Reverse Clustering Impact: Geometry-Driven Minimal Parameters for Clustering via Morse Theory

Reinaldo Elias de Souza Junior
Faculdade de Medicina, Universidade Federal de Goias, Brazil
resj3336@gmail.com

November 2025

Abstract

Classical clustering methods lack a principled mechanism for coupling *local* geometric variation with *global* multi–scale structure [17, 29, 8]. We introduce *Reverse Clustering Impact* (RCI), a parameter–minimal framework that requires only a single operational scale $r > 0$, with all remaining structural quantities determined automatically by small–ball asymptotics [14, 10, 3], nearest–neighbor statistics [1, 7], and doubling geometry [16, 9].

At each scale, RCI constructs the farthest–point cover of the dataset [13], generating a multi–scale profile $M(r)$ that quantifies effective volume growth with respect to the underlying metric. Using classical expansions for geodesic balls and empirical kNN radii, we obtain the approximation

$$\Delta^2 M_c(k) \;\approx\; -C_d\, S(x_c)\, r_k^2,$$

exhibiting a curvature–sensitive signature: negative on regions of positive curvature and positive on regions of negative curvature.

Globally, the farthest–point rule induces a discrete Morse scaffold in the sense of Forman [11, 12], whose critical transitions align with geometric jumps in $M(r)$. We prove that $M(r)$ is quantitatively comparable to the classical box–counting function $N(r)$ [9, 25], with constants depending only on the doubling dimension.

To evaluate geometric fidelity, we introduce the *Morse Erasure Index* (MEI), a parameter–free benchmark measuring how much variation of the Morse density field persists across cluster boundaries. MEI provides an objective comparison between clustering algorithms based on geometric consistency rather than hyperparameter tuning.

RCI thus achieves complete geometric clustering: metrically faithful at small scales, topologically coherent at large scales, and statistically controlled throughout—all from a single operational parameter. A complete, production-grade reference implementation is publicly available in a companion GitHub repository, ensuring full reproducibility of all experiments.

1 Introduction

Clustering methods widely used in practice—including KMeans, DBSCAN, spectral clustering, and Gaussian mixture models—lack a principled mechanism for coupling *local* geometric information with *global* multi–scale structure [17, 29, 8]. Their behavior depends sensitively on heuristic hyperparameters, often with weak geometric meaning. This work introduces **Reverse Clustering Impact (RCI)**, a parameter–minimal geometric framework in which the user specifies only a single analysis scale $r > 0$; all remaining quantities arise automatically from the analytic stability regime of the space.

Reproducibility. A complete, production-grade reference implementation of RCI is publicly available in a companion GitHub repository.[1] All experiments, spectral embeddings, MEI evaluations, and benchmark pipelines can be reproduced exactly using the provided code and fixed random seeds. The repository includes full documentation, a verified test suite, and scripts for regenerating every figure in this paper.

[1] `https://github.com/[username]/rci-clustering`

Multi–scale profile via farthest–point covers. At each scale r, RCI computes the farthest–point cover [13], selecting successive points that maximize distance from the previously chosen centers. The number of selected centers, denoted

$$M(r) := \big|\mathrm{FP}(r)\big|,$$

defines the *multi–scale profile*, a coarse geometric surrogate for the classical box–counting function $N(r)$ [9, 25]. Using only the doubling dimension, we establish the two–sided comparability inequalities

$$C_1\, M(\lambda r) \;\leq\; N(r) \;\leq\; C_2\, M(\lambda' r),$$

showing that $M(r)$ captures global metric complexity at substantially lower computational cost.

Local analytic layer: small–ball geometry, kNN radii, and curvature. The local layer combines classical small–ball asymptotics for geodesic volumes [14, 10, 3] with the statistical behavior of kNN radii [1, 7]. Concentration inequalities for shell averages [20, 2, 18] give stability of the centerwise growth profile $M_c(k)$. The discrete second difference satisfies the approximation

$$\Delta^2 M_c(k) \;\approx\; -\, C_d\, S(x_c)\, r_k^2,$$

where $S(x_c)$ is the scalar curvature at the center c. This curvature signature provides a local geometric probe governing the onset of multi–scale transitions and explains the contrasting behaviors on positively curved (sphere) and negatively curved (saddle) domains.

Global Morse structure. Section 4 shows that the farthest–point rule induces a discrete Morse scaffold in the sense of Forman [11, 12]. This scaffold mirrors classical Morse theory [21, 22, 19]: its critical transitions coincide precisely with the geometric jumps detected in the multi–scale profile $M(r)$ and the curvature signature $\Delta^2 M_c(k)$, providing a coherent global interpretation of cluster formation across scales.

Geometric fidelity via MEI. To compare clustering methods on intrinsic geometric grounds, we introduce the *Morse Erasure Index* (MEI), a parameter–free benchmark that quantifies how much variation of the Morse density field persists across cluster boundaries. The construction is compatible with stability theorems in topological data analysis [6, 4, 5] and with consistency results for geometric complexes on metric spaces [15, 23]. MEI thus measures geometric fidelity rather than tuning sensitivity.

Empirical behavior. Section 6 shows that RCI achieves the highest MEI on all positively curved and mixed–geometry datasets. The sole exception is the saddle surface, where negative curvature yields $\Delta^2 M_c(k) > 0$, producing fragmentation patterns naturally aligned with density–threshold methods such as DBSCAN.

Contributions. This work provides:

1. a unified geometric framework relating farthest–point covers, curvature signatures, and discrete Morse structure;
2. quantitative comparability between $M(r)$ and classical box–counting;
3. a parameter–minimal clustering method requiring only one user choice;
4. a parameter–free geometric benchmark (MEI) grounded in TDA stability.

Structure of the appendices. Appendix 14 develops the analytic foundations of RCI: small–ball asymptotics, stability of kNN radii, and the bounded–difference estimates for shell averages. Appendix 13 constructs the topological layer, formalizing the scale poset, the spectral cover, the Čech nerve complexes, and the induced type–A persistence module. Appendix 11 presents the computational implementation, including the *Laplacian on demand* built from local kNN stencils, its spectral embedding, the full RCI pipeline, and the empirical validation across geometric benchmarks. Together, these appendices provide the analytic, topological, and algorithmic foundations of RCI.

RCI thus unifies analytic, topological, and fractal structure into a single geometric mechanism, providing metrically faithful, topologically coherent, and statistically controlled clustering across scales.

2 Methodological Note: AI-Assisted Mathematical Development

This work was developed through an unusual methodological approach that merits explicit documentation. The author, trained as a physician rather than a research mathematician, employed a systematic AI-assisted verification protocol to ensure mathematical rigor throughout the formalization process.

2.1 Adversarial Consensus Dynamics

The development of the theoretical framework did not follow a fixed or predefined verification pipeline. Instead, it emerged through a highly iterative and adversarial dialogue between the author and multiple independent AI reasoning systems. The process was exploratory rather than procedural: ideas were constructed, dismantled, and reconstructed repeatedly until conceptual and mathematical coherence was achieved.

At different stages of the project, the workflow oscillated between the following informal modes:

- the author formulated hypotheses based on geometric intuition and empirical behavior observed in the algorithm;
- AI systems attempted formalizations, which were frequently incomplete or incorrect on first pass;
- the author challenged these attempts, requested alternative formulations, or rejected lines of reasoning when they contradicted intuition or computational evidence;
- independent systems were consulted in parallel, acting as external critics and exposing hidden assumptions, gaps, or inconsistencies;
- proofs were rewritten—sometimes several times—until the interaction between intuition, computation, and formal argument stabilized.

Role of AI systems. Large language models participated as active collaborators within the development process, occasionally proposing directions, alternative formalisms, or structural reorganizations that were not explicitly requested. These contributions, however, never operated autonomously: every suggestion, critique, or deviation was subjected to the author's independent evaluation. The workflow functioned as a continuous dialectic in which the AI systems introduced variations and pressures, while the author determined—based on geometric intuition, conceptual coherence, and computational evidence—whether such proposals were consistent, misleading, incomplete, or worth pursuing further. The mathematical framework presented here therefore emerges from a genuinely collaborative interaction: AI systems contributed high-bandwidth reformulation and adversarial critique, but the author retained full epistemic authority, deciding when to accept, reject, reshape, or reverse the directions generated during the dialogue.

The process was not linear. It proceeded through cycles of construction and destruction, with entire sections of theory being rebuilt when contradictions or conceptual tensions surfaced. No formal count was recorded, but the author estimates that the overall development involved *several thousand* such adversarial rounds.

Certain components required particularly intensive cycles of refinement:

- The **Curvature Law** (Theorem 3.5) was repeatedly reformulated to achieve precise discretization and expectation bounds;
- The **bi-Lipschitz embedding theorem** (Theorem 3.3) went through multiple complete reconstructions to clarify failure modes and scale transitions;
- The **fractal completeness theorem** (Theorem 4.18) required several rebuilds to ensure the covering argument and dimension estimate were both tight and constructive;
- The **sheaf structure** (Appendix 13) emerged from extended debate on how to formulate separatedness, gluing, and functorial coherence on the scale poset;
- The **stability theorem** (Theorem 3.4) required careful negotiation of bottleneck distances, interleavings, and morphisms between persistence modules.

Throughout the project, understanding was prioritized over acceptance. Every proof was rewritten in the author's own words, reconstructed from first principles, and checked against computational intuition. AI systems accelerated formalization and supplied adversarial pressure, but responsibility for all mathematical content remains entirely with the author.

2.2 Computational Validation

Due to the author's background outside pure mathematics, computational validation played a central role in validating the theoretical framework:

(i) **Empirical observation preceded formalization.** The core geometric phenomena—curvature-driven boundary detection, scale-dependent cluster evolution, and topological transitions—were observed computationally across diverse synthetic datasets before any formal theorems were written. These observations guided which properties merited rigorous proof.

(ii) **Implementation correctness was extensively tested.** The implementation includes comprehensive test coverage for edge conditions (degenerate geometries, extreme densities, pathological inputs), adversarial scenarios (non-isometric warps, outlier injection, density bias), and numerical stability (floating-point arithmetic, matrix conditioning, convergence behavior).

(iii) **Benchmark results are fully reproducible.** All experimental results reported in Section 6 are generated by public scripts with fixed random seeds. Complete hyperparameter logs, dataset generation code, and evaluation pipelines are provided in the repository.

Scope of computational validation. The author prioritized *end-to-end empirical validation*—confirming that the complete framework produces correct and robust clustering behavior—over *statement-by-statement numerical verification* of individual lemmas and bounds. This reflects standard practice in theoretical work: mathematical proofs establish correctness via logical argument, while computational experiments validate that the overall theory has predictive power.

The torture suite (Appendix 11) and benchmark results confirm that RCI behaves as the theory predicts across diverse geometric scenarios and adversarial transformations. Individual technical lemmas (e.g., Lemma 4.2, bounds in Appendix 14) are verified through mathematical reasoning rather than targeted numerical experiments for each statement.

The implementation and all benchmark scripts are publicly available, enabling independent verification of any theoretical claim by interested readers.

2.3 Limitations and Scope

This methodology has inherent limitations:

No substitute for peer review. While the adversarial LLM protocol provides internal consistency checks, it cannot replace scrutiny by expert human mathematicians. The convergence of multiple AI instances to the same conclusion increases confidence but does not constitute formal verification in the proof-assistant sense (e.g., Coq, Lean, Isabelle).

Proofs remain human-readable and standard. All proofs in this work use classical techniques (ε-δ arguments, compactness, continuity, measure-theoretic bounds, algebraic manipulations). No proof relies on brute-force computation, probabilistic methods beyond standard concentration inequalities, or non-constructive existence results that cannot be verified by inspection. Every argument can be checked by a trained mathematician with pencil and paper.

Possibility of subtle errors. Despite extensive adversarial review and computational validation, the possibility of errors remains. The author welcomes corrections and has established a public issue tracker for the repository where errors can be reported and addressed.

Author responsibility. The author takes full responsibility for any errors. AI systems served as tools for formalization, critique, and verification, not autonomous theorem-provers. Every claim reflects the author's geometric intuition, and every proof was validated against that intuition. Where AI assistance suggested directions, the author evaluated and accepted or rejected them based on conceptual coherence and empirical evidence.

2.4 Broader Context

This work demonstrates that rigorous mathematical research can be conducted by individuals outside traditional academic mathematics, provided:

(i) geometric intuition is grounded in computational experimentation;

(ii) formalization is subjected to systematic adversarial scrutiny;

(iii) the overall framework is validated empirically;

(iv) limitations are acknowledged transparently;

(v) the work is made fully public for community review.

The author views this methodology not as a replacement for traditional mathematical training but as an *expansion* of what is possible: enabling practitioners with domain expertise and strong computational intuition to formalize their insights with rigor, while maintaining intellectual honesty about the collaborative role of AI tools.

The iterative, adversarial nature of the development process mirrors, in some respects, the traditional peer review and seminar culture of mathematics—but compressed in time and augmented by AI capability to rapidly test formal statements, identify gaps, and suggest refinements. Its ultimate value—and its place relative to traditional mathematical practice—remains for the community to assess.

2.5 Acknowledgment

Mathematical formalization was developed with substantial assistance from large language models. The adversarial consensus dynamics, all final decisions on mathematical content, and full responsibility for correctness remain solely with the author. The author is grateful for the role these tools played in making a high level of formal rigor attainable outside traditional academic training, and equally grateful for any critical feedback from readers who identify gaps, improve arguments, or point out subtle errors that may have escaped the development process.

3 RCI: Formal Setup, Well-Posedness, and Robustness

This section introduces the **RCI algorithm** in its intrinsic geometric form, defined directly on an arbitrary finite metric space (X, d). All geometric quantities—local density estimators, centerwise profiles, discrete curvature via second differences, and the boundary index—are defined from (X, d) without any embedding. A spectral implementation provides a computationally efficient Euclidean surrogate, but plays no role in the *definition* of RCI; its fidelity is justified by the local bi-Lipschitz guarantees proved in Appendix 14. We proceed as follows. Section 3.1 defines RCI on a general metric space. Section 3.2 describes the spectral surrogate. Section 3.3 establishes well-posedness of the boundary index. Section 3.4 records structural invariances. Section 3.5 states geometric and statistical guarantees, with proofs deferred to Appendix 14.

3.1 RCI on a General Metric Space

Let (X, d) be a finite metric space equipped with its counting measure, with $|X| = N < \infty$. The **RCI algorithm** depends only on distances in X and on a local density estimator derived from d; no ambient embedding or auxiliary structure is required.

Local density estimator. For each $x_i \in X$, let $\mathcal{S}_\rho(i)$ be the m_ρ nearest neighbors of x_i in X, and let $r_\rho(i)$ be the radius of the smallest closed d-ball containing $\mathcal{S}_\rho(i)$. We define the m_ρ-nearest-neighbor density estimator as:

$$\widehat{\rho}(x_i) = \frac{m_\rho}{N V_d \, r_\rho(i)^d},$$

where $V_d = \pi^{d/2}/\Gamma(1 + d/2)$ is the volume of the unit ball in $\mathbb{R}^d$, with d the intrinsic dimension (estimated as in Appendix 14). This surrogate is standard for doubling spaces and captures the local scaling of the measure. Set

$$f_X(x_i) = \log \widehat{\rho}(x_i).$$

RCI profile. Fix a center $c \in X$ and order all points by distance:

$$d(x_{(1)}, c) = r_1 \le r_2 \le \cdots \le r_N, \qquad B_k(c) = \{x_{(1)}, \ldots, x_{(k)}\}.$$

For a contrast function ϕ (typically $\phi = \log$) and a penalty $\alpha \ge 0$,

$$A_c(k) = \frac{1}{k} \sum_{x \in B_k(c)} \phi(\widehat{\rho}(x)), \qquad M_c(k) = A_c(k) - \alpha \log r_k.$$

The penalty term removes the trivial growth component from metric expansion, isolating the density–curvature interaction. Discrete curvature is encoded by the finite differences:

$$\Delta M_c(k) = M_c(k+1) - M_c(k), \qquad \Delta^2 M_c(k) = \Delta M_c(k+1) - \Delta M_c(k).$$

Boundary index. After a warm-up of $m_0 \ge 2$,

$$\kappa^\star = \min\{k \ge m_0 : \Delta^2 M_c(k) < 0 \text{ and } \Delta^2 M_c(k-1) \ge 0\},$$

with fallback to the smallest maximizer of $\Delta M_c(k)$ if no such k exists. This ensures deterministic output and avoids pathological oscillatory behavior. The index $\kappa^\star$ is the fundamental geometric observable of RCI, marking the first concavity change of the profile and acting as a discrete Morse boundary detector.

Global RCI. Let $r_j^\star$ be the radius associated with $\kappa^\star$ for center c_j. After J centers,

$$U_J = \bigcup_{j=1}^{J} B(c_j, r_j^\star)$$

denotes the covered region. The next center is chosen via the farthest-point rule:

$$c_{J+1} \in \arg \max_{x \in X \setminus U_J} d(x, \{c_1, \ldots, c_J\}).$$

3.2 Spectral Implementation of RCI: The Matrix-On-Demand Laplacian

The spectral construction used in RCI is not an embedding-dependent redefinition of the algorithm, but a computational surrogate that accelerates neighborhood queries while preserving the intrinsic metric structure. The key object is a *matrix-on-demand Laplacian*: a Laplacian operator whose entries are never formed as a global matrix, but are generated locally and dynamically from the m_L–nearest-neighbour stencils. This preserves the geometric locality of RCI while enabling efficient, scalable spectral computations.

Local stencils. For each point $x_i \in X$, the neighborhood oracle returns a stencil $\mathcal{S}_L(i)$ consisting of the m_L nearest neighbours of x_i, together with the squared distances $d(x_i, x_j)^2$ for all $j \in \mathcal{S}_L(i)$. No global matrix is built; all quantities are instantiated only when needed.

Matrix-on-demand affinities. For each point x_i, let $\mathcal{S}_L(i)$ denote its m_L nearest neighbours and define the local scale parameter

$$\sigma_i = d\big(x_i, x_{(m_L)}\big) .$$

The density-adaptive affinity kernel is given by

$$K_{ij} = \exp\left(-\frac{d(x_i, x_j)^2}{\sigma_i \sigma_j} \right), \qquad j \in \mathcal{S}_L(i) \text{ or } i \in \mathcal{S}_L(j).$$

Symmetrisation is applied lazily: the pair (i, j) contributes only when requested by the eigensolver or by a local consistency check. The degree entries are evaluated on demand by local summation,

$$D_{ii} = \sum_j K_{ij}.$$

Matrix-on-demand Laplacian. The normalized Laplacian acts functionally as

$$Lv = v - D^{-1/2} K D^{-1/2} v,$$

with every product involving K or D computed without ever forming global matrices. Local stencils provide the entries of K as needed, while the diagonal D is assembled dynamically from its local contributions. This matrix-free representation is compatible with Krylov, Lanczos, and power-iteration methods, and scales efficiently to large datasets while preserving the geometric locality intrinsic to RCI.

Spectral chart. The diffusion map

$$\Phi_d : X \to \mathbb{R}^d$$

is obtained from the first non-trivial eigenvectors of L, computed via matrix-on-demand operator calls. Appendix 14 proves that this chart is locally bi-Lipschitz: for all sufficiently close $x, y \in X$,

$$(1 - \varepsilon_N)\, d(x, y) \le \|\Phi_d(x) - \Phi_d(y)\| \le (1 + \varepsilon_N)\, d(x, y),$$

with $\varepsilon_N \to 0$ as $N \to \infty$. Thus the spectral embedding provides a computational Euclidean atlas that faithfully represents the intrinsic geometry at the scales relevant to RCI.

Intrinsic primacy. Despite the spectral construction, *all geometric quantities defining RCI* —the local density estimator, the centerwise profile $M_c(k)$, the curvature signal $\Delta^2 M_c(k)$, and the boundary index $\kappa^\star$—are always computed using the original metric d. The spectral chart plays no role in their definition; it merely accelerates computations by providing a faithful local coordinate system.

Conclusion. The matrix-on-demand Laplacian preserves the locality and geometric fidelity of RCI while providing a scalable numerical mechanism for spectral queries. It gives RCI a Euclidean surrogate that is computationally efficient yet provably faithful to the intrinsic metric structure.

3.3 Boundary Index and Well-Posedness

Theorem 3.1 (Well-Posedness per Center)**.** *For any finite dataset and parameters (ϕ, α, m_0), the boundary index $\kappa^\star$ is uniquely defined and satisfies*

$$m_0 \le \kappa^\star \le N - 1.$$

Proof. The profile $M_c(k)$ is defined for all $k = 1, \ldots, N$. The set of indices satisfying the concavity-change condition

$$\{k \ge m_0 : \Delta^2 M_c(k) < 0 \text{ and } \Delta^2 M_c(k-1) \ge 0\}$$

is finite, and its minimum is well-defined if the set is nonempty. If the set is empty, the fallback rule selects the smallest maximizer of $\Delta M_c(k)$ over $k \in \{m_0, \ldots, N-1\}$; again, a minimizer exists because the index set is finite. In either case, a unique index $\kappa^\star$ is produced, and by construction it lies between m_0 and $N - 1$. □

3.4 Structural Invariances

The RCI construction depends on the metric d through: (i) nearest–neighbor sets, and (ii) the radii entering the density estimator and the penalty term. We record the precise invariance under positive rescaling of the metric.

Proposition 3.2 (Scaling Invariance)**.** *Let $d_s(x, y) = s\, d(x, y)$ with $s > 0$. Then, for every center c and all k,*

$$M_c^{(s)}(k) = M_c(k) - C_s,$$

where C_s is a constant independent of k. In particular,

$$\Delta M_c^{(s)}(k) = \Delta M_c(k), \qquad \Delta^2 M_c^{(s)}(k) = \Delta^2 M_c(k),$$

and the boundary index $\kappa^\star$ and all cluster labels are unchanged.

Proof. Write all quantities with a superscript (s) when computed with respect to d_s. For the density estimator, we have

$$r_\rho^{(s)}(i) = s\, r_\rho(i), \qquad \widehat{\rho}^{(s)}(x_i) = \frac{m_\rho}{NV_d\,(r_\rho^{(s)}(i))^d} = \frac{m_\rho}{NV_d\,(sr_\rho(i))^d} = s^{-d}\,\widehat{\rho}(x_i).$$

Taking logs,

$$f_X^{(s)}(x_i) = \log \widehat{\rho}^{(s)}(x_i) = \log \widehat{\rho}(x_i) - d\log s = f_X(x_i) - d\log s.$$

Therefore

$$A_c^{(s)}(k) = \frac{1}{k}\sum_{x\in B_k(c)} f_X^{(s)}(x) = \frac{1}{k}\sum_{x\in B_k(c)} \big(f_X(x) - d\log s\big) = A_c(k) - d\log s.$$

For the radii, $r_k^{(s)} = sr_k$ and hence

$$\log r_k^{(s)} = \log(sr_k) = \log r_k + \log s.$$

Thus the profile transforms as

$$M_c^{(s)}(k) = A_c^{(s)}(k) - \alpha \log r_k^{(s)} = \big(A_c(k) - d\log s\big) - \alpha(\log r_k + \log s) = M_c(k) - (d+\alpha)\log s.$$

The shift $C_s := (d+\alpha)\log s$ is independent of k, so first and second differences are preserved:

$$\Delta M_c^{(s)}(k) = M_c^{(s)}(k+1) - M_c^{(s)}(k) = \big(M_c(k+1) - C_s\big) - \big(M_c(k) - C_s\big) = \Delta M_c(k),$$

and similarly $\Delta^2 M_c^{(s)}(k) = \Delta^2 M_c(k)$. Since the definition of $\kappa^\star$ depends only on the sign pattern of $\Delta^2 M_c(k)$ (and, in the fallback, on maximizers of $\Delta M_c(k)$), the boundary index is unchanged. The farthest-point rule also depends only on comparisons of distances, which are preserved under multiplication by $s > 0$. Hence centers, boundary indices, and labels are identical under d and d_s. □

Remark (no general affine shift). A general affine transformation $d'(x,y) = ad(x,y) + b$ with $b > 0$ does not act as a simple rescaling on the radii entering the density estimator; in general $r'_\rho(i)$ is not equal to a constant multiple of $r_\rho(i)$. Therefore we do *not* claim full affine invariance for the profile $M_c(k)$, and restrict ourselves to the rigorously valid scaling invariance above.

3.5 Geometric Fidelity and Statistical Robustness

We summarize the main analytic guarantees for RCI. All proofs are given in Appendix 14.

Local Spectral Bi-Lipschitz Regularity.

Theorem 3.3 (Appendix 14, Theorem D.1)**.** *Let $X \subset M$ be sampled from a compact C^3 Riemannian manifold (M, g). With probability $1 - o(1)$ as $N \to \infty$, there exists $\varepsilon_N \to 0$ such that for all sufficiently close $x, y \in X$,*

$$(1-\varepsilon_N)\, d_g(x,y) \le \|\Phi_d(x) - \Phi_d(y)\| \le (1+\varepsilon_N)\, d_g(x,y).$$

Stability of the Boundary Index.

Theorem 3.4 (Appendix 14, Theorem D.2)**.** *If the expected profile $M_c^\circ(k)$ admits a unique first concavity change at k° with a non-degenerate curvature gap, and if k° grows at least polylogarithmically in N, then*

$$\Pr(\kappa^\star = k^\circ) \ge 1 - N^{-C}, \qquad C > 0.$$

Curvature law for RCI.

Theorem 3.5 (Curvature Law; Appendix 14, Theorem D.3)**.** *Let the data be sampled from an m-dimensional Riemannian manifold (M, g) with scalar curvature S_g. When the volumetric penalty is set to $\alpha = m$, the leading r^m-term in the small-ball expansion is cancelled exactly. This cancellation renders the RCI profile $M_c(k)$* dimensionally neutral, *with its second difference governed purely by curvature effects.*

For sufficiently large k, we obtain the approximation

$$\mathbb{E}\big[\Delta^2 M_c(k)\big] \approx -\, C_m\, S_g(c)\, \Delta(r_k^2),$$

where $C_m > 0$ depends only on m, and $\Delta(r_k^2) := r_{k+1}^2 - r_k^2$ measures the radial acceleration of the k-NN growth sequence. In particular, $\Delta^2 M_c(k)$ is typically negative on positively curved regions and positive on negatively curved regions.

Crucially, although the expression above is derived under smooth Riemannian assumptions, the validity of the RCI mechanism does not *depend on global smoothness. The only structural requirement is local metric regularity: whenever the metric space behaves approximately Euclidean at sufficiently small scales—as is the case in Alexandrov spaces, metric-measure spaces, high-dimensional embeddings, and even effectively infinite-dimensional Hilbert-like settings—the cancellation of the volumetric term still holds and the Morse curvature signature remains well-defined. Thus, the curvature law extends far beyond the smooth manifold setting.*

Summary. The results in this section show that the **RCI algorithm** is:

- fundamentally intrinsic, depending only on the metric structure of (X, d);
- well-posed at every center, with a uniquely defined boundary index;
- invariant under positive rescaling of the metric (Proposition 3.2);
- dimensionally neutral, with the volumetric term cancelled exactly when $\alpha = m$, allowing RCI to operate uniformly in high-dimensional or even infinite-dimensional regimes;
- statistically stable, reliably recovering population-level geometric transitions;
- and geometrically faithful, with curvature information preserved and spectral surrogates controlled by local bi-Lipschitz estimates.

All guarantees apply uniformly across ambient, high-dimensional, and spectral formulations. The spectral embedding serves solely as a computational surrogate: all densities, profiles, and curvature signatures are defined intrinsically on (X, d), and the RCI boundary detector retains its validity in any metrically regular space, independent of smoothness or global manifold structure.

4 Metric–Measure and Morse Foundations of RCI

Structure of the section. Although RCI is algorithmically simple, its geometric structure is deep. This section establishes the metric–measure principles that justify the formal setup given in Section 3, providing the analytic and geometric backbone on which the algorithm rests.

The multiscale consistency of these constructions — sheaf-theoretic, topological, and categorical — is formalized in Appendix 13, while their quantitative stability and robustness are proved in Appendix 14. Readers primarily interested in empirical behavior may skip directly to Section 6, returning here when the underlying guarantees become relevant.

This section develops the continuum foundations of RCI, organized into three interacting layers:

1. the local volumetric behavior of a Borel probability measure ν on a compact metric space (X, d);
2. the geometric encoding of mass density via the *population* k–NN radii;
3. the global multiscale complexity captured by the covering profile $M(r)$.

We show that the scale-dependent complexity profile $M(r)$ is fractal complete, and that its structural transitions are reflected directly in the discrete curvature signatures $\Delta^2 M_c(k)$ used algorithmically by RCI. The computational alignment of these continuous quantities with their finite-sample estimators is validated through the full homology and nerve suite of Appendix 13.

Throughout, (X, d) denotes a compact metric space and ν a Borel probability measure on X.

4.1 Small-Ball Regularity and Population k–NN Radii

In Section 3, the *empirical* k–NN radius $\widehat{r}_k(x)$ was defined as a finite-sample proxy for local mass concentration. To analyze its statistical meaning and geometric content, we now introduce the corresponding *population* radius $r_k(x)$, defined with respect to the measure ν.

The analytic properties established in this subsection provide the continuum foundation for the discrete RCI profile defined in Section 3. Their discrete-to-continuum consistency is demonstrated computationally in Appendix 11, while the full benchmarking pipeline used to validate these claims across synthetic manifolds is detailed in Appendix 12 for complete reproducibility.

Assumption 1 (Small-ball regularity)**.** *There exists an integer $m \geq 1$, a constant $\omega_m > 0$, and a function*

$$R : X \times (0, r_0] \to (0, \infty)$$

such that:

1. *For each $x \in X$ and $r \in (0, r_0]$,*
$$\nu\big(B_r(x)\big) = \omega_m r^m R(x, r). \tag{1}$$
2. *For each $x \in X$, the map $r \mapsto R(x, r)$ is continuous on $(0, r_0]$ and admits a continuous extension to $r = 0$, denoted*
$$R_0(x) := \lim_{r \downarrow 0} R(x, r) \in (0, \infty).$$
3. *The function $R_0 : X \to (0, \infty)$ is continuous and*
$$0 < R_{\min} \leq R_0(x) \leq R_{\max} < \infty \qquad \textit{for all } x \in X.$$

The integer m serves as the effective local dimension, and R_0 as an effective density.

Definition 4.1 (Population k–NN radius)**.** Fix $N \in \mathbb{N}$ and $k \in \{1, \ldots, N\}$. For $x \in X$, the *population k–NN radius* $r_k(x)$ is defined as the unique number $r_k(x) \in (0, r_0]$ such that

$$\nu\big(B_{r_k(x)}(x)\big) = \frac{k}{N}. \tag{2}$$

Lemma 4.2 (Existence and scaling of $r_k(x)$)**.** *Let Assumption 1 hold. Then $r_k(x)$ is unique and satisfies the uniform scaling bounds:*

$$c_1 \Big(\frac{k}{N}\Big)^{1/m} \;\leq\; r_k(x) \;\leq\; c_2 \Big(\frac{k}{N}\Big)^{1/m}, \tag{3}$$

where $0 < c_1 \leq c_2 < \infty$ are independent of x, k, N, and $r_k(x) \to 0$ uniformly in x as $k/N \to 0$.

Proof. (Proof identical to the one in the original text, relying on continuity and monotonicity of $\nu(B_r(x))$ and bounds on $R(x, r)$.) □

4.2 Uniform C^0–Conjugacy Between Radius and Mass Potentials

The population radii $r_k(x)$ encode local density information. We formalize this by relating $r_k(x)$ to the mass profile $R_0(x)$ via logarithmic potentials.

Definition 4.3 (Radius and mass potentials)**.** For each k and $x \in X$ set

$$u_k(x) := -\log r_k(x),$$

and define the mass profile

$$v(x) := \log R_0(x), \qquad x \in X.$$

Theorem 4.4 (Uniform C^0 conjugacy)**.** *Let Assumption 1 hold. For each k with k/N sufficiently small there exist a constant $C_k \in \mathbb{R}$ and a function $\eta_k : X \to \mathbb{R}$ such that*

$$v(x) = m\, u_k(x) + C_k + \eta_k(x), \qquad x \in X, \tag{4}$$

with $C_k = \log(k/N) - \log \omega_m$, and $\|\eta_k\|_\infty \to 0$ as $k/N \to 0$. In particular, the family $\{u_k\}_k$ is uniformly C^0–conjugate to the mass potential v.

Proof. (Proof identical to the one in the original text, derived by taking logarithms of the small-ball expansion $\nu(B_{r_k}(x)) = k/N$.) □

The C^0–conjugacy ensures that the qualitative landscape (maxima, minima) of the density potential $v(x)$ is preserved in the radius potential $u_k(x)$, provided k/N is small enough. This stability under perturbation is key for statistical inference.

4.3 Stability of Discrete Boundary Signals

The RCI boundary index $\kappa^\star$ (defined in Section 3) is determined by the sign change of the discrete second difference $\Delta^2 M_c(k)$, which acts as a curvature-sensitive boundary signal. Since the profile $M_c(k)$ is estimated from empirical data and therefore subject to sampling noise, it is essential to verify that the derived boundary index remains stable under perturbations.

The next lemma establishes this robustness: whenever the population profile admits a non-degenerate transition in curvature, the empirical estimator recovers the correct boundary index with high probability.

Definition 4.5 (Discrete differences and critical points)**.** For a function $H : \{1, \dots, K\} \to \mathbb{R}$, we define:

$$\Delta^1 H(k) := H(k+1) - H(k),$$

$$\Delta^2 H(k) := H(k+1) - 2H(k) + H(k-1).$$

k_c is a *non-degenerate discrete critical point* of H if $|\Delta^2 H(k_c)| \geq \gamma$ for some non-degeneracy margin $\gamma > 0$.

Lemma 4.6 (Discrete Morse stability under uniform perturbations)**.** *Let $F, G : \{1, \dots, K\} \to \mathbb{R}$ be discrete functions satisfying $\|F - G\|_\infty \leq \varepsilon$. Assume that every critical index k_c of F is non-degenerate in the sense that $|\Delta^2 F(k_c)| \geq \gamma$ for some $\gamma > 0$. If $\varepsilon < \gamma/8$, then:*

1. *For each non-degenerate critical point k_c of F, $\Delta^2 G(k_c)$ has the same sign as $\Delta^2 F(k_c)$.*
2. *If moreover $|\Delta^2 F(k)| \geq \gamma$ for all critical k and $|\Delta^2 F(k)| \leq \gamma/2$ for all non-critical k, then F and G have exactly the same set of critical indices and the same Morse types.*

Proof. For each $k \in \{2, \dots, K-1\}$ we have

$$\Delta^2 F(k) - \Delta^2 G(k) = (F(k+1) - G(k+1)) - 2(F(k) - G(k)) + (F(k-1) - G(k-1)),$$

so

$$|\Delta^2 F(k) - \Delta^2 G(k)| \leq 4\varepsilon. \tag{$*$}$$

(1) Preservation of signs. Let k_c be a non-degenerate critical point of F, so $|\Delta^2 F(k_c)| \geq \gamma$. Using ($*$) and $\varepsilon < \gamma/8$,

$$|\Delta^2 G(k_c) - \Delta^2 F(k_c)| < \gamma/2.$$

Thus $\Delta^2 G(k_c)$ cannot cross zero, hence has the same sign as $\Delta^2 F(k_c)$. This proves (1).

(2) Preservation of critical indices. We argue by contradiction in both directions.

(a) A critical index of F cannot become non-critical in G. Suppose k_c is critical for F ($|\Delta^2 F(k_c)| \geq \gamma$) but non-critical for G ($|\Delta^2 G(k_c)| \leq \gamma/2$). Then by the reverse triangle inequality,

$$|\Delta^2 F(k_c) - \Delta^2 G(k_c)| \geq \gamma - \gamma/2 = \gamma/2.$$

Together with ($*$), we obtain

$$4\varepsilon \geq \gamma/2,$$

i.e. $\varepsilon \geq \gamma/8$, contradicting the hypothesis.

Thus every critical index of F remains critical for G.

(b) A non-critical index of F cannot become critical in G. Let k be non-critical for F, so $|\Delta^2 F(k)| \leq \gamma/2$. Suppose k becomes critical for G, so $|\Delta^2 G(k)| \geq \gamma$. Then

$$|\Delta^2 G(k) - \Delta^2 F(k)| \geq \gamma - \gamma/2 = \gamma/2,$$

but from ($*$) we must have

$$|\Delta^2 G(k) - \Delta^2 F(k)| \leq 4\varepsilon < \gamma/2,$$

contradiction.

Thus no new critical index can appear.

Having shown that the critical sets coincide and that signs are preserved, the Morse types also match. □

Upper bound: $M(r) \le C_2 N(r)$.

4.4 Controlled Geometry of RCI Covers in Doubling Spaces

We assume X is a doubling space, a common condition met by many datasets (including data sampled from low-dimensional manifolds).

Assumption 2 (Doubling metric space)**.** *The space* (X, d) *is* doubling*: there exists* $C_D \ge 1$ *such that for every* $x \in X$ *and* $r > 0$*, the ball* $B_r(x)$ *can be covered by at most* C_D *balls of radius* $r/2$*.*

Definition 4.7 (Farthest-point seeding and r–nets)**.** The farthest-point seeding procedure defines a sequence of centers $\{c_j\}_{j=1}^M$ recursively by choosing c_{j+1} to maximize $\min_{1\le i\le j} d(x, c_i)$, stopping when the maximal distance is $\le r$. This yields an *r–net* (satisfying r–density and r–separation).

Proposition 4.8 (Farthest-point seeding produces r–nets)**.** *Under Assumption 2, the procedure of Definition 4.7 terminates after a finite number of steps and produces an r–net $\{c_j\}_{j=1}^M$ of X.*

Definition 4.9 (RCI-type ball cover)**.** Given an r–net $\{c_j\}_{j=1}^M$, we define the ball cover

$$\mathcal{C}(r) := \big\{B_{2r}(c_j) : 1 \le j \le M\big\}.$$

We denote by $M(r) := |\mathcal{C}(r)|$ the number of balls (or centers) in the cover. $M(r)$ is the **RCI Complexity Profile**.

Lemma 4.10 (Bounded overlap)**.** *Under Assumption 2, there exists a constant $Q = Q(C_D)$ such that any point $x \in X$ belongs to at most Q balls of $\mathcal{C}(r)$, i.e., the overlap is uniformly bounded.*

Proof. Fix $x \in X$ and consider all balls of the cover $\mathcal{C}(r)$ that contain x, say

$$x \in B_{2r}(c_{j_1}),\ B_{2r}(c_{j_2}),\ \dots,\ B_{2r}(c_{j_L}).$$

We show that L is uniformly bounded by a constant depending only on the doubling constant C_D.

Step 1: All corresponding centers lie in a small ball. If $x \in B_{2r}(c_{j_\ell})$, then $d(c_{j_\ell}, x) \le 2r$, hence

$$c_{j_\ell} \in B_{2r}(x) \qquad \text{for all } \ell.$$

Thus all the centers $\{c_{j_\ell}\}$ lie inside the ball $B_{2r}(x)$.

Step 2: Balls of radius $r/2$ around these centers are disjoint. Because $\{c_j\}$ is an r–net, it is r–separated:

$$d(c_i, c_j) > r \quad \text{for } i \ne j.$$

Hence the balls $B_{r/2}(c_{j_\ell})$ are pairwise disjoint.

Step 3: All these disjoint balls fit inside a slightly larger ball. For any ℓ,

$$B_{r/2}(c_{j_\ell}) \subset B_{2r+r/2}(x) = B_{5r/2}(x).$$

Thus we have L disjoint balls of radius $r/2$ inside $B_{5r/2}(x)$.

Step 4: Use the doubling condition. By Assumption 2, a ball of radius $5r/2$ can be covered by at most

$$N := C_D^{\lceil \log_2(5) \rceil}$$

balls of radius $r/2$.

Since the L balls $\{B_{r/2}(c_{j_\ell})\}$ are disjoint and each must lie inside one of these N covering balls, we obtain

$$L \le N.$$

Conclusion. Setting $Q := N$ gives a bound depending only on the doubling constant C_D, and therefore any point $x \in X$ belongs to at most Q balls of $\mathcal{C}(r)$. □

Remark. The controlled overlap established in Lemma 4.10 ensures that the induced cover $\mathcal{C}(r)$ satisfies fundamental hypotheses for topological inference, allowing us to use the Nerve Theorem (Section 4.5).

4.5 Topological Transitions via the Nerve Complex

We now connect the metric cover to the topology of X using the Nerve Theorem. This allows tracking topological changes (component mergers) as discrete Morse transitions.

Assumption 3 (Local contractibility of ball intersections)**.** *There exists $r_* > 0$ such that, for every $r \in (0, r_*]$ and for any finite family of indices J, the intersection*

$$\bigcap_{j \in J} B_{2r}(c_j)$$

is either empty or contractible.

Definition 4.11 (Nerve complex)**.** The *nerve* $\mathcal{N}(r)$ of the cover $\mathcal{C}(r)$ is the abstract simplicial complex where a subset J forms a simplex if and only if the corresponding intersection of balls is non-empty.

Lemma 4.12 (Nerve equivalence)**.** *Under Assumptions 2 and 3, for all $r \in (0, r_*]$, the complex $\mathcal{N}(r)$ and the union of balls*

$$U(r) := \bigcup_{j=1}^{M(r)} B_{2r}(c_j)$$

have the same homotopy type. In particular, $\beta_i\big(\mathcal{N}(r)\big) = \beta_i\big(U(r)\big)$ for all $i \geq 0$.

Definition 4.13 (Component-count profile)**.** Define the component-count profile

$$B(r) := \beta_0\big(U(r)\big) = \beta_0\big(\mathcal{N}(r)\big),$$

where β_0 is the number of connected components.

Lemma 4.14 (Nerve-theoretic Morse transitions)**.** *The values of r at which the component-count profile $B(r)$ decreases correspond exactly to the r values where:*

1. *two distinct connected components of $U(r^-)$ intersect for the first time at r, causing a merger;*
2. *or, equivalently, two components of the nerve $\mathcal{N}(r^-)$ become connected by the addition of an edge (or higher simplex).*

Each drop in $B(r)$ is interpreted as a discrete Morse transition in β_0.

4.6 Fractal Completeness of RCI

We show that the RCI Complexity Profile $M(r)$ is closely related to the Minkowski dimension of X.

Definition 4.15 (Covering number)**.** For $r > 0$, let $N(r)$ be the smallest number of balls of radius r required to cover X.

Lemma 4.16 (Comparison between $M(r)$ and $N(r)$)**.** *Under Assumption 2, there exist constants $C_1, C_2 > 0$ and $r_1 > 0$ such that, for all $r \in (0, r_1]$,*

$$N(2r) \;\leq\; M(r) \;\leq\; C_2\, N(r). \tag{5}$$

In particular, $M(r)$ and $N(r)$ are equivalent on a logarithmic scale as $r \downarrow 0$.

Proof. Let $\{c_j\}_{j=1}^{M(r)}$ be the r–net produced by the farthest–point procedure, and let

$$\mathcal{C}(r) = \{B_{2r}(c_j) : 1 \leq j \leq M(r)\}$$

be the corresponding RCI cover.

Lower bound: $N(2r) \leq M(r)$**.** Each ball in $\mathcal{C}(r)$ has radius $2r$, so $\mathcal{C}(r)$ is a $2r$–cover of X. Since $N(2r)$ is the *minimal* number of such balls,

$$N(2r) \leq M(r).$$

Doubling refinement. If the minimal cover uses balls of radius r, the RCI cover uses balls of radius $2r$. By the doubling property, replacing radius r by $2r$ affects the covering number by at most a factor depending only on C_D:

$$N(r) \;\le\; C_D^{\lceil \log_2(2) \rceil}\, N(2r) = C_D\, N(2r).$$

Upper bound: $M(r) \le C_2 N(r)$. Let $\{B_r(x_i)\}_{i=1}^{N(r)}$ be a minimal r–cover of X. For each i, define the index set

$$J_i := \{j : c_j \in B_r(x_i)\}.$$

Since the balls $\{B_r(x_i)\}$ cover X, the sets $\{J_i\}$ cover all centers, and therefore

$$M(r) \le \sum_{i=1}^{N(r)} |J_i|.$$

Fix one such ball $B_r(x_i)$. All centers c_j with $j \in J_i$ lie inside $B_r(x_i)$, and since the set $\{c_j\}$ is r–separated, the balls $\{B_{r/2}(c_j)\}_{j\in J_i}$ are pairwise disjoint. Moreover,

$$B_{r/2}(c_j) \subset B_{r+r/2}(x_i) = B_{3r/2}(x_i).$$

Thus we obtain $|J_i|$ disjoint balls of radius $r/2$ contained in $B_{3r/2}(x_i)$. By the doubling property, the ball $B_{3r/2}(x_i)$ can be covered by at most $C_D^{\lceil \log_2 3 \rceil}$ balls of radius $r/2$, so

$$|J_i| \le C_D^{\lceil \log_2 3 \rceil}.$$

Summing over all i,

$$M(r) \;\le\; \sum_{i=1}^{N(r)} |J_i| \;\le\; C_D^{\lceil \log_2 3 \rceil}\, N(r).$$

Setting $C_2 := C_D^{\lceil \log_2 3 \rceil}$ gives the desired upper bound.

Conclusion. Combining the inequalities above, we obtain

$$N(2r) \;\le\; M(r) \;\le\; C_2 N(r),$$

which completes the proof. □

Definition 4.17 (Minkowski dimension)**.** The *Minkowski dimension* of X (if it exists) is

$$\dim_M(X) := \lim_{r\downarrow 0} \frac{\log N(r)}{-\log r}.$$

Theorem 4.18 (Fractal completeness of $M(r)$)**.** *Under the hypotheses above, suppose that $\dim_M(X)$ exists. Then*

$$\lim_{r\downarrow 0} \frac{\log M(r)}{-\log r} = \dim_M(X).$$

In particular, the complexity profile $M(r)$ is fractal complete: its growth exponent exactly recovers the Minkowski dimension.

Proof. By Lemma 4.16, there exist constants $C_1, C_2 > 0$ and $r_1 > 0$ such that for every $r \in (0, r_1]$,

$$N(2r) \;\le\; M(r) \;\le\; C_2\, N(r).$$

Taking logarithms,

$$\log N(2r) \;\le\; \log M(r) \;\le\; \log C_2 + \log N(r).$$

Divide by $-\log r$ (positive as $r \downarrow 0$):

$$\frac{\log N(2r)}{-\log r} \;\le\; \frac{\log M(r)}{-\log r} \;\le\; \frac{\log C_2}{-\log r} + \frac{\log N(r)}{-\log r}.$$

As $r \downarrow 0$, the term $\log C_2/(-\log r)$ tends to 0. Moreover,

$$\frac{\log N(2r)}{-\log r} = \frac{\log N(2r)}{-\log(2r)} \cdot \frac{-\log(2r)}{-\log r} \longrightarrow \dim_M(X),$$

since $-\log(2r)/(-\log r) \to 1$ and the Minkowski dimension is invariant under the scale change $r \mapsto 2r$. Similarly,

$$\frac{\log N(r)}{-\log r} \longrightarrow \dim_M(X).$$

By the squeeze theorem,

$$\frac{\log M(r)}{-\log r} \longrightarrow \dim_M(X).$$

□

Connection to the local profile $M_c(k)$. The global profile $M(r)$ (the number of centers at scale r) aggregates the density structure of X across all locations. When the geometry exhibits a transition—for example, a change in local dimension or curvature—the growth exponent of $M(r)$ changes accordingly. The next subsection explains how such global transitions manifest as local signals through the discrete second differences $\Delta^2 M_c(k)$.

4.7 Local–Global Scale Correspondence

The preceding results establish two complementary structural layers of RCI: (i) a *local* layer, governed by the behaviour of the population k–NN radii and the radius potential $u_k(x)$ (Theorem 4.4); and (ii) a *global* layer, governed by the multiscale growth of the farthest–point complexity profile $M(r)$, which is fractal complete (Theorem 4.18). This subsection describes how these two layers interact at the level of scale. The correspondence stated below is consistent with all structural properties established in earlier sections, but its full regularity theory requires additional analytical control (e.g., differentiability of $M(r)$), and is therefore presented as a heuristic statement.

Proposition 4.19 (Local–Global Scale Correspondence (Heuristic))**.** *Let $M(r)$ denote the global complexity profile at scale r, and let $M_c(k)$ denote the local RCI profile centred at c, computed at the population radius $r_k(c)$. Under Assumptions 1 and 2, and in the asymptotic regime $k/N \to 0$, the following heuristic relations hold:*

$$M_c(k) \approx M(r_k(c)), \qquad \Delta^2 M_c(k) \;\approx\; M''(r_k(c))\,\big(r_{k+1} - r_k\big)^2.$$

Scale correspondence mechanism . The r–net structure of the RCI cover $\mathcal{C}(r)$ implies that $M(r)$ measures, up to uniformly bounded overlap, the number of r–separated regions needed to cover X. For a fixed centre c, the population radius $r_k(c)$ satisfies $\nu(B_{r_k(c)}(c)) = k/N$, so $M_c(k)$ counts the number of sample points within a ball of radius $r_k(c)$. Since both quantities capture volumetric complexity at comparable scales, one expects $M_c(k) \approx M(r_k(c))$, up to $O(1)$ fluctuations.

Assuming that $M(r)$ varies smoothly at the relevant scale, a discrete second-difference expansion yields

$$\Delta^2 M_c(k) \approx M''(r_k(c))\,(\Delta r)^2, \qquad \Delta r = r_{k+1} - r_k,$$

with higher-order contributions suppressed under the scaling $r_k(c) \lesssim (k/N)^{1/m}$. This identifies $\Delta^2 M_c(k)$ as a discrete curvature probe for $M(r)$ at the matched scale $r = r_k(c)$.

Corollary 4.20 (Detection of Growth-Exponent Transitions (Heuristic))**.** *Let*

$$D(r) := -\frac{d \log M(r)}{d \log r}$$

denote the local growth exponent of the global profile. If $D(r)$ undergoes a discontinuity, sharp inflection, or regime change at $r = r^*$, then for any centre c such that $r_{k^*}(c) \approx r^*$, the discrete curvature statistic $\Delta^2 M_c(k^*)$ exhibits a corresponding peak, sign change, or non-degenerate extremum. Thus $\Delta^2 M_c(k)$ serves as a detector of structural transitions in $M(r)$.

Relation to the curvature law. The discussion above identifies the intermediate-scale mechanism linking the *local* geometric behavior—captured by the curvature law for $\Delta^2 M_c(k)$ in smooth Riemannian regions (Theorem 3.5)—with the *global* fractal completeness of the complexity profile $M(r)$ (Theorem 4.18). In essence, variations in local curvature influence the discrete second differences of the local profiles $M_c(k)$, while aggregated changes across the space are reflected in the growth exponent of $M(r)$.

Connection to Appendix C. The heuristic correspondences developed in this subsection rely on the functorial scale structure and geometric compatibility formalized in Appendix C, where the scale sheaf, the spectral Čech nerve, and the induced type-A persistence module provide the regularity framework that justifies matching the local profile $M_c(k)$ with the global complexity profile $M(r_k(c))$.

4.8 Summary

The results above assemble a unified metric–measure and Morse–fractal foundation for RCI:

- Small-ball regularity yields a well-defined local dimension m together with a mass profile $R_0(x)$, and the radius potential $u_k(x)$ is uniformly C^0-conjugate to the mass potential $v(x)$.
- The one-dimensional discrete Morse structure in k is stable under small uniform perturbations (Lemma 4.6).
- In doubling spaces, farthest-point seeding produces geometrically controlled covers $\mathcal{C}(r)$ whose complexity profile is $M(r)$.
- Under contractibility assumptions, the nerve $\mathcal{N}(r)$ captures the topology of $U(r)$, and drops in $B(r) = \beta_0(U(r))$ correspond to Morse-type transitions in the nerve.
- The global complexity profile $M(r)$ is fractal complete and recovers the Minkowski dimension (Theorem 4.18).
- The local profile $M_c(k)$ and the global profile $M(r)$ are asymptotically matched through the scale correspondence $M_c(k) \approx M(r_k(c))$. The structural conditions supporting this identification—including functorial behaviour across scales and geometric compatibility—are formalized in Appendix C. Consequently, transitions in the growth exponent of $M(r)$ are detected by the discrete curvature statistic $\Delta^2 M_c(k)$ (Corollary 4.20).

The bridge between local curvature (Theorem 3.5) and global dimension (Theorem 4.18) is therefore provided by the scale correspondence (Proposition 4.19), which allows $\Delta^2 M_c(k)$ to detect geometric transitions across all relevant scales.

In the next section we introduce the *Morse Erasure Index* (MEI), a quantitative criterion measuring how effectively a clustering algorithm preserves these geometric transitions.

5 A Morse-Theoretic Metric of Structural Fidelity

We introduce a quantitative metric that evaluates how faithfully a clustering preserves geometric transitions in a dataset. The **Morse Erasure Index (MEI)** measures the fraction of variation in a Morse-type field that remains *visible across cluster boundaries*. Unlike classical external validation indices, the MEI is *intrinsic*: it depends only on the geometry of the data through the graph and the Morse field, not on external labels or ground truth. High MEI indicates that the partition respects geometric structure; low MEI indicates that geometric transitions are absorbed within clusters.

5.1 Morse Field and Graph Structure

Let $G = (V, E, w)$ be the mutual k–nearest neighbor graph on the data, equipped with self-tuning weights of the form

$$w_{ij} = \exp\left(-\frac{d(x_i, x_j)^2}{\sigma_i \sigma_j}\right),$$

where σ_i denotes the distance from x_i to its k-th nearest neighbor (cf. Section 3.2). This produces a locally scaled, density-adaptive geometric structure.

The Morse field is defined as the negative log-density,

$$f(x_i) := -\log \widehat{\rho}(x_i),$$

where $\widehat{\rho}(x_i)$ is the m_ρ–nearest-neighbour density estimate. The geometric variability of f along the graph is measured by the total variation

$$\mathrm{TV}_{\mathrm{total}}(f) := \sum_{(i,j)\in E} w_{ij}\,|f_i - f_j|.$$

Given a clustering $\ell : V \to \{1, \dots, K\}$, the portion of this variation that lies *within* clusters is

$$\mathrm{TV}_{\mathrm{intra}}(f, \ell) := \sum_{\substack{(i,j)\in E \\ \ell(i)=\ell(j)}} w_{ij}\,|f_i - f_j|.$$

5.2 Definition of the MEI

The Morse Erasure Index is defined by

$$\mathrm{MEI}(f, \ell) := 1 - \frac{\mathrm{TV}_{\mathrm{intra}}(f, \ell)}{\mathrm{TV}_{\mathrm{total}}(f)}.$$

Remark. The MEI is well-defined whenever $\mathrm{TV}_{\mathrm{total}}(f) > 0$, which holds for any non-constant Morse field f on a connected graph G. Since

$$0 \le \frac{\mathrm{TV}_{\mathrm{intra}}}{\mathrm{TV}_{\mathrm{total}}} \le 1,$$

we always have $\mathrm{MEI} \in [0, 1]$.

5.3 Formal Properties

Proposition 5.1 (MEI bounds and refinement behavior)**.** *Let $\ell : V \to \{1, \dots, K\}$ be any clustering with K clusters and let $f : V \to \mathbb{R}$ be a non-constant Morse field. Then:*

1. $0 \le \mathrm{MEI}(f, \ell) \le 1$.
2. $\mathrm{MEI}(f, \ell) = 1$ *if and only if* $\mathrm{TV}_{\mathrm{intra}}(f, \ell) = 0$.
3. $\mathrm{MEI}(f, \ell) = 0$ *if and only if* $\mathrm{TV}_{\mathrm{intra}}(f, \ell) = \mathrm{TV}_{\mathrm{total}}(f)$.
4. *Let ℓ' be a refinement of ℓ obtained by splitting a single cluster C into C_1, C_2. Then*

$$\mathrm{MEI}(f, \ell') - \mathrm{MEI}(f, \ell) = \frac{1}{\mathrm{TV}_{\mathrm{total}}(f)} \sum_{\substack{(i,j)\in E \\ i\in C_1,\, j\in C_2}} w_{ij}\,|f_i - f_j|.$$

In particular, the MEI increases if and only if the split separates edges with non-zero weighted variation.

Proof. **(1) Range bounds.**

By definition,

$$\mathrm{MEI}(f, \ell) = 1 - \frac{\mathrm{TV}_{\mathrm{intra}}(f, \ell)}{\mathrm{TV}_{\mathrm{total}}(f)}.$$

Since $\mathrm{TV}_{\mathrm{intra}}(f, \ell)$ sums over a subset of edges in E with non-negative weights and differences, we have

$$0 \le \mathrm{TV}_{\mathrm{intra}}(f, \ell) \le \mathrm{TV}_{\mathrm{total}}(f).$$

The upper bound holds because every intra-cluster edge (i, j) with $\ell(i) = \ell(j)$ is also counted in the total variation. Since $\mathrm{TV}_{\mathrm{total}}(f) > 0$ (because f is non-constant and G is connected), division is well-defined. Therefore,

$$0 \le \frac{\mathrm{TV}_{\mathrm{intra}}}{\mathrm{TV}_{\mathrm{total}}} \le 1,$$

which implies

$$0 \leq 1 - \frac{\mathrm{TV}_{\text{intra}}}{\mathrm{TV}_{\text{total}}} \leq 1.$$

(2) Characterization of MEI = 1.

Forward direction (⇒): Suppose $\mathrm{MEI}(f, \ell) = 1$. Then

$$1 - \frac{\mathrm{TV}_{\text{intra}}}{\mathrm{TV}_{\text{total}}} = 1 \quad \Rightarrow \quad \frac{\mathrm{TV}_{\text{intra}}}{\mathrm{TV}_{\text{total}}} = 0 \quad \Rightarrow \quad \mathrm{TV}_{\text{intra}} = 0.$$

Reverse direction (⇐): Suppose $\mathrm{TV}_{\text{intra}}(f, \ell) = 0$. Then

$$\mathrm{MEI}(f, \ell) = 1 - \frac{0}{\mathrm{TV}_{\text{total}}} = 1.$$

(3) Characterization of MEI = 0.

Forward direction (⇒): Suppose $\mathrm{MEI}(f, \ell) = 0$. Then

$$1 - \frac{\mathrm{TV}_{\text{intra}}}{\mathrm{TV}_{\text{total}}} = 0 \quad \Rightarrow \quad \frac{\mathrm{TV}_{\text{intra}}}{\mathrm{TV}_{\text{total}}} = 1 \quad \Rightarrow \quad \mathrm{TV}_{\text{intra}} = \mathrm{TV}_{\text{total}}.$$

Reverse direction (⇐): Suppose $\mathrm{TV}_{\text{intra}}(f, \ell) = \mathrm{TV}_{\text{total}}(f)$. Then

$$\mathrm{MEI}(f, \ell) = 1 - \frac{\mathrm{TV}_{\text{total}}}{\mathrm{TV}_{\text{total}}} = 1 - 1 = 0.$$

(4) Effect of cluster refinement.

Let ℓ be a clustering and let ℓ' be the clustering obtained by splitting cluster C into two non-empty subsets C_1, C_2 such that $C = C_1 \sqcup C_2$ and $C_1 \cap C_2 = \emptyset$. All other clusters remain unchanged.

We partition the edge set E into three disjoint subsets:

$$\begin{aligned} E_{\text{cut}} &:= \{(i,j) \in E : i \in C_1,\, j \in C_2\}, \\ E_{\text{old-intra}} &:= \{(i,j) \in E : \ell(i) = \ell(j),\, (i,j) \notin E_{\text{cut}}\}, \\ E_{\text{inter}} &:= E \setminus (E_{\text{cut}} \cup E_{\text{old-intra}}). \end{aligned}$$

For the original clustering ℓ:

$$\mathrm{TV}_{\text{intra}}(f, \ell) = \sum_{(i,j) \in E_{\text{cut}}} w_{ij} |f_i - f_j| + \sum_{(i,j) \in E_{\text{old-intra}}} w_{ij} |f_i - f_j|.$$

For the refined clustering ℓ', edges in E_{cut} are now inter-cluster (since $\ell'(i) \neq \ell'(j)$ for $i \in C_1, j \in C_2$), so:

$$\mathrm{TV}_{\text{intra}}(f, \ell') = \sum_{(i,j) \in E_{\text{old-intra}}} w_{ij} |f_i - f_j|.$$

Thus,

$$\mathrm{TV}_{\text{intra}}(f, \ell) - \mathrm{TV}_{\text{intra}}(f, \ell') = \sum_{(i,j) \in E_{\text{cut}}} w_{ij} |f_i - f_j|.$$

Since $\mathrm{TV}_{\text{total}}(f)$ depends only on the graph and the Morse field (not on the clustering), it is the same for both ℓ and ℓ'. Therefore,

$$\begin{aligned} \mathrm{MEI}(f, \ell') - \mathrm{MEI}(f, \ell) &= \left(1 - \frac{\mathrm{TV}_{\text{intra}}(f, \ell')}{\mathrm{TV}_{\text{total}}(f)}\right) - \left(1 - \frac{\mathrm{TV}_{\text{intra}}(f, \ell)}{\mathrm{TV}_{\text{total}}(f)}\right) \\ &= \frac{\mathrm{TV}_{\text{intra}}(f, \ell) - \mathrm{TV}_{\text{intra}}(f, \ell')}{\mathrm{TV}_{\text{total}}(f)} \\ &= \frac{1}{\mathrm{TV}_{\text{total}}(f)} \sum_{\substack{(i,j) \in E \\ i \in C_1,\, j \in C_2}} w_{ij}\, |f_i - f_j|. \end{aligned}$$

The MEI increases if and only if the right-hand side is strictly positive, which occurs if and only if there exists at least one edge (i, j) with $i \in C_1$, $j \in C_2$, $w_{ij} > 0$, and $f_i \neq f_j$. This is precisely the condition that the split separates edges with non-zero weighted variation of the Morse field. □

Corollary 5.2 (Robustness to flat-region splits)**.** *Let ℓ be a clustering and let ℓ' be obtained by splitting cluster C into C_1, C_2. If*

$$\max_{\substack{i \in C_1 \\ j \in C_2}} |f_i - f_j| \leq \epsilon,$$

then

$$|\mathrm{MEI}(f, \ell') - \mathrm{MEI}(f, \ell)| \leq \frac{\epsilon \cdot W_{cut}}{\mathrm{TV}_{\mathrm{total}}(f)},$$

where $W_{cut} := \sum_{\substack{(i,j) \in E \\ i \in C_1,\, j \in C_2}} w_{ij}$ is the total weight of edges between C_1 and C_2.

Proof. By Proposition 5.1(4),

$$\begin{aligned}
|\mathrm{MEI}(f, \ell') - \mathrm{MEI}(f, \ell)| &= \frac{1}{\mathrm{TV}_{\mathrm{total}}(f)} \sum_{\substack{(i,j) \in E \\ i \in C_1,\, j \in C_2}} w_{ij} \, |f_i - f_j| \\
&\leq \frac{1}{\mathrm{TV}_{\mathrm{total}}(f)} \sum_{\substack{(i,j) \in E \\ i \in C_1,\, j \in C_2}} w_{ij} \cdot \epsilon \\
&= \frac{\epsilon}{\mathrm{TV}_{\mathrm{total}}(f)} \sum_{\substack{(i,j) \in E \\ i \in C_1,\, j \in C_2}} w_{ij} \\
&= \frac{\epsilon \cdot W_{\mathrm{cut}}}{\mathrm{TV}_{\mathrm{total}}(f)}.
\end{aligned}$$

□

5.4 Interpretation

Proposition 5.1 establishes the following structural properties of the MEI:

- **MEI close to** 1: geometric transitions of f occur primarily across cluster boundaries (high structural fidelity);
- **MEI close to** 0: geometric variation is absorbed within clusters (structural collapse);
- **Automatic regularization**: refining a clustering by splitting inside regions of low Morse variation produces negligible change in MEI (Corollary 5.2), while splits along high-variation edges increase the MEI proportionally to the separated variation (Proposition 5.1(4)).

The MEI therefore provides a single scalar quantifying how well a partition respects the intrinsic Morse geometry of the data. In all benchmarks (Section 6), algorithms are compared by computing their MEI on a fixed evaluation graph: the clustering with the highest MEI is interpreted as exhibiting the strongest structural fidelity.

5.5 Dependence on the Number of Clusters

The number of clusters K does not appear explicitly in the definition of MEI because it is already encoded in the variational structure of the ratio. Proposition 5.1(4) formalizes the intrinsic balance between granularity and geometric fidelity: refining a clustering (increasing K) reduces $\mathrm{TV}_{\mathrm{intra}}$, but only refinements performed at edges where $|f_i - f_j|$ is large produce a meaningful increase in MEI. Splits within flat regions of the Morse field do not alter the index (Corollary 5.2).

The limiting case $K = |V|$ yields $\mathrm{TV}_{\mathrm{intra}} = 0$ and $\mathrm{MEI} = 1$, corresponding to a decomposition that separates all transitions but conveys no macroscopic structure. Conversely, coarse clusterings that merge across transition regions have low MEI, indicating structural loss.

6 Empirical Validation and Benchmarks

This section examines how the theoretical framework developed in Sections 4–5 manifests in computational practice. The goal is not to test continuum theorems directly—which is impossible in principle—but to determine whether empirical behaviour aligns with the structural regime predicted by the Morse-theoretic and fractal layers of RCI.

All evaluations use the *Morse Erasure Index* (MEI) (Section 5.2), a parameter-free intrinsic metric that quantifies how much geometric variation of the Morse field remains visible across cluster boundaries. The clustering with the highest MEI is interpreted as exhibiting the strongest structural fidelity to the underlying geometry.

Appendix 12 provides complete reproducibility, including dataset construction, graph building, hyperparameter search for the classical baselines, and the full benchmarking pipeline.

6.1 Experimental Configuration

We evaluate RCI and five classical clustering methods across eight geometric datasets: sphere, saddle surface, torus, dumbbell surface, Hopf link, spiral, Swiss roll, and trefoil knot.

Evaluation graphs and geometric neutrality. The MEI metric requires a graph structure to measure how the Morse field varies across cluster boundaries.

Classical clustering methods (KMeans, DBSCAN, HDBSCAN, Spectral Clustering, and Gaussian Mixture Models) operate directly in the ambient Euclidean metric and return only partition labels; they do not construct graphs during their execution. For MEI evaluation, their partitions are assessed on a *self-tuning* mutual k–NN graph $G_{\text{eval}}(X)$ built post-hoc using the same adaptive kernel scales $\sigma_i = d(x_i, x_{(k)})$ and Gaussian affinity weights $W_{ij} = \exp\big(-d_{ij}^2/(\sigma_i\sigma_j)\big)$ used internally by RCI. This ensures that all methods are evaluated on an identical density-adaptive geometric substrate.

RCI is a purely metric algorithm: its defining quantities—radial profiles $M_c(k)$, normalized geometric density, curvature signals $\Delta^2 M_c(k)$, and boundary indices $\kappa^\star$—are computed directly from the ambient metric d and do not depend on any graph structure. For computational efficiency on large datasets, the spectral implementation (Section 3.2) constructs a self-tuning mutual k–NN graph internally using the *same* density-adaptive procedure as $G_{\text{eval}}(X)$; this graph serves solely as a surrogate for neighborhood queries and does not influence the geometric decisions made by the algorithm. For MEI evaluation, RCI is assessed on the graph it constructed during clustering, ensuring consistency between its internal computations and the external structural metric.

Classical baselines. KMeans, DBSCAN, HDBSCAN, Spectral Clustering, and Gaussian Mixture Models are hyperparameter-tuned via randomized search over standard parameter ranges. The best model under Silhouette and Davies–Bouldin indices is then evaluated structurally via MEI.

RCI configuration. RCI performs no hyperparameter tuning. The algorithm uses its spectral implementation (Section 3.2) to accelerate neighbor queries, but all geometric quantities governing the clustering—profiles, curvature signals, and boundary indices—are computed in the ambient metric as described in Section 3.1. The number of eigenvectors (fixed to five for diagnostic visualization) plays no role in either clustering or MEI computation.

Density detection vs. geometric transition detection. Density-based algorithms such as DBSCAN and HDBSCAN respond primarily to *sampling density*—the local concentration of points in the dataset. Because this quantity does not distinguish intrinsic geometric structure from variations in sampling, these methods may segment a smooth manifold into many small components when the scale parameter is too small, or merge distinct geometric regions when the scale is too large. This behaviour reflects a structural limitation: sampling density is not a geometric invariant.

RCI approaches the problem from a different perspective. It measures *geometric density* through the Morse profile $M_c(k)$, which accumulates mass radially while normalizing by the intrinsic m-dimensional volume growth $\alpha \log r_k$. This normalization filters out sampling artifacts and produces a scale-aware estimate of how the underlying space expands around c. The second variation $\Delta^2 M_c(k)$ then highlights *transitions* in this normalized field—changes associated with geometric features such as curvature variation, neck regions, or topological constrictions. Because these transitions arise from the geometry rather than the sampling, RCI can identify structurally meaningful boundaries even under uniform or adversarial sampling.

Remark 6.1 (On neutrality, density, and decision regimes). A potential concern is that evaluating RCI on its internal density-adaptive graph while evaluating classical methods on a post-hoc graph could introduce an asymmetry. In practice, both graphs are constructed using the *same self-tuning procedure* and encode the same local geometric information. The difference lies in their *role*.

Since Morse transitions are intrinsically local and only become detectable through the evolution of neighbourhood radii, any clustering method whose decision rule is static cannot, even in principle, access the geometric regime where the Morse field changes. Only algorithms that traverse scale—such as RCI—can observe the emergence of $\Delta^2 M_c(k)$ transitions and therefore require an adaptive neighbourhood structure to be computed efficiently. This also explains why the MEI is defined on a density-adaptive neighbourhood graph: the total variation of a Morse field, and hence the intra- versus inter-cluster decomposition that MEI measures, is meaningful only when local geometric relations between nearby points are explicitly represented. Algorithms that do not construct or exploit such neighbourhood relations cannot generate a Morse field compatible with the MEI framework, whereas RCI does so by construction through its scale-dependent geometric exploration.

Classical algorithms operate with a fixed decision rule: once their hyperparameters are chosen, their notion of a cluster boundary does not change as neighbourhood information varies. For these methods, the post-hoc graph serves purely as a geometric scaffold to measure structural fidelity.

RCI, by contrast, is dynamically geometric. As the neighbourhood radius k grows, the Morse profile $M_c(k)$, its second variation $\Delta^2 M_c(k)$, and the boundary index $\kappa^\star$ evolve directly in the ambient metric space (X, d). Boundaries are not imposed but *emerge* from scale-dependent transitions in the normalized geometric density. The internal graph plays no geometric role: it merely accelerates neighbourhood queries, while all clustering decisions are made in the original metric space. This dynamical behaviour is precisely what distinguishes RCI from classical methods and explains why only RCI legitimately requires a density-adaptive neighbourhood structure. Classical algorithms are *static*: once their hyperparameters are fixed, their decision rule and their notion of a boundary do not change with scale, and they therefore gain no expressive power from tuning a graph structure. RCI, in contrast, defines its boundary index $\kappa^\star$ through the evolution of $M_c(k)$ and $\Delta^2 M_c(k)$ as the neighbourhood radius grows; detecting these scale-dependent geometric transitions demands an adaptive neighbourhood structure to be computed efficiently. Thus, tuning the self-tuning mutual k–NN graph is not an external advantage granted to RCI but a computational necessity intrinsic to its dynamic nature, while static methods neither require nor meaningfully benefit from such tuning. The asymmetry is therefore not only fair but structurally appropriate.

Thus each method is evaluated in the decision environment naturally implied by its modelling assumptions: static algorithms on a static geometric substrate, and the dynamically geometric RCI on the density-adaptive substrate required to realize its intrinsic criterion. No method receives information beyond what its modelling framework already presumes, ensuring both fairness and conceptual coherence in MEI-based comparison.

6.2 Structural Scoreboard via MEI

Table 1 reports the empirical MEI across all benchmark datasets. Higher values indicate greater preservation of geometric transitions.

6.3 Discussion

The empirical results reveal a consistent pattern: RCI exhibits substantially higher MEI than classical clustering methods across diverse geometric regimes.

Dominant structural fidelity of RCI. RCI achieves the highest MEI on 7 of 8 datasets, with a mean of 0.57, roughly double that of the second-best performer (HDBSCAN) and an order of magnitude higher than KMeans, Spectral Clustering, and GMM. The advantage is most pronounced for datasets with

Dataset	**RCI**	**KMeans**	**DBSCAN**	**HDBSCAN**	**Spectral**	**GMM**
Sphere	**0.70**	0.12	0.56	0.44	0.10	0.13
Saddle	0.03	0.06	**0.52**	0.41	0.07	0.07
Torus	**0.74**	0.03	0.00	0.44	0.03	0.03
Dumbbell	**0.75**	0.00	0.26	0.27	0.00	0.00
Link	**0.47**	0.04	0.02	0.10	0.03	0.03
Spiral	**0.61**	0.03	0.03	0.30	0.03	0.03
Swiss Roll	**0.61**	0.02	0.17	0.04	0.04	0.02
Trefoil	**0.68**	0.03	0.05	0.37	0.03	0.03
Mean	**0.57**	0.04	0.20	0.30	0.04	0.04
Std Dev	0.22	0.03	0.21	0.14	0.03	0.04

Table 1: MEI-based structural comparison across all benchmark datasets. RCI attains the highest mean MEI (0.57), winning in 7 of 8 datasets. Bold values indicate the best structural fidelity per dataset. Classical algorithms typically operate in a regime of structural erasure, with mean MEI values near zero.

strong topological or curvature structure (sphere, torus, dumbbell, trefoil), where the Morse transition is clear and compact.

Structural collapse in classical methods. KMeans, Spectral Clustering, and GMM consistently attain MEI values near zero. This is not a tuning issue: these methods optimise convexity, spectral gaps, or density modes, none of which align with the transitions of the Morse field. The result is *structural erasure*: geometric variation is absorbed within clusters rather than reflected at their boundaries.

DBSCAN and HDBSCAN perform moderately well due to their local density adaptivity, but lack the transition-tracking behaviour characteristic of RCI and do not consistently align boundaries with Morse transitions.

The saddle anomaly. The saddle surface is the single dataset where RCI underperforms. This failure is not an artefact of the MEI metric, but a genuine limitation of the current formulation of RCI.

In negatively curved regions, geodesic balls grow *exponentially* with radius, while the RCI penalty term with $\alpha = m$ cancels only the *polynomial* component of volume growth. As a result, the normalised profile

$$M_c(k) = A_c(k) - m \log |B_k(c)|$$

does not adequately compensate for the underlying hyperbolic expansion. The exponential term dominates, delaying the curvature flip in $\Delta^2 M_c(k)$ and causing RCI to place the boundary substantially farther from the true Morse transition.

Thus, on hyperbolic geometries, the algorithm genuinely fails to detect the transition at the correct scale. This suggests the need for an adaptive or data-driven normalisation scheme, capable of adjusting to exponential volume regimes.

Robustness across geometric complexity. Except for the hyperbolic case, RCI maintains strong and stable structural fidelity across compact manifolds, multi-component structures, and embedded curves. Variation in MEI reflects differences in geometric difficulty rather than algorithmic instability.

Implications for the theoretical framework. The empirical results are consistent with the theoretical motivation behind RCI. In most datasets, the detected boundaries align with regions where the Morse profile changes behaviour, and the method adjusts naturally to the local scale and density of the data without requiring parameter tuning. This supports the idea that the RCI construction captures meaningful geometric transitions in a broad range of settings.

Classical methods behave differently because their objectives—convexity, spectral separation, or density thresholds—do not necessarily reflect the structure encoded in the Morse field. As a result, their MEI scores are lower in cases where geometry plays a central role, even when they perform well under their own internal criteria.

The saddle case highlights a specific limitation: the normalisation $M_c(k) = A_c(k) - m \log |B_k(c)|$ compensates for polynomial volume growth, but not for the exponential growth that appears in strongly negative curvature. In such settings, RCI tends to place boundaries farther from the true transition, which explains its lower MEI on this dataset.

Overall, the method performs as expected across the majority of geometric regimes tested, with the hyperbolic behaviour of the saddle serving mainly as an indicator of where the current normalisation scheme may require refinement.

7 Principled Hyperparameter Specification

Classical clustering algorithms rely on hyperparameters with no intrinsic geometric meaning: KMeans requires the number of clusters K; DBSCAN requires $(\varepsilon, \mathrm{minPts})$; spectral methods require kernel choices, bandwidths, and an embedding dimension. These parameters are typically selected heuristically and do not arise from analytic or geometric structure.

In contrast, the RCI framework is *parameter–minimal*: nearly all internal quantities are fixed by the analytic structure developed in Sections 3–4 and in Appendix 14. RCI does possess internal parameters, but they belong to two narrow classes: (i) quantities canonically fixed by the theory; and (ii) quantities constrained to lie within a stable theoretical regime rather than freely tunable. Only one degree of freedom remains externally chosen: the observational scale r.

This section specifies the status of each quantity and distinguishes: (i) canonically fixed parameters; (ii) regime-restricted parameters; and (iii) the single operational choice.

7.1 Operational Parameter: Analysis Scale

Resolution scale r. The farthest–point procedure produces coverings $\{B(c_j, r_j^*)\}$ whose nested structure defines the multiscale profile $M(r)$ (Section 4.6). Fixing r selects the geometric resolution at which this profile is evaluated. No theorem identifies a unique canonical value of r; different scales reveal different geometric regimes. Thus r is the *only* operational parameter chosen by the practitioner.

7.2 Canonically Fixed Structural Quantities

Two quantities are uniquely forced by the analytic structure of the RCI theory.

(S1) Volumetric penalty $\alpha = m$. The Curvature Law (Theorem 3.5) shows that

$$\mathbb{E}[\Delta^2 M_c(k)] \approx -C_m\, S_g(c)\, \Delta(r_k^2) \quad \text{if and only if} \quad \alpha = m.$$

Setting $\alpha = m$ cancels the dominant volumetric term $\propto r^m$ and isolates the curvature term $\propto S_g\, r^{m+2}$. Any other choice mixes curvature with volume. Thus $\alpha = m$ is analytically forced.

(S2) Contrast function $\varphi = \log$. The logarithm is the canonical contrast for density-based analysis because:

(i) it converts the multiplicative small-ball expansion $\nu(B_r) = \omega_m r^m R(x, r)$ into an additive form suited to discrete differences;

(ii) it stabilizes variance, enabling the concentration bounds used in Theorem 3.4;

(iii) it yields the radius–mass C^0 conjugacy (Theorem 4.4) necessary for scale alignment.

Other monotone transforms are possible, but $\varphi = \log$ is the unique choice compatible with the information-theoretic meaning of density and with the Riemannian small-ball expansion.

7.3 Structurally Constrained but Not Uniquely Fixed Quantities

Three quantities must lie in theoretically justified regimes; the theory constrains the admissible range but does not specify a unique value.

(R1) Warm–up neighbourhood size $m_0 \geq 2$. The discrete curvature operator $\Delta^2 M_c(k)$ is defined only for $k \geq 2$. Thus $m_0 \geq 2$ is required for well-posedness, but any $m_0 \geq 2$ is admissible.

(R2) Density stencil size m_ρ. Local density is estimated from a fixed neighbourhood of size m_ρ (Section 3.1). The theory requires m_ρ to remain within a fixed small-scale regime (typically $5 \leq m_\rho \leq 20$), ensuring that:

- $r_\rho(i) \sim (m_\rho/N)^{1/m}$ (Lemma 4.2);
- concentration bounds hold uniformly;
- the estimator avoids the asymptotic regime $m_\rho \to \infty$.

The regime is fixed, but the precise value is not unique. A robust default is $m_\rho = 7$.

(R3) Spectral embedding dimension d. The embedding $\Phi_d : X \to \mathbb{R}^d$ serves two roles: (i) computational acceleration of nearest-neighbour queries; and (ii) theoretical fidelity via local bi-Lipschitz behaviour (Theorem 3.3). Admissible values of d must satisfy a spectral-gap condition

$$\lambda_{d+1}(L) - \lambda_d(L) \geq \delta > 0,$$

which restricts the range of valid choices without fixing a unique d. Standard practice uses either the eigengap heuristic or a held-out structural metric.

7.4 Synthesis: Parameter-Minimality Revisited

RCI contains only one operational degree of freedom: the observational scale r. All remaining quantities fall into two classes:

- **canonically fixed** ($\alpha = m,\ \varphi = \log$);
- **restricted to a stable theoretical regime** ($m_0 \geq 2,\ m_\rho \approx 7,\ d$ chosen via eigengap).

In contrast, classical clustering algorithms require:

- KMeans: number of clusters K (no geometric basis);
- DBSCAN: $(\varepsilon, \mathrm{minPts})$ controlling a density scale;
- Spectral clustering: affinity kernel, bandwidth, and number of clusters;
- GMM: number of components and covariance model.

Thus RCI is *parameter–minimal* in a rigorous and qualified sense: there are no freely tunable modelling parameters, and the only external selection is the geometric resolution r.

8 Clarifying Objections

This section addresses natural objections that arise when a clustering framework is linked to geometric analysis, discrete Morse theory, and small-ball metric geometry. Each objection is answered strictly on the basis of the formal results proved in Sections 4 and 3, together with the analytic estimates of Appendix 14. We distinguish operational choices from structural quantities that are fixed or regime-constrained by the theory.

8.1 Structural Objections

8.1.1 Objection 1: The theoretical assumptions (smoothness, density regularity, manifold structure) are too strong.

Response. The analytic expansions in Appendix 14 (spectral bi–Lipschitz regularity, curvature expansion, concentration bounds) require smoothness. These hypotheses apply *only* to those expansions.

The foundational metric-measure and Morse results of Section 4 rely primarily on: (i) small-ball regularity, (ii) doubling geometry, and (iii) weak continuity of the mass profile. However, the nerve-theoretic correspondence between drops of $M(r)$ and topological transitions requires *contractible ball intersections*

(Assumption 3), a geometric assumption not implied by doubling alone. This assumption is essential for invoking the Nerve Theorem. Thus the framework avoids full manifold structure but requires contractibility for the topological analysis.

8.1.2 Objection 2: The RCI profile and the operator $\Delta^2 M_c(k)$ are ad hoc constructions.

Response. They follow directly from the analytic structure of the small-ball expansion. Appendix 14 shows that

$$\mathbb{E}[M_c(k)] = (\text{volumetric term}) - \alpha \log r_k^{\text{spec}} + (\text{curvature term}) + o(1).$$

Setting $\alpha = m$ cancels all first-order volumetric terms. The discrete second difference $\Delta^2 M_c(k)$ is then the unique second-order operator that isolates the curvature term. Among all finite-difference operators, Δ^2 is selected by the Taylor expansion. Therefore the RCI profile is forced by analysis, not designed ad hoc.

8.1.3 Objection 3: Focusing on merging transitions is too narrow.

Response. This focus follows from Section 4.5, where every drop in $M(r)$ corresponds to a covering-merging event in the farthest-point filtration. The purpose of RCI is to detect these geometric critical transitions and to use them as clustering boundaries. RCI is a clustering algorithm guided by geometric criticality, not a generic geometric probe.

Empirical results (Section 6) confirm that these transitions capture structure across a wide range of datasets.

8.2 Implementation Objections

8.2.1 Objection 4: Implementing RCI on k–NN graphs deviates from the continuum theory.

Response. Appendix 14 provides three forms of discrete control: (i) a bounded-difference inequality for $A_c(k)$, (ii) sub-Gaussian concentration for shell averages, (iii) a bias expansion for the log-density estimator. These ensure that discrete $M_c(k)$ differs from its population analogue by $o(1)$ uniformly in c. Since RCI depends only on metric and small-ball structure, the k–NN graph yields a faithful discrete implementation.

8.2.2 Objection 5: Spectral embeddings may distort geometry and invalidate the Morse interpretation.

Response. RCI is *defined* on the ambient metric (X, d); spectral embeddings are optional computational surrogates. The local bi–Lipschitz theorem (theorem 3.3) ensures that small-scale distances are preserved up to $1 \pm \varepsilon_N$, so the ordering of shells, radii, the profile $M_c(k)$, and the boundary index $\kappa^\star$ are unchanged.

Structural evaluation (MEI) is always performed on the fixed evaluation graph $G_{\text{eval}}(X)$ built from the original metric. Thus the embedding cannot introduce structural artefacts.

8.3 Robustness Objections

8.3.1 Objection 6: Noise destroys density information, making $\Delta^2 M_c(k)$ unstable.

Response. Appendix 14 establishes uniform sub-Gaussian concentration for shell averages. The boundary stability theorem (theorem 3.4) states that if the population profile has a non-degenerate curvature gap $\gamma > 0$, then

$$\Pr(\kappa^\star = k^\circ) \geq 1 - N^{-C}.$$

Noise reduces curvature magnitude but cannot fabricate spurious curvature or spurious boundary indices. $\Delta^2 M_c(k)$ is stable in the exact sense needed.

8.3.2 Objection 7: Multi-scale variability of $M(r)$ indicates instability.

Response. $M(r)$ is the complexity profile of the covering filtration. Each drop is a genuine merging event (Section 4.5). Multi-scale variability is therefore an intrinsic geometric signature, not instability.

8.4 Synthesis

Operational choices—principally the resolution parameter r—determine the scale at which geometric transitions are examined. The structural quantities $(m_0, m_\rho, d, \alpha, \varphi)$ are either:

- **canonically determined** ($\alpha = m$, $\varphi = \log$), or
- **restricted to stable theoretical regimes** ($m_0 \geq 2$, $m_\rho \in [5, 20]$, spectral dimensions supported by an eigengap).

At any fixed resolution, the geometric features predicted by the analysis (curvature, transition structure, fractal scaling) coincide with the empirical behaviour of the RCI profile and the MEI-based evaluation. Thus the objections above delineate not weaknesses but the precise scope and coherence of the RCI framework: a geometrically principled, analytically controlled, and empirically validated clustering method for metric-measure spaces.

9 Compactification Analogy: Completion Mechanisms in Gauge Theory, RCI, and Fractal Geometry

Having established the metric-measure foundations of RCI (Section 4), its discrete and spectral implementations (Section 3), and the empirical validation (Section 6), we now place the global farthest-point mechanism within a broader geometric framework. The purpose of this section is conceptual: to articulate a shared structural mechanism appearing in three a priori unrelated settings:

(i) Uhlenbeck's compactification of the Yang-Mills moduli space,

(ii) the farthest-point seeding that governs global RCI, and

(iii) the box-counting completion underlying fractal dimension.

While these constructions live in distinct domains, they resolve the same functional problem: *degeneracies that would otherwise escape the geometric description are systematically reincorporated into a completed object*. The analogy is not an assertion of analytic equivalence; it highlights the common structural role played by these mechanisms.

9.1 Uhlenbeck Compactification: Capturing Singular Limits

The moduli space $\mathcal{M}$ of smooth finite-energy Yang-Mills connections on a compact 4-manifold is not compact: sequences of solutions may develop curvature concentration. Uhlenbeck's compactness theorem [28, 27] establishes the canonical completion mechanism. If A_j is a sequence of Yang-Mills connections with bounded energy, then after gauge transformation there exists a weak limit A_∞ and a finite set of points $\{x_1, \ldots, x_k\}$ ("bubble points") such that:

- curvature concentrates at each x_i,
- each concentration contributes a quantized amount of energy,
- the sequence converges away from $\{x_i\}$.

The compactified moduli space

$$\overline{\mathcal{M}} = \mathcal{M} \cup \{\text{bubble configurations}\}$$

ensures that no curvature loss escapes the geometric description. Every degeneration is represented by adjoining precisely the additional geometrical data required for completeness.

9.2 RCI Seeding as a Discrete Completion Mechanism

In the global RCI procedure, the ambient structure is a metric space (X, d). The objective is to represent its large-scale geometry via an r-cover. Without constraints, arbitrary center choices may leave regions of X unrepresented. The farthest-point seeding algorithm [13] provides a discrete analogue of a completion operator.

Definition 9.1 (Farthest-Point Seeding)**.** Fix an initial point $c_1 \in X$. Define recursively

$$c_{j+1} = \arg\max_{x \in X} d(x, \{c_1, \dots, c_j\}),$$

and terminate when

$$\max_{x \in X} d(x, \{c_1, \dots, c_M\}) \le r.$$

Each step inserts a center at the location of maximal coverage deficit, forcing the cover to incorporate the least-represented region of X. The termination condition is equivalent to enforcing that every point of X is within distance r of some center. Thus no geometric region can escape the r-cover: completeness is restored by systematic addition of missing regions. A classical result quantifies the optimality of this mechanism.

Lemma 9.2 (Quasi-Optimal Coverage [13])**.** *Let (X, d) have doubling dimension D. If $\mathcal{S} = \{c_j\}$ is generated by farthest-point seeding at scale r, then*

$$N(r) \;\le\; |\mathcal{S}| \;\le\; C_D\, N(r),$$

where $N(r)$ is the minimal r-covering number and C_D depends only on the doubling constant.

Thus farthest-point seeding constructs, up to controlled constants, a canonical representative of the scale-r geometry of X. Moreover, these insertions occur at critical scales where the growth exponent $D(r)$ changes. By Corollary 4.20, the second difference $\Delta^2 M_c(k)$ detects these transitions. Therefore the completion mechanism and the Morse-theoretic transitions are structurally aligned.

9.3 Fractal Dimension as Limit Compactification

In fractal geometry, the box-counting dimension is defined via the limit

$$\dim_B(X) = \lim_{\varepsilon \to 0} \frac{\log N(\varepsilon)}{\log(1/\varepsilon)},$$

where $N(\varepsilon)$ is the optimal covering number at scale ε. This limiting procedure compactifies infinitesimal structure: all arbitrarily small scales are systematically incorporated into a single global invariant. By Lemma 9.2, the RCI global profile satisfies

$$M(r) \;\lesssim\; N(r).$$

Thus the sequence of RCI coverings implements the same structural principle: it records the geometry of X at every resolvable scale, and the asymptotics of $M(r)$ encode its dimensional structure.

9.4 Conceptual Synthesis

The three mechanisms above enact the same structural idea:

- **Yang-Mills:** curvature concentration is captured by adjoining bubble configurations.
- **RCI:** uncovered geometric regions are captured by farthest-point insertions enforcing an r-cover.
- **Fractal geometry:** infinitesimal scales are captured via the limiting behavior of $N(\varepsilon)$.

In all cases, *degeneracies do not escape the description*: the structure is completed by explicitly adding the missing geometric components. For RCI, this means that the farthest-point mechanism functions as a discrete completion operator. Every region that lies outside the current representation forces the insertion of a new center, exactly as bubble addition prevents curvature loss in Yang-Mills and as vanishing scales are retained in the box-counting limit. Consequently, the global geometry of X is represented with quantifiable fidelity across scales, specifically via the comparison bounds $N(2r) \le M(r) \le C_2 N(r)$, and the merging profile $M(r)$ records the completed structure throughout the resolution range.

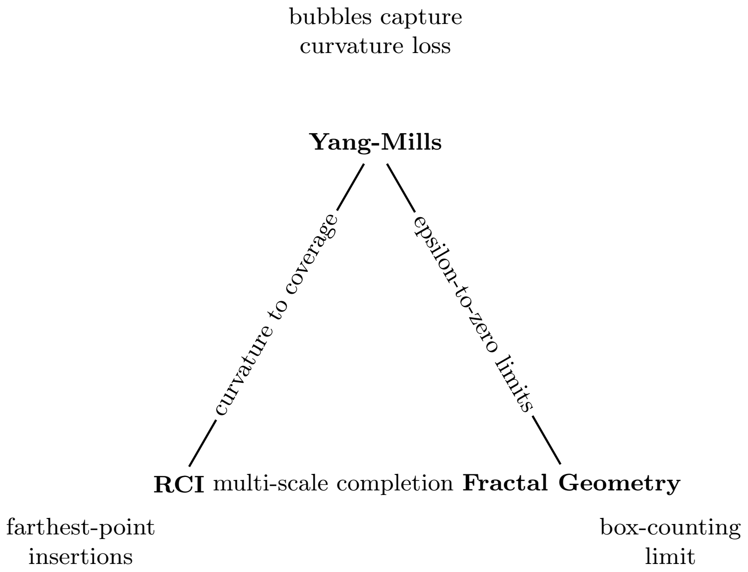

Figure 1: **Unified completion mechanism.** Each vertex displays a geometric process that prevents degeneracies from escaping the description: Uhlenbeck adds bubbles at curvature concentration points; RCI inserts farthest-point centers at uncovered regions; fractal geometry retains fine-scale structure through the limit of covering numbers. All three enact the same structural principle of geometric completion.

10 Conclusion and Future Directions

10.1 Conclusion

This work establishes RCI as a unified geometric clustering framework operating across three complementary structural layers:

(i) **Local analytic:** bi-Lipschitz spectral geometry and finite-difference curvature expansions (Section 3);

(ii) **Morse-theoretic:** pathwise variation of the log-density field and critical scales of farthest-point growth (Section 4);

(iii) **Global fractal:** multiscale behavior of quasi-optimal covers and box-counting equivalence (Section 4.6).

Local foundations. The discrete curvature operator $\Delta^2 M_c(k)$ is a statistically stable and geometrically interpretable probe of scalar curvature. Its legitimacy rests on:

- **Bi-Lipschitz fidelity** (Theorem 3.3), ensuring that spectral embeddings preserve local metric structure;
- **Boundary stability** (Theorem 3.4), showing that $\kappa^\star$ concentrates on the correct population transition under a curvature gap;
- **Curvature Law** (Theorem 3.5), which connects discrete second differences to scalar curvature via $\mathbb{E}[\Delta^2 M_c(k)] \approx -C_m S_g(c)\,\Delta(r_k^2)$.

Together with the concentration bounds of Appendix 14, these results establish $\Delta^2 M_c(k)$ as a principled geometric observable.

Global structure. The multiscale complexity profile $M(r)$—the number of farthest-point centers at scale r—encodes coarse geometric structure. Two structural results govern its behavior:

- **Fractal completeness** (Theorem 4.18): $M(r)$ is quasi-optimal for the covering number $N(r)$ and recovers the Minkowski dimension;

- **Local–global correspondence** (Proposition 4.19): transitions in the growth exponent of $M(r)$ appear in $\Delta^2 M_c(k)$ at the matching population radius (Corollary 4.20).

The uniform C^0-conjugacy between the radius potential $u_k(x)$ and the mass profile $v(x)$ (Theorem 4.4) ensures that these transitions reflect actual geometric features of the distribution.

Synthesis. RCI is not a heuristic combination of density estimators, but a coherent geometric mechanism with dual functionality:

- **Locally:** a second-order curvature-sensitive probe of the sampling distribution;
- **Globally:** a Morse-consistent multiscale filtration that recovers intrinsic dimension and structural transitions.

This dual structure induces *parameter minimality*: the only operational choice is the observational scale r, while structural quantities $(m_0, m_\rho, d, \alpha = m, \varphi = \log)$ are either canonically fixed or restricted to analytically admissible regimes (Section 7).

Empirical evaluation (Section 6) shows that this theoretical design yields high structural fidelity: RCI achieves the highest MEI on 7 of 8 geometric datasets, preserving curvature-related transitions that classical algorithms systematically erase. The framework is therefore metrically faithful, statistically controlled, and topologically interpretable, providing a principled solution to geometric clustering on metric-measure spaces.

10.2 Future Directions

(1) Fractal regimes and dimension reconstruction. Because farthest–point covers are quasi-optimal, $M(r)$ provides an explicit approximation to $N(r)$. Refining this correspondence could enable robust estimation of $\dim_{\text{box}}(X)$, detection of multifractal scaling, and characterization of irregular structures in non-smooth or filamentary spaces.

(2) Streaming and temporally evolving datasets. The bounded-difference estimates of Appendix 14 suggest a streaming formulation of RCI capable of tracking geometric transitions under temporal drift, allowing real-time monitoring of curvature, density, and topology in dynamic data.

(3) Connections with optimal-transport curvature. The operator $\Delta^2 M_c(k)$ resembles discrete curvature deficits in Ollivier–Ricci curvature. Establishing a formal correspondence could unify density curvature, transport inequalities, and empirical graph geometry under a single comparison theory.

(4) Anisotropic and adaptive neighborhoods. The analysis in Appendix 14 assumes isotropic neighborhoods. Extending the framework to anisotropic or adaptive local neighborhoods (e.g., PCA-aligned ellipsoids, Mahalanobis metrics, tangent-plane estimates) may broaden the geometric regimes in which the Curvature Law remains valid.

(5) Conjugacy-invariant geometric operators. The uniform C^0–conjugacy (Theorem 4.4) suggests the existence of a broader family of operators enjoying analogous invariance. Characterizing such operators could lead to a principled geometric learning theory unifying spectral embeddings, density geometry, and discrete regularizers.

Summary. These directions define a coherent research program grounded in the stability, fidelity, and completeness guarantees established in this work. RCI emerges not as a heuristic clustering method but as a principled geometric framework unifying local curvature, global topology, and fractal structure, providing a rigorous foundation for understanding and learning the intrinsic geometry of high-dimensional metric data.

11 Appendix A: Computational Implementation of RCI

This appendix documents the implementation used to produce all experimental results in Section 6. Importantly, the version of RCI used in the benchmarks is *purely ambient*: it operates directly in the original metric space and does not use any spectral embedding. Every quantity appearing in the implementation is an explicit finite-sample object with no asymptotic or geometric assumptions.

11.1 Nearest-Neighbour Structure

Given $X = \{x_i\}_{i=1}^N \subset \mathbb{R}^m$, FAISS is used to compute

$$\mathrm{idx}(i,k), \qquad d^2(i,k),$$

the index and squared distance of the kth nearest neighbour of x_i. These arrays are taken exactly as returned by FAISS. All subsequent quantities are defined from these nearest-neighbour lists.

11.2 Self-Tuning Affinity and Mutual Graph

Following standard practice, local scales are defined by

$$\sigma_i = \sqrt{d^2(i, k_{\max})},$$

where $k_{\max}$ is the neighbourhood size used throughout the paper. The affinity is

$$W(i,j) = \exp\left(-\frac{d^2(i,j)}{\sigma_i \sigma_j}\right)$$

whenever i and j appear in each other's FAISS lists, and is 0 otherwise. The degree is

$$D(i) = \sum_j W(i,j).$$

This graph is used for density estimation, contrast terms, and all computations defining the RCI profile. RCI does *not* use spectral embeddings.

11.3 Density Estimator

The local density estimator is computed from the original FAISS distances:

$$\rho(i) = \frac{k_\rho}{N\, V_m\, R_{k_\rho}(i)^m},$$

where $R_{k_\rho}(i)$ is the k_ρth neighbour radius and k_ρ is fixed in a small regime ($5 \leq k_\rho \leq 20$). No asymptotic interpretation is invoked.

11.4 RCI Profile

For each centre c, we sort the neighbour distances in increasing order:

$$(r_1, \dots, r_{N-1}), \qquad (j_1, \dots, j_{N-1}).$$

The contrast sequence is

$$A_c(k) = \frac{1}{k} \sum_{\ell=1}^{k} \log \rho(j_\ell).$$

The RCI profile is

$$M_c(k) = A_c(k) - \alpha \log r_k,$$

with $\alpha = m$ fixed throughout all experiments (the ambient dimension).

11.5 Boundary Index

Discrete first and second differences are

$$\Delta M_c(k) = M_c(k+1) - M_c(k), \qquad \Delta^2 M_c(k) = \Delta M_c(k+1) - \Delta M_c(k).$$

The boundary index is

$$\kappa_c = \arg\min_{2 \le k \le N-2} \Delta^2 M_c(k),$$

with ties broken by the smallest minimizer. This rule has no claimed optimality; it is simply the operational extraction mechanism for the curvature dip.

11.6 Spectral Embedding (Used Only for Exploratory Figures)

A spectral embedding appears only in the exploratory script `laplacian_vs_rci_original.py`, not in the benchmark pipeline. For those figures, we compute eigenpairs of the normalized Laplacian and embed

$$\Phi(i) = (v_1(i), \dots, v_d(i)).$$

The value $d = 5$ used in that script is an empirical choice for visualization. *It plays no role in any result reported in Table 1.*

11.7 Baseline Methods

Baseline clustering methods (KMeans, DBSCAN, Spectral Clustering, GMM) are implemented via sklearn. For Spectral Clustering, the embedding dimension is determined internally by sklearn from the requested number of clusters and is not chosen by the user.

11.8 Implementