = 0$). (i) The Shannon entropy of a probability vector on r outcomes lies in $[0, \ln r]$, and $r = \text{rank}(\mathbf{D}_i^\top \mathbf{D}_i) = \text{rank}(\mathbf{D}_i) \leq \min(k, D)$. (ii) The entropy equals $\ln r$ if and only if the distribution is uniform on its r outcomes, that is, the non-zero eigenvalues are equal, and it equals 0 if and only if the distribution is a point mass, that is, $r = 1$. (iii) A rotation $\mathbf{Q}$ maps $\mathbf{C}_i$ to $\mathbf{Q}\mathbf{C}_i\mathbf{Q}^\top$, which has the same eigenvalues; translations leave the displacements unchanged; and a rescaling of the data by $\gamma > 0$ multiplies all eigenvalues by γ^2 , which leaves every $q_{i,t}$ unchanged. $\square$

Rényi entropy at $\alpha \rightarrow 1$. Write $q_t = q_{i,t}$ for a fixed point i , with the sums below running over the t with $q_t > 0$. The Rényi entropy $H_i^{(\alpha)} = \frac{1}{1-\alpha} \ln \sum_t q_t^\alpha$ is of the form $0/0$ at $\alpha = 1$, because $\sum_t q_t = 1$. By L'Hôpital's rule,

$$\lim_{\alpha \rightarrow 1} H_i^{(\alpha)} = - \lim_{\alpha \rightarrow 1} \frac{\sum_t q_t^\alpha \ln q_t}{\sum_t q_t^\alpha} = - \sum_t q_t \ln q_t = H_i.$$

For $\alpha = 0$, $\exp(H_i^{(0)}) = r$ is the rank; for $\alpha = 2$, $\exp(H_i^{(2)}) = (\sum_t \lambda_{i,t})^2 / \sum_t \lambda_{i,t}^2$ is the participation ratio. Since $H_i^{(\alpha)}$ is non-increasing in α , $\exp(H_i^{(2)}) \leq \exp(H_i) \leq \text{rank}(\mathbf{C}_i)$.

B Evaluation metrics

Let $\{U_1, \dots, U_K\}$ be the true classes of the N evaluated points and $\{V_1, \dots, V_{K'}\}$ a clustering of the same points. The contingency counts are $N_{kl} = |U_k \cap V_l|$, with row sums $N_k = \sum_l N_{kl}$ (class sizes) and column sums $N_l = \sum_k N_{kl}$ (cluster sizes). Points labeled as noise by DBSCAN or HDBSCAN form one additional cluster for ARI and VI.

Clustering accuracy (ACC). Because cluster indices are arbitrary, accuracy uses the one-to-one assignment f of clusters to classes that maximizes the number of matched points, found with the Hungarian algorithm [14]:

$$\text{ACC} = \max_f \frac{1}{N} \sum_{i=1}^N \chi(y_i = f(\hat{y}_i)), \quad (14)$$

where $y_i \in \{1, \dots, K\}$ is the class of point i , $\hat{y}_i \in \{1, \dots, K'\}$ its cluster, and the maximum runs over one-to-one maps f from clusters to classes. Points in clusters left without a class, and noise points, count as errors.

Adjusted Rand index (ARI). The Rand index is the fraction of point pairs on which two partitions agree. The ARI [12] corrects it for the agreement expected by chance:

$$\text{ARI} = \frac{\sum_{kl} \binom{N_{kl}}{2} - \left[\sum_k \binom{N_{k\cdot}}{2} \sum_l \binom{N_{\cdot l}}{2} \right] / \binom{N}{2}}{\frac{1}{2} \left[\sum_k \binom{N_{k\cdot}}{2} + \sum_l \binom{N_{\cdot l}}{2} \right] - \left[\sum_k \binom{N_{k\cdot}}{2} \sum_l \binom{N_{\cdot l}}{2} \right] / \binom{N}{2}}. \quad (15)$$

It equals 1 for identical partitions and is close to 0 for independent ones.

