BEYOND DENSITY: BOUNDARY DETECTION VIA LOCAL MEAN CURVATURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Boundary detection in multivariate datasets remains a fundamental challenge in unsupervised learning. Existing methods rely predominantly on density estimation, which conflates flat low-density regions with genuinely curved geometric transitions and degrades on nonlinear manifolds, anisotropic distributions, and high-dimensional sparse data. We propose the mean curvature boundary points (MCBP) method, a geometry-driven framework that identifies boundary samples by estimating pointwise mean curvature via a discrete approximation of the shape operator on local k-nearest neighbor patches, without requiring explicit manifold parametrization. We establish three theoretical pillars: (i) a variational characterization showing that high-curvature regions coincide with geometric boundaries via the first variation of the Hausdorff measure; (ii) a connection to the Laplace–Beltrami operator that grounds the MCBP within spectral manifold theory; and (iii) a geometric-statistical duality proving that the curvature score jointly amplifies manifold bending and local sparsity, yielding a richer boundary signal than density alone. We further prove pointwise statistical consistency of the discrete estimator. The curvature signal induces a principled data decomposition into smooth interior and high-curvature boundary subsets, acting as a nonlinear geometric filter. Experiments on more than 40 real-world datasets show consistent improvements in terms of internal clustering metrics, with average silhouette coefficient gains of 49% for k-means++ and 58% for HDBSCAN, with curvature-informed centroid initialization outperforming standard k-means++ seeding.

1 INTRODUCTION

Samples lying at the interfaces between distinct data regions (*boundary points*) encode critical structural information for clustering, classification, and manifold learning Xia et al. (2006); Qiu & Cao (2016). In high-dimensional settings their identification is complicated by distance concentration, sparsity, and the degradation of local neighborhood structure Beyer et al. (1999); Bellman (1961), while their intrinsic ambiguity increases sensitivity to noise Li & Maguire (2011); Li et al. (2017).

Existing approaches predominantly rely on *density estimation*: DBSCAN and HDBSCAN Ester et al. (1996); Campello et al. (2015) label boundaries implicitly through core-point reachability; BPF Khalique & Kitagawa (2022) and ARDOD Rahmati et al. (2024) combine density and distance into unified scores; BRIM Qiu et al. (2007) exploits ϵ-neighborhood asymmetry. Despite success in well-separated, homogeneous settings, these approaches share a fundamental limitation: they operate on *extrinsic*, scalar-density signals that cannot distinguish a flat low-density region from a geometrically curved transition zone. On non-linear manifolds, anisotropic distributions, or datasets with heterogeneous sampling densities, this conflation produces imprecise and unstable boundary estimates Donoho & Grimes (2003); Coifman & Lafon (2006).

A second line of work incorporates structural cues beyond density. In geometric reconstruction, BPD and Raw Boundary Smoothing detect concave and convex edges in 3-D point clouds using FFT-based filters Mineo et al. (2019), and BPG3D extends this idea to LiDAR-based 3-D object detection via boundary-aware pooling modules Tang et al. (2024). For high-dimensional tabular data, directed Markov-tree models capture local feature dependencies and perform boundary discovery through combinatorial graph traversal Cao (2021). These methods demonstrate the value of boundary information across domains, but each is tied to a specific data modality, discrete traversal

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

scheme, or task-specific architecture. None provides a general, parameter-light geometric criterion for boundary detection in arbitrary feature spaces.

Geometric machine learning studies how manifold structure, differential operators, and topological features can serve as principled inductive biases for learning Bronstein et al. (2017); Weber (2025); Papillon et al. (2025). Within this framework, curvature is a natural descriptor of local shape: it quantifies how a data manifold bends in the ambient space and can reveal interfaces, folds, and transition regions that density estimates obscure. Hessian Eigenmaps Donoho & Grimes (2003) and diffusion maps Coifman & Lafon (2006) exploit second-order and spectral structure for dimensionality reduction, and graph Laplacians on k-NN graphs are known to converge to the Laplace–Beltrami operator Belkin & Niyogi (2006); von Luxburg (2007), whose embedding identity $\Delta_{\mathcal{M}}\iota = dH\mathbf{n}$ directly connects mean curvature to the spectrum of the data graph. Despite these connections, curvature has remained underexplored as a *practical* boundary criterion in unsupervised learning.

We address this gap with **MCBP** (*Mean Curvature Boundary Points*), a framework grounded in differential geometry and aligned with the emerging paradigm of Geometric Machine Learning Bronstein et al. (2017); Weber (2025); Papillon et al. (2025). MCBP characterizes boundary points as loci of high *mean curvature*, a second-order *intrinsic* quantity estimated from local k-NN patches without requiring an explicit manifold parametrization. The main contributions of this work are:

1. **Theoretically grounded criterion.** We establish mean curvature as a principled boundary surrogate via: (a) the first variation of the Hausdorff measure; (b) the Laplace–Beltrami embedding identity linking curvature to graph Laplacian convergence; and (c) a geometric-statistical duality showing that the estimator jointly amplifies manifold bending and local sparsity.
2. **Consistent discrete estimator.** We prove pointwise statistical consistency of the discrete shape operator under standard manifold sampling assumptions.
3. **Curvature-driven clustering pipeline.** Three novel preprocessing strategies: direct filtering, curvature-informed initialization, and a hybrid geometric-density decomposition, validated on 40+ real-world datasets against density-based baselines.

The remainder of the paper is organized as follows. Section 2 presents the proposed Mean Curvature Boundary Points (MCBP) framework in details. Section 3 reports the computational experiments and the obtained results, and Section 4 presents the conclusions and final remarks. The geometric foundations and theoretical justification are detailed in Appendices A, B and C.

2 The MCBP Framework

The proposed Mean Curvature Boundary Points (MCBP) algorithm estimates local boundary structure through a discrete approximation of the shape operator. In the following, we discuss the proposed method in details.

2.1 Discrete Curvature Estimator

For each sample $\mathbf{x}_i$, we form the local patch $P_i = \{\mathbf{x}_i\} \cup \eta_i$, where η_i is its k-NN set.

First fundamental form proxy. The local covariance

$$\mathbf{\Sigma}_i = \tfrac{1}{k} \sum_{\mathbf{x}_j \in \eta_i} (\mathbf{x}_j - \mathbf{x}_i)(\mathbf{x}_j - \mathbf{x}_i)^\top \tag{1}$$

is symmetric positive-definite under mild conditions and converges to a scaled version of the Riemannian metric $\mathbf{G}(\mathbf{x}_i)$. Its inverse induces a local Mahalanobis geometry that adapts to feature anisotropy Li & Tian (2018).

Second fundamental form proxy. The second fundamental form B is known as the curvature tensor. Let $\{u_\ell\}$ be the eigenvectors of $\mathbf{\Sigma}_i$ (local tangent basis). We build a design matrix $\mathbf{X}_i$ encoding constant, linear, and quadratic terms in these coordinates (following Hessian Eigenmaps Donoho & Grimes (2003)), and extract the second-order block $\mathbf{H}_i$:

$$\mathcal{H}_i = \mathbf{H}_i \, \mathbf{H}_i^\top . \tag{2}$$

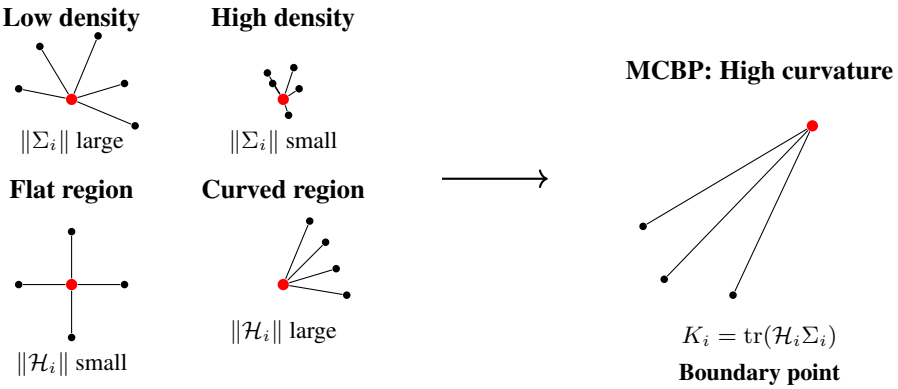


Figure 1: Mean Curvature Boundary Points (MCBP) principle. Top: covariance captures local density. Middle: Hessian captures geometric deformation. Bottom: high curvature emerges from the combination of low density and strong curvature, identifying boundary points.

Discrete shape operator and curvature score. Mirroring the continuous formula $\mathbf{S} = \mathbf{G}^{-1}\mathbf{B}$:

$$\widehat{S}_i = \mathcal{H}_i\,\mathbf{\Sigma}_i, \qquad \widehat{K}_i = \mathrm{tr}(\widehat{S}_i). \tag{3}$$

Scores are min-max normalized to $[0, 1]$. Points with $\widehat{K}_i \geq T_p$ (where T_p is the p-th empirical percentile) are labeled boundary ($B_i = 1$); the rest form the smooth interior S.

Geometric-statistical duality. In the k-NN regime, $\varepsilon(\mathbf{x}_i) \asymp (k/nf(\mathbf{x}_i))^{1/d}$ Devroye et al. (1996), implying $\|\mathbf{\Sigma}_i\|_F \propto f(\mathbf{x}_i)^{-2/d}$. Hence:

$$\widehat{K}_i \;\approx\; \underbrace{\|\mathcal{H}_i\|_F}_{\text{geometric bending}} \;\times\; \underbrace{\|\mathbf{\Sigma}_i\|_F}_{\text{local sparsity}}, \tag{4}$$

assigning high scores to points simultaneously exhibiting low density and strong geometric variation, the defining co-occurrence of boundary regions Genovese et al. (2012). Figure 1 provides an intuitive illustration of the interplay between local density and geometric deformation in the proposed curvature-based framework.

2.2 Statistical Consistency

Proposition 1 (Pointwise consistency). *Let $\mathcal{M}$ be a C^3-smooth compact d-dimensional submanifold of $\mathbb{R}^m$, and let f be a density on $\mathcal{M}$ bounded and bounded away from zero. Let $k = k(n) \to \infty$, $k/n \to 0$, k-NN radius $\varepsilon \to 0$, and $k\varepsilon^2 \to \infty$ as $n \to \infty$. Then for any interior point $\mathbf{x}_i \in \mathcal{M}\backslash\partial\mathcal{M}$:*

$$\widehat{K}_i \;\xrightarrow{\;\mathbb{P}\;}\; H(\mathbf{x}_i) \quad \text{as } n \to \infty. \tag{5}$$

Proof sketch. By a second-order Taylor expansion of the parametrization around $\mathbf{x}_i$ and the law of large numbers over the k-NN ball, $\mathbf{\Sigma}_i \xrightarrow{\mathbb{P}} \frac{\varepsilon^2}{d+2}\mathbf{G}(\mathbf{x}_i)$ at rate $O(k^{-1/2})$. The Hessian proxy $\mathbf{H}_i$, obtained via local polynomial regression on P_i, converges to $c\,\mathbf{B}(\mathbf{x}_i)$ for $c > 0$ depending only on d, consistent with curvature estimation theory for point clouds Aamari & Levrard (2019). Continuity of the trace and the relation $H = \mathrm{tr}(\mathbf{G}^{-1}\mathbf{B})$ then yield $\widehat{K}_i \to H(\mathbf{x}_i)$ after normalization. □ □

Remark 1. *The conditions $k \to \infty$, $k/n \to 0$, $k\varepsilon^2 \to \infty$ are standard in k-NN manifold estimation Aamari & Levrard (2019) and are satisfied by, e.g., $k = \lfloor \log_2 n \rfloor$.*

2.3 Algorithm and Computational Complexity

Algorithm 1 presents the complete MCBP procedure. k-NN graph construction costs $O(nm \log n)$ with KD-trees. The per-sample loop costs $O(m^2k)$ for $\mathbf{\Sigma}_i$, $O(m^3)$ for eigendecomposition, and

$O(m^3)$ for matrix products, giving overall complexity $O(nm \log n) + O(nm^3)$. All per-sample operations are independent and trivially parallelizable. For $m > 50$, a PCA projection to $d \ll m$ components reduces the dominant term to $O(nd^3)$, yielding substantial speedups while preserving the relevant geometric structure.

Algorithm 1 Mean Curvature Boundary Points (MCBP)

Require: $X \in \mathbb{R}^{n \times m}$, neighbors k, percentile $p \in (0, 1)$
Ensure: interior/smooth points S; boundary points B

$A \leftarrow$ kNN-GRAPH(X, k)
for $i = 1$ **to** n **do**
  $\mathbf{\Sigma}_i \leftarrow \frac{1}{k} \sum_{\mathbf{x}_j \in \eta_i} (\mathbf{x}_j - \mathbf{x}_i)(\mathbf{x}_j - \mathbf{x}_i)^\top$
  $U_i \leftarrow \text{eigvecs}(\mathbf{\Sigma}_i)$ ▷ local tangent basis
  Build $\mathbf{X}_i$: const., linear, quadratic in U_i-coords
  $\mathbf{H}_i \leftarrow$ second-order columns of $\mathbf{X}_i$
  $\mathcal{H}_i \leftarrow \mathbf{H}_i \mathbf{H}_i^\top$
  $\widehat{K}_i \leftarrow \text{tr}(\mathcal{H}_i \, \mathbf{\Sigma}_i)$
end for
$\widehat{K} \leftarrow (\widehat{K} - \min \widehat{K}) / (\max \widehat{K} - \min \widehat{K})$
$T_p \leftarrow \text{percentile}(\widehat{K}, p)$
$B_i \leftarrow \mathbf{1}[\widehat{K}_i \geq T_p]$ for all i
$B_i \leftarrow \mathbf{0}[\widehat{K}_i < T_p]$ for all i
$S_p \leftarrow \{x_i : B_i = 0\}$; $B_p \leftarrow \{x_i : B_i = 1\}$
return B, S

Beyond boundary detection, MCBP induces a simple dataset decomposition. For a chosen threshold T_p, we define $S_p = \{\mathbf{x}_i \in X : K_i < T_p\}$ and $B_p = \{\mathbf{x}_i \in X : K_i \geq T_p\}$, where S_p is the *smooth/inner set* and B_p is the *boundary set*. In clustering applications, algorithms such as k-means, spectral clustering, or density-based methods can first be applied to S_p, where samples are less ambiguous and cluster cores are more stable. Boundary samples in B_p can then be discarded for filtering, analyzed as transition points, or assigned labels afterward using a local interpolation rule such as 1-nearest neighbor. This strategy treats boundaries not as arbitrary noise, but as a separate high-curvature component of the data geometry.

3 Computational Experiments

We evaluate MCBP along two axes. First, we use synthetic two-dimensional data to verify whether the curvature scores highlight visually meaningful boundary regions. Second, we evaluate MCBP as a preprocessing step for clustering on real datasets. The goal is not to propose a new clustering algorithm, but to test whether removing or separately handling high-curvature samples yields cleaner cluster structure for standard unsupervised methods.

3.1 Synthetic Boundary Visualization

We validate the geometric intuition of MCBP on four controlled 2-D benchmarks: a single isotropic Gaussian blob ($n = 400$), two isotropic Gaussian clusters ($n = 500$), two elliptical (anisotropic) clusters ($n = 500$), and two-moons ($n = 1000$). In all four benchmarks, high-curvature scores concentrate at cluster peripheries and inter-cluster transition zones, while dense interior regions receive near-zero scores. On the anisotropic dataset, MCBP correctly adapts to the principal axes of each ellipse, assigning elevated scores to extremities and minor-axis tips, behavior inaccessible to isotropic density estimators. On two-moons, the algorithm traces boundary points along the strongly curved arcs without any global convexity assumption, demonstrating sensitivity to underlying manifold topology. Figure 2 shows a representative two-moons dataset. The first panel shows the estimated curvature field, and the right panel shows the high-curvature samples selected as boundary points. MCBP highlights the outer arcs, low-density extremities, and transition regions where local neighborhoods are less well approximated by a single tangent direction. This qualitative behavior is consistent with the role of curvature: low scores correspond to smooth cluster interiors, while high scores identify geometrically unstable regions.

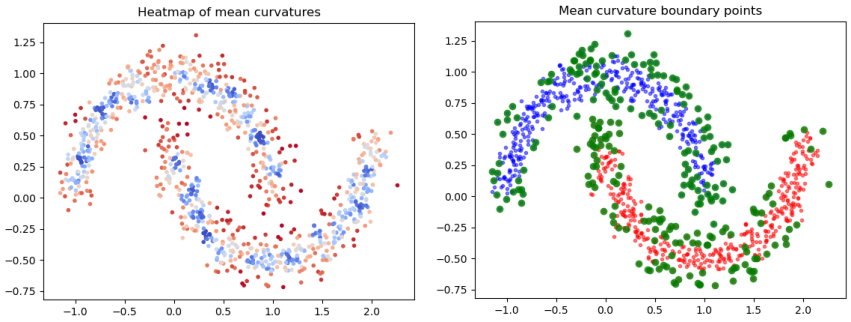


Figure 2: Synthetic boundary detection with MCBP on a two-moons dataset: estimated mean-curvature heatmap, and high-curvature boundary samples selected by percentile thresholding ($T = 0.75$ percentile).

Table 1: Complete list of datasets used in the computational experiments.

Dataset	Samples	Feat.	Cls.	Dataset	Samples	Feat.	Cls.	Dataset	Samples	Feat.	Cls.
AP_Breast_Colon	630	10935	2	**heart-h**	294	13	2	**satimage**	6430	36	6
AP_Omentum_Kidney	337	10935	2	**Indian_pines**	9144	220	8	**seeds**	210	7	3
AP_Omentum_Ovary	275	10935	2	**ionosphere**	351	34	2	**segment**	2310	19	7
arsenic-male-bladder	559	5	14	**ipums_la_98-small**	7487	60	7	**solar-flare**	1066	12	6
artificial-characters	10218	7	10	**ipums_la_99-small**	8844	60	7	**steel-plates-fault**	1941	33	2
balance-scale	625	4	3	**iris**	150	4	3	**tae**	151	5	3
cardiotocography	2126	35	10	**LED-display-domain-7digit**	500	7	10	**tecator**	240	124	2
coil-20	1440	1024	20	**led7**	3200	7	10	**texture**	5500	40	11
dermatology	366	34	6	**letter**	20000	16	26	**thoracic_surgery**	470	16	2
diabetes	768	8	2	**mammography**	11183	6	2	**thyroid-dis**	2800	26	7
digits	1797	64	10	**mfeat-factors**	2000	216	10	**tr11.wc**	414	6430	9
DLBCL	77	5470	2	**mfeat-morphological**	2000	6	10	**tr23.wc**	204	5833	6
doggle_table_a2	310	8	9	**mfeat-pixel**	2000	240	10	**tr41.wc**	878	7455	10
Engine1	383	5	3	**mfeat-zernike**	2000	47	10	**UMIST_Faces_Cropped**	575	10304	20
Fashion-MNIST	70000	784	10	**optdigits**	5620	64	10	**USPS**	9298	256	10
fl2000	67	15	5	**pendigits**	10992	16	10	**vote**	435	17	2
Flare	1066	11	6	**penguins**	344	6	3	**vowel**	990	12	11
gas-drift	13910	128	6	**prnn_synth**	250	2	2	**wine**	178	13	3
glass	214	9	6	**prnn_viruses**	61	18	4	**wine-quality-red**	1599	11	6
hayes-roth	160	4	3	**Satellite**	5100	36	2				

3.2 Real-World High-Dimensional Benchmarks

We evaluate on 40+ publicly available OpenML datasets spanning diverse domains, sample sizes n, dimensions m, and class counts c. All datasets are standardized to zero mean and unit variance; when $m > 50$, PCA projection to 50 components precedes curvature estimation. MCBP uses $k = \lfloor \log_2 n \rfloor$ and $p = 0.75$ uniformly across all experiments. Quality is assessed by three complementary internal indices: Silhouette Coefficient (SC↑), Calinski–Harabasz (CH↑), and Davies–Bouldin (DB↓). Complete per-dataset tables are omitted from the main text and should be moved to the supplementary material. Table 1 lists the complete set of datasets used in the computational experiments. Each block of rows reports the dataset name, number of samples, number of attributes, and number of classes.

Strategy A: Geometric filtering for k-means. Table 2 compares k-means++ applied to raw X versus filtered interior S. Filtering yields consistent improvements across all 40 datasets used in this experiment: average SC increases by $+49\%$ ($0.193 \to 0.288$), CH by $+37\%$ ($693.7 \to 948.1$), and DB decreases by -26% ($2.335 \to 1.737$). Gains are particularly pronounced on datasets with high noise or class overlap (`gas-drift`: SC$+63\%$; `tr23.wc`: SC$+441\%$), supporting the hypothesis that high-curvature samples correspond to structurally ambiguous regions that degrade centroid estimation.

Strategy B: Curvature-informed initialization. Table 3 compares standard k-means++ seeding against S-centroid initialization (seeds from k-means++ on S, applied to full X). S-centroids raise average SC by $+29\%$ ($0.190 \to 0.246$) and reduce DB by -19%. The strategy is especially effective on small sample size extreme high-dimensional genomics datasets (`DLBCL`, `AP_Omentum`: SC

Table 2: Internal indices obtained after k-means++ clustering of the original raw data X and the filtered data S.

	Original data			Filtered data		
Datasets	**SC**	**CH**	**DB**	**SC**	**CH**	**DB**
iris	0.4799	157.3602	0.7894	**0.5303**	**193.6954**	**0.6574**
wine	0.2840	70.9400	1.3890	**0.3700**	**86.8090**	**1.0590**
digits	0.0171	**28.4674**	3.8744	**0.0570**	27.1675	**3.5870**
prnn_viruses	0.2448	13.7203	1.5892	**0.4139**	**29.4246**	**0.8939**
diabetes	0.1646	146.7425	2.0994	**0.2324**	**156.9266**	**1.7507**
LED-display-domain-7digit	0.3793	104.6648	1.2722	**0.6796**	**252.6741**	**0.6035**
dermatology	0.1566	67.3073	2.2340	**0.1883**	**73.4740**	**2.1060**
led7	0.3864	693.9301	1.2855	**0.5801**	**993.8856**	**0.7647**
steel-plates-fault	0.2767	515.7229	1.2606	**0.3037**	**565.4308**	**1.1417**
Flare	0.2114	180.6994	1.4335	**0.3787**	**293.3462**	**1.0613**
pendigits	0.2738	2,307.2487	1.3821	**0.3076**	**2,496.4991**	**1.2895**
satimage	0.3522	4,797.1089	1.0234	**0.4314**	**6,171.9679**	**0.7742**
mfeat-pixel	0.0333	**30.8179**	3.8825	**0.0548**	29.7255	**3.5305**
mfeat-morphological	0.5378	2,836.8208	0.7260	**0.5834**	**3,458.4824**	**0.6470**
optdigits	0.0140	**90.1927**	3.8637	**0.0540**	87.5224	**3.4536**
mammography	0.4904	5,977.7711	0.9609	**0.6374**	**11,530.2653**	**0.6572**
USPS	-0.0549	147.7980	3.8465	**-0.0129**	**179.0574**	**3.0914**
Satellite	0.4301	5,216.9835	0.9202	**0.5139**	**6,185.1103**	**0.7433**
ipums_la_99-small	0.0007	144.1783	4.5469	**0.0499**	**169.0616**	**3.5386**
ipums_la_98-small	0.0050	115.3043	4.7772	**0.0427**	**146.0484**	**3.7011**
cardiotocography	0.1614	201.3592	1.8590	**0.2163**	**199.6298**	**1.6167**
gas-drift	0.0860	201.9786	3.4622	**0.1403**	**549.3173**	**2.1817**
AP_Breast_Colon	0.0312	8.5064	**2.8206**	**0.0551**	**9.2042**	5.5461
vowel	0.1745	**125.5085**	1.6898	**0.2163**	113.8239	**1.5497**
ionosphere	0.2704	95.9125	1.6819	**0.3865**	**137.0405**	**1.2671**
solar-flare	0.2163	188.5326	1.4525	**0.3615**	**273.0862**	**1.2735**
vote	0.2862	181.3880	1.4480	**0.4086**	**277.0933**	**1.0295**
fl2000	0.4121	40.2213	0.7060	**0.5295**	**56.2753**	**0.5445**
seeds	0.4024	249.6550	0.9221	**0.4761**	**266.3524**	**0.7563**
coil-20	0.0575	26.2887	2.6548	**0.1799**	**36.3722**	**2.1835**
letter	0.1454	**1,291.1190**	1.6533	**0.1678**	1,121.6117	**1.5953**
Fashion-MNIST	0.0448	1,188.3357	3.7282	**0.0642**	**1,257.6910**	**3.2775**
thyroid-dis	0.1513	206.9558	2.2627	**0.2910**	**334.5878**	**1.0182**
UMIST_Faces_Cropped	0.0932	10.3407	2.5635	**0.1496**	**12.3307**	**2.1875**
tr23.wc	0.1073	4.2307	2.1373	**0.5799**	**5.9179**	**0.2816**
tr41.wc	0.1361	18.5660	1.4231	**0.1740**	**85.9308**	**1.1306**
DLBCL	0.0195	1.0655	0.9407	**0.0878**	**1.3696**	**0.8255**
thoracic_surgery	0.1704	**55.7060**	2.3208	**0.2235**	54.3333	**1.4912**
AP_Omentum_Kidney	0.0513	**4.9412**	6.9474	**0.2617**	3.3424	**1.4653**
AP_Omentum_Ovary	0.015	**4.198**	7.584	**0.137**	3.391	**3.223**
Average	0.193	693.715	2.335	**0.288**	**948.132**	**1.737**
Median	0.163	134.843	1.686	**0.247**	**162.994**	**1.281**

gains $> 4\times$) and high-dimensional text corpora (`tr23.wc`: $0.107 \rightarrow 0.481$, $+349\%$), where boundaries are geometrically complex and density-based seeds are unreliable.

Strategy C: Geometric filtering for HDBSCAN Table 4 evaluates HDBSCAN on raw X versus filtered S (`min_cluster_size` selected from $\{5, 10, 20\}$ to maximize SC per dataset). Filtering raises average SC by $+58\%$ ($0.227 \rightarrow 0.359$) and reduces DB by -20%. The CH spike (from 183.4 to 7187) reflects datasets like `led7` and `satimage`, where boundary removal eliminates bridge samples that previously caused cluster merging in the density hierarchy.

Strategy D: Hybrid approach with k-means++ with HDBSCAN S-centroids Table 5 evaluates HDBSCAN on raw X versus a hybrid approach based on k-means++ initialized with the S-centroids obtained after HDBSCAN on S (S_H). The replacement of the regular k-means++ initialization centroids by the HDBSCAN S-centroids raises average SC by $+112\%$ ($0.1427 \rightarrow 0.3025$), the average CH by $+518\%$ ($94.5314 \rightarrow 584.9311$) and reduces DB by -47% ($2.8756 \rightarrow 1.5113$). Datasets like `satimage`, `mammography`, `letter` and `artificial-characters` where boundaries are highly non-linear with overlapping classes.

Strategy E: Hybrid approach with HDBSCAN on S and 1-NN classification on B Table 6 reports the complementary HDBSCAN-on-S + 1-NN-on-B strategy. The 1-NN classifier is trained with the labels learned by HDBSCAN. Average SC doubles ($0.105 \rightarrow 0.231$, $+120\%$) and average CH triples ($103.9 \rightarrow 325.6$, $+213\%$). The hybrid approach is particularly effective on highly overlapping datasets (`satimage`: SC $-0.217 \rightarrow 0.497$; `ipums_la_98`: SC $0.029 \rightarrow 0.420$).

The experimental results across $40+$ datasets reveal three consistent and practically relevant patterns.

Table 3: Internal indices obtained after kmeans clustering of the original raw data X using different centroid initialization strategies: kmeans++ versus *kmeans++ S-centroids* (centroids obtained by kmeans++ in the filtered data S).

	kmeans++			kmeans++ (S-centroids)		
Datasets	**SC**	**CH**	**DB**	**SC**	**CH**	**DB**
diabetes	0.1646	146.7425	2.0994	**0.1952**	**151.6192**	**2.0179**
LED-display-domain-7digit	0.3793	104.6648	1.2722	**0.4143**	**120.2487**	**1.1057**
led7	0.3864	693.9301	1.2855	**0.4067**	**729.8705**	**1.1127**
Flare	0.2114	**180.6994**	1.4335	**0.2691**	172.3430	**1.4871**
pendigits	0.2738	2,307.2487	1.3821	**0.2794**	**2,362.1967**	**1.3388**
cardiotocography	0.1614	201.3592	1.8590	0.2146	228.1636	1.5477
AP_Breast_Colon	0.0312	8.5064	**2.8206**	**0.1707**	**8.7406**	5.6799
vowel	0.1745	125.5085	1.6898	**0.1890**	**131.5905**	**1.6603**
Fashion-MNIST	0.0448	1,188.3357	**3.7282**	**0.0509**	**1,209.6117**	3.8089
thyroid-dis	0.1513	206.9558	2.2627	**0.2282**	**208.1820**	**1.8495**
tr23.wc	0.1073	4.2307	2.1373	**0.4806**	**4.3882**	**0.3662**
balance-scale	0.1652	127.2775	1.7583	0.1718	135.4921	1.7033
texture	0.2907	3,475.1554	**1.1527**	**0.2998**	**3,502.9030**	1.1800
Indian_pines	0.0235	136.8340	4.5198	**0.0284**	**142.8617**	**4.3562**
tae	0.2192	42.3612	1.6399	**0.2663**	**49.4178**	**1.4677**
doggle_table_a2	0.4335	453.7007	0.8389	**0.4641**	**487.4629**	**0.8026**
hayes-roth	0.1929	35.3936	1.6519	**0.2054**	**37.2918**	**1.5809**
prnn_viruses	0.2448	13.7203	1.5892	**0.2893**	**18.7570**	**1.4653**
DLBCL	0.0195	1.0655	0.9407	**0.1042**	**1.4553**	**0.8051**
artificial-characters	0.2107	2,033.8132	1.3717	**0.2339**	**2,095.2515**	**1.2515**
wine-quality-red	0.1808	275.3569	1.4589	**0.1923**	**279.7846**	**1.4160**
AP_Omentum_Kidney	0.0513	4.9412	6.9474	**0.1758**	**2.3146**	**1.6035**
AP_Omentum_Ovary	0.0154	4.1981	7.5836	**0.1023**	**2.5550**	**3.1436**
coil-20	0.0575	26.2887	**2.6548**	**0.0670**	**26.5775**	2.6930
tr11.wc	0.5532	9.2731	1.5022	**0.6565**	**9.2771**	**1.4780**
Average	0.190	472.302	2.303	**0.246**	**484.734**	**1.877**
Median	0.174	127.278	1.652	**0.215**	**135.492**	**1.487**

Table 4: Internal indices obtained after HDBSCAN clustering of the original raw data X and the filtered data S.

	Original data			Filtered data		
Datasets	**SC**	**CH**	**DB**	**SC**	**CH**	**DB**
iris	0.489	136.990	0.535	**0.648**	**313.973**	**0.508**
wine	0.134	24.689	3.132	**0.261**	**33.415**	**2.352**
digits	0.096	9.368	7.201	**0.119**	**10.434**	**6.821**
prnn_viruses	0.320	13.819	1.456	**0.387**	**21.063**	**1.381**
LED-display-domain-7digit	0.7154	111.9395	1.0400	**0.8896**	**529.3976**	**0.6375**
dermatology	0.2467	52.4151	2.6514	**0.3381**	**80.3264**	**1.7295**
led7	0.9646	3,038.0936	0.9334	**0.9992**	**229,218.6462**	**0.0622**
steel-plates-fault	0.0477	43.2375	1.4368	**0.1294**	**77.1314**	**1.3803**
pendigits	0.0892	421.9440	1.4500	**0.2071**	**745.0504**	**1.4050**
satimage	-0.2169	315.8412	1.5841	**0.5531**	**2,026.9532**	**0.5444**
mfeat-morphological	0.4424	974.4174	1.0192	**0.5349**	**1,799.6987**	**0.5764**
optdigits	0.1104	34.6735	8.4633	**0.1628**	**20.3868**	**7.3756**
mammography	0.3062	114.9760	1.7156	**0.4754**	**460.7328**	**1.5017**
Satellite	0.1002	81.5235	**2.6807**	**0.2137**	**455.4285**	3.2347
ipums_la_99-small	0.1559	66.2524	4.2294	**0.2093**	**80.7173**	**2.4364**
ipums_la_98-small	0.3268	74.5943	5.3899	**0.4534**	**54.2004**	**0.7727**
gas-drift	-0.0219	16.0178	1.8449	**0.0816**	**58.0739**	**1.4101**
vowel	0.3149	31.3645	**1.2386**	**0.3486**	**33.0159**	1.2417
ionosphere	0.0297	18.0542	1.8830	**0.1763**	**36.1569**	**1.7077**
solar-flare	0.6394	24.0040	1.4333	**0.8586**	**207.1133**	**1.0736**
fl2000	0.3590	14.4283	1.4358	**0.5700**	**38.7709**	**0.8878**
seeds	0.0873	41.2133	2.0715	**0.2427**	**103.2725**	**2.4370**
coil-20	0.0822	22.3140	2.5420	**0.1871**	**26.7769**	**1.8239**
thyroid-dis	0.2477	92.2377	1.5941	**0.3112**	**203.4515**	**1.2777**
UMIST_Faces_Cropped	0.0802	8.8354	2.4356	**0.1285**	**10.7997**	**2.2106**
artificial-characters	0.0216	30.1161	1.2112	**0.1240**	**43.4042**	**1.1666**
thoracic_surgery	0.0810	21.2529	2.1302	**0.1380**	**34.4886**	**2.0854**
prnn_synth	0.0772	39.7613	1.8218	**0.3623**	**104.9119**	**1.3923**
Engine1	0.0848	31.9535	**2.4220**	**0.2162**	**45.5290**	4.0726
glass	0.3887	24.8534	2.0361	**0.5544**	**55.2599**	**0.5266**
tea	0.2242	20.8471	1.4813	**0.3307**	**27.3312**	**1.2394**
heart-h	0.0077	12.6209	**2.2074**	**0.1001**	**18.1224**	2.2838
penguins	0.4745	88.3367	1.6968	**0.5509**	**205.7073**	**1.6600**
Average	0.227	183.424	2.315	**0.359**	**7,187.265**	**1.855**
Median	0.134	34.673	1.822	**0.311**	**58.074**	**1.405**

Gains are largest when datasets exhibit overlapping classes, nonlinear cluster geometry, or heterogeneous sampling density, precisely the settings where density-based methods degrade. In these configurations, inter-cluster transition zones are geometrically complex: points near them simultaneously display low local density *and* high manifold bending, producing elevated curvature scores

Table 5: Internal indices obtained by clustering the original raw data X after HDBSCAN and a hybrid approach based on k-means++ with *HDBSCAN S-centroids* (centroids obtained by HDBSCAN in the filtered data S)

	HDBSCAN			kmeans (S-centroids)		
Datasets	**SC**	**CH**	**DB**	**SC**	**CH**	**DB**
iris	0.489	136.990	0.535	**0.582**	**251.349**	**0.593**
wine	0.134	24.689	3.132	**0.268**	**69.486**	**1.448**
digits	0.096	9.368	7.201	**0.135**	**31.363**	**2.534**
diabetes	**0.256**	35.951	2.478	0.180	**156.517**	**1.683**
dermatology	0.2467	52.4151	2.6514	**0.2776**	**120.6953**	**1.4836**
steel-plates-fault	0.0477	43.2375	1.4368	**0.2237**	**151.7721**	**1.4705**
pendigits	0.0892	421.9440	1.4500	**0.2477**	**1,652.3285**	**1.4119**
satimage	-0.2169	315.8412	1.5841	**0.4900**	**2,886.3704**	**0.7594**
mfeat-pixel	-0.0294	5.3032	7.3308	**0.0525**	**24.7033**	**3.3115**
mfeat-morphological	0.4424	974.4174	1.0192	**0.5678**	**2,690.6942**	**0.6375**
optdigits	0.1104	34.6735	8.4633	**0.2034**	**89.4938**	**2.1026**
mammography	0.3062	114.9760	1.7156	**0.4654**	**1,564.7656**	**0.9931**
Satellite	0.1002	81.5235	2.6807	**0.4301**	**5,216.9835**	**0.9202**
ipums_la_99-small	0.1559	66.2524	4.2294	**0.1835**	**146.5300**	**1.6689**
ipums_la_98-small	0.3268	74.5943	5.3899	**0.3599**	**133.6102**	**1.0969**
gas-drift	-0.0219	16.0178	1.8449	**0.2321**	**138.3420**	**1.5309**
vowel	0.3149	31.3645	1.2386	**0.3185**	**64.8741**	**1.0710**
ionosphere	0.0297	18.0542	1.8830	**0.2514**	**47.0749**	**1.4971**
fl2000	0.3590	14.4283	1.4358	**0.6919**	**60.9354**	**0.7527**
seeds	0.0873	41.2133	2.0715	**0.4658**	**255.8548**	**0.7969**
coil-20	0.0822	22.3140	2.5420	**0.1300**	**27.8713**	**2.0897**
letter	-0.1965	43.6669	**1.3480**	**0.1808**	**336.1503**	1.4207
Fashion-MNIST	0.0279	73.9698	5.3302	**0.1000**	**1,112.9422**	**4.7994**
thyroid-dis	0.2477	92.2377	1.5941	**0.2545**	**178.4132**	**1.2854**
artificial-characters	0.0216	30.1161	1.2112	**0.3053**	**508.7349**	**1.0043**
thoracic_surgery	0.0810	21.2529	2.1302	**0.1137**	**28.3253**	**2.0568**
prnn_synth	0.0772	39.7613	1.8218	**0.4741**	**295.0686**	**0.6855**
Engine1	0.0848	31.9535	2.4220	**0.3311**	**194.4901**	**1.2556**
optdigits	0.1104	34.6735	8.4633	**0.2034**	**89.4938**	**2.1026**
tea	0.2242	20.8471	1.4813	**0.3542**	**43.7423**	**0.8723**
heart-h	0.0077	12.6209	2.2074	**0.1386**	**31.8714**	**2.1028**
penguins	**0.4745**	88.3367	1.6968	0.4677	**116.9486**	**0.9227**
Average	0.1427	94.5314	2.8756	**0.3025**	**584.9311**	**1.5113**
Median	0.0981	37.8563	1.9772	**0.2614**	**142.4360**	**1.4163**

Table 6: Internal indices obtained by clustering the original raw data X after HDBSCAN and a hybrid approach based on HDBSCAN on S and 1-NN on the B. The partition of X into S and B was performed using the threshold T (50% or 75% mean curvature percentiles).

	HDBSCAN			HDBSCAN on S + 1-NN on B			
Datasets	**SC**	**CH**	**DB**	**SC**	**CH**	**DB**	**T**
iris	0.4891	136.9898	0.5345	**0.5818**	**251.3493**	**0.5933**	75
wine	0.1338	24.6893	3.1319	**0.2198**	**32.5265**	**2.3224**	75
digits	-0.0414	29.5779	**2.0500**	**0.0113**	**44.9314**	2.1752	75
dermatology	0.2467	52.4151	2.6514	**0.2926**	**104.5889**	**1.3689**	75
pendigits	0.0892	421.9440	1.4500	**0.1581**	**744.4372**	**1.6623**	75
satimage	-0.2169	315.8412	1.5841	**0.4967**	**2,858.1257**	**0.7130**	75
mfeat-zernike	0.1324	85.8501	2.4330	**0.1591**	**94.7533**	**2.2354**	50
mfeat-factors	0.0546	11.9177	4.3286	**0.1149**	**143.3338**	**1.9928**	50
optdigits	0.2976	79.9268	1.8549	**0.3060**	**102.5913**	**1.0660**	50
mammography	0.3062	114.9760	**1.7156**	**0.3375**	**139.2778**	1.7466	75
Satellite	0.1002	81.5235	2.6807	**0.1765**	**544.5387**	**2.5399**	75
ipums_la_98-small	0.0288	55.9285	1.6963	**0.4201**	**167.1648**	**0.8247**	75
ionosphere	0.0297	18.0542	**1.8830**	**0.0913**	**20.8204**	2.2230	75
seeds	0.0873	41.2133	2.0715	**0.2578**	**124.3263**	**2.3347**	75
prnn_synth	0.0772	39.7613	1.8218	**0.2712**	**102.4417**	**1.4543**	75
Engine1	0.0848	31.9535	**2.4220**	**0.1632**	**40.1255**	4.1937	75
texture	-0.0617	365.0684	**1.5455**	**0.0620**	**859.6780**	1.8435	50
arsenic-male-bladder	0.1826	105.3658	1.4827	**0.2752**	**149.8674**	**1.2652**	75
segment	0.0681	80.8490	**1.4577**	**0.1338**	**215.3052**	2.2941	50
AP_Breast_Colon	-0.0441	26.8911	**3.2052**	**0.0195**	**27.2369**	4.8201	50
tecator	0.1545	61.6940	**2.2033**	**0.2986**	**71.1271**	3.0470	75
Average	0.1047	103.9253	2.1049	**0.2308**	**325.6451**	**2.0341**	
Median	0.0873	61.6940	**1.8830**	**0.2198**	**124.3263**	1.9928	

(Eq. 4). Removing or isolating these points before clustering eliminates a primary source of centroid displacement and density-hierarchy instability, explaining the pronounced gains observed on datasets such as `gas-drift`, `satimage`, and `tr23.wc`. Conversely, when clusters are already well-separated and approximately isotropic, such as `penguins` or `seeds`, the curvature signal is weaker and the filtering benefit is modest but never harmful.

The percentile threshold p governs the size of the boundary set B and thus the aggressiveness of filtering. Our sensitivity analysis shows that performance is robust across $p \in [0.70, 0.85]$, with the default $p = 0.75$ performing consistently near-optimally. This behavior reflects the structure of the curvature distribution: the bulk of interior points occupy a compact low-curvature range, while boundary points form a heavier tail that is well-separated from the interior mass, making percentile-based thresholding a natural and stable criterion. For practitioners, we recommend starting with $p = 0.75$ and adjusting only if domain knowledge suggests a substantially different boundary proportion.

A recurring finding is that discarding boundary samples entirely is not always optimal. The HDBSCAN-on-S + 1-NN-on-B hybrid strategy (Strategy E, Table 6) consistently outperforms both standalone HDBSCAN on X and filtering followed by label restriction to S alone. This demonstrates that boundary samples carry genuine structural information: once cluster cores are robustly identified in the geometrically smooth interior, boundary points can be assigned labels via proximity with minimal distortion. Treating the set B as an ambiguous but labelable region, rather than noise to discard, can be both theoretically natural and empirically superior.

MCBP is not a replacement for density or graph-based clustering; it is a geometry-aware preprocessing layer that improves the conditions under which any downstream algorithm operates. Its primary cost is the cubic dependence on ambient dimension m, which we address via PCA projection to $d \leq 50$ components, a step that preserves the relevant geometric structure while reducing computation. In the few cases where MCBP yields marginal or mixed results (e.g., `letter`, `vowel`, slight DB degradation in some HDBSCAN runs), the cause is typically a well-separated dataset where curvature is already uniformly low, so filtering produces negligible change rather than harm. No dataset in our benchmark exhibited a statistically meaningful deterioration across all three metrics simultaneously, suggesting that MCBP can be applied as a default preprocessing step with low risk.

4 Conclusion

We introduced MCBP, a curvature-driven framework for boundary detection and geometry-aware dataset preprocessing grounded in differential geometry and manifold learning. In contrast to classical density-based approaches, MCBP characterizes boundary points through local geometric deformation, using a discrete approximation of mean curvature derived from first- and second-order neighborhood structure.

The proposed formulation is supported by three complementary perspectives. From a geometric viewpoint, boundary regions correspond to locations of high curvature and strong sensitivity to local perturbations. From a statistical viewpoint, the proposed estimator combines curvature and local sparsity through the interaction between the Hessian approximation and neighborhood covariance structure. From a spectral viewpoint, the connection between mean curvature and the Laplace–Beltrami operator links MCBP to graph-based manifold representations and diffusion geometry. Together, these results establish MCBP as a principled geometric alternative to purely density-driven boundary detection methods.

Empirically, the curvature-based partition $X = S \cup B$ consistently improves clustering performance across more than 40 datasets using both centroid-based and density-based algorithms. The gains are particularly significant in high-dimensional, noisy, and geometrically non-linear settings, suggesting that curvature provides a robust signal for identifying transition regions that are poorly captured by density alone.

More broadly, the results reinforce the idea that geometric quantities traditionally studied in differential geometry can play a practical role in modern machine learning pipelines. We believe this perspective opens several promising directions at the intersection of geometric machine learning, manifold estimation, and representation learning, including adaptive neighborhood geometry, scalable curvature estimation, graph neural architectures, and curvature-regularized latent representations.

AI USE STATEMENT

In this work, we used generative AI tools to polish writing. We have not used generative AI tools to generate synthetic data sets, interpret results, implement methods, or tasks regarding the support for mathematical proofs. Lastly, we take responsibility for the final content of this work.

REFERENCES

Eddie Aamari and Clément Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. *The Annals of Statistics*, 47(1):177 – 204, 2019. doi: 10.1214/18-AOS1685. URL https://doi.org/10.1214/18-AOS1685.

Eddie Aamari, Jisu Kim, Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Estimating the reach of a manifold. *Electronic Journal of Statistics*, 13(1):1359 – 1399, 2019. doi: 10.1214/19-EJS1551. URL https://doi.org/10.1214/19-EJS1551.

Yariv Aizenbud and Barak Sober. Non-parametric estimation of manifolds from noisy data. *arXiv preprint arXiv:2105.04754*, 2021.

Mikhail Belkin and Partha Niyogi. Convergence of laplacian eigenmaps. In B. Schölkopf, J. Platt, and T. Hoffman (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper_files/paper/2006/file/5848ad959570f87753a60ce8be1567f3-Paper.pdf.

Richard Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? *International conference on database theory*, pp. 217–235, 1999.

Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4): 18–42, 2017. doi: 10.1109/MSP.2017.2693418.

Ricardo JGB Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data*, 10(1):1–51, 2015.

X Cao. High-dimensional cluster boundary detection using directed markov tree. *Pattern Analysis and Applications*, 24:35–47, 2021.

Fan R. K. Chung. *Spectral Graph Theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, Providence, RI, 1997. ISBN 0821803158 9780821803158.

Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, New York, 1996. ISBN 978-0-387-94618-4. doi: 10.1007/978-1-4612-0711-5.

Manfredo P. do Carmo. *Differential Geometry of Curves and Surfaces: Revised and Updated Second Edition*. Dover Publications, New York, UNITED STATES, 2016. ISBN 978-0-486-81797-2.

David L Donoho and Carrie Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003.

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)*, pp. 226–231, 1996.

Christopher R. Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Manifold estimation and singular deconvolution under Hausdorff loss. *The Annals of Statistics*, 40(2):941 – 963, 2012. doi: 10.1214/12-AOS994. URL https://doi.org/10.1214/12-AOS994.

Mark Kac. Can one hear the shape of a drum? *The American Mathematical Monthly*, 73(4):1–23, 1966. ISSN 00029890, 19300972. URL http://www.jstor.org/stable/2313748.

Vijdan Khalique and Hiroyuki Kitagawa. Bpf: An effective cluster boundary points detection technique. In Christine Strauss, Alfredo Cuzzocrea, Gabriele Kotsis, A. Min Tjoa, and Ismail Khalil (eds.), *Database and Expert Systems Applications*, pp. 404–416, Cham, 2022. Springer International Publishing.

Dewei Li and Yingjie Tian. Survey and experimental study on metric learning methods. *Neural Networks*, 105:447–462, 2018. ISSN 0893-6080.

Xiaojie Li, Xi Wu, Jiancheng Lv, Jia He, Jianping Gou, and Mao Li. Automatic detection of boundary points based on local geometrical measures. *Soft Computing*, 22:3663–3674, 2017.

Yuhua Li and Liam Maguire. Selecting critical patterns based on local geometrical and statistical information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(6):1189–1201, 2011.

Carmelo Mineo, Stephen Gareth Pierce, and Rahul Summan. Novel algorithms for 3d surface point cloud boundary detection and edge reconstruction. *Journal of Computational Design and Engineering*, 6(1):81–91, 2019.

Boaz Nadler, Stéphane Lafon, Ronald R. Coifman, and Ioannis G. Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, 21(1):113–127, 2006.

Mathilde Papillon, Sophia Sanborn, Johan Mathe, Louisa Cornelis, Abby Bertics, Domas Buracas, Hansen J. Lillemark, Christian Shewmake, Fatih Dinc, Xavier Pennec, and Nina Miolane. Beyond Euclid: An illustrated guide to modern machine learning with geometric, topological, and algebraic structures. *Machine Learning: Science and Technology*, 6, 2025. doi: 10.1088/2632-2153/adf375.

Bao-Zhi Qiu, Feng Yue, and Jun-Yi Shen. Brim: An efficient boundary points detecting algorithm. In Zhi-Hua Zhou, Hang Li, and Qiang Yang (eds.), *Advances in Knowledge Discovery and Data Mining*, pp. 761–768, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

Baozhi Qiu and Xiaofeng Cao. Clustering boundary detection for high dimensional space based on space inversion and hopkins statistics. *Knowledge-Based Systems*, 98:216–225, 2016.

F. Rahmati, R.H. Gharaei, and H. Nezamabadi-pour. Ardod: adaptive radius density-based outlier detection. *Evolutionary Intelligence*, 17:3707–3722, 2024.

Leon Simon. *Lectures on Geometric Measure Theory*. Australian National University Press, 1 edition, 1983. URL https://openresearch-repository.anu.edu.au/items/7151bcb4-6ce1-4958-8576-049e09dc79b1.

A. Singer and H.-T. Wu. Vector diffusion maps and the connection laplacian. *Communications on Pure and Applied Mathematics*, 65(8):1067–1144, 2012. doi: https://doi.org/10.1002/cpa.21395.

Amit Singer. From graph to manifold Laplacian: the convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, 2006.

Michael Spivak. *A Comprehensive Introduction to Differential Geometry*. Publish or Perish, 1999.

Qingsong Tang, Mingzhi Yang, Ziyi Wang, Wenhao Dong, and Yang Liu. Boundary points guided 3d object detection for point clouds. *Applied Soft Computing*, 165:112117, 2024.

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

Melanie Weber. Geometric machine learning. *AI Magazine*, 46(1):e12210, 2025. doi: https://doi.org/10.1002/aaai.12210.

Chenyi Xia, W. Hsu, M.L. Lee, and B.C. Ooi. Border: efficient computation of boundary points. *IEEE Transactions on Knowledge and Data Engineering*, 18(3):289–303, 2006.