Variation of information (VI). VI [19] is a metric distance between the two partitions,

$$\text{VI} = \mathcal{H}(U) + \mathcal{H}(V) - 2\mathcal{I}(U, V), \quad (16)$$

where $\mathcal{H}(U) = -\sum_k \frac{N_{k\cdot}}{N} \ln \frac{N_{k\cdot}}{N}$ and $\mathcal{H}(V) = -\sum_l \frac{N_{\cdot l}}{N} \ln \frac{N_{\cdot l}}{N}$ are the entropies of the class and cluster sizes, and $\mathcal{I}(U, V) = \sum_{kl} \frac{N_{kl}}{N} \ln \frac{N_{kl}}{N_{k\cdot} N_{\cdot l}}$ is their mutual information. Lower values are better, and VI is 0 for identical partitions.

Silhouette coefficient. For a clustered point $\mathbf{x}_i$, let $\delta_{\text{in}}(i)$ be its mean Euclidean distance to the other points of its cluster and $\delta_{\text{out}}(i)$ the smallest mean Euclidean distance from $\mathbf{x}_i$ to the points of another cluster. Its silhouette [20] is

$$\text{sil}(i) = \frac{\delta_{\text{out}}(i) - \delta_{\text{in}}(i)}{\max(\delta_{\text{in}}(i), \delta_{\text{out}}(i))} \in [-1, 1], \quad (17)$$

and the silhouette coefficient is the mean of $\text{sil}(i)$ over all points not labeled as noise, computed in the 20-dimensional input space. It uses no labels and measures how compact and well separated the clusters are; it is undefined when fewer than two clusters remain.

References

- [1] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.
- [2] Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining (PAKDD)*, LNCS 7819, pages 160–172, 2013.
- [3] Frédéric Chazal, Leonidas J. Guibas, Steve Y. Oudot, and Primoz Skraba. Persistence-based clustering in Riemannian manifolds. *Journal of the ACM*, 60(6):41, 2013.
- [4] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995.
- [5] Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.
- [6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 226–231, 1996.
- [7] Elena Facco, Maria d’Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7:12140, 2017.
- [8] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.
- [9] K. Chidananda Gowda and G. Krishna. Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern Recognition*, 10(2):105–112, 1978.

- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- [11] Mark O. Hill. Diversity and evenness: A unifying notation and its consequences. *Ecology*, 54(2):427–432, 1973.
- [12] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.
- [13] R. A. Jarvis and E. A. Patrick. Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on Computers*, C-22(11):1025–1034, 1973.
- [14] Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.
- [15] Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems 17*, 2004.
- [16] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.
- [17] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- [18] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, 2001.
- [19] Marina Meilă. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007.
- [20] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.
- [21] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 547–561, 1961.
- [22] Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.
- [23] Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *Proceedings of the 15th European Signal Processing Conference (EUSIPCO)*, Poznań, Poland, 2007. doi:10.5281/zenodo.40328.
- [24] Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery*, 2(2):169–194, 1998.
- [25] Uri Shaham, Kelly Stanton, Henry Li, Boaz Nadler, Ronen Basri, and Yuval Kluger. SpectralNet: Spectral clustering using deep neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.
- [26] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.
- [27] Vincent A. Traag, Ludo Waltman, and Nees Jan van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9:5233, 2019.
- [28] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

- [29] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. SCAN: Learning to classify images without labels. In *European Conference on Computer Vision (ECCV)*, LNCS 12355, pages 268–285, 2020.
- [30] Andrea Vedaldi and Stefano Soatto. Quick shift and kernel methods for mode seeking. In *European Conference on Computer Vision (ECCV)*, pages 705–718, 2008.
- [31] Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.
- [32] Joe H. Ward, Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.
- [33] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.
- [34] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, PMLR 48, pages 478–487, 2016.
- [35] Lih Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems 17*, pages 1601–1608, 2004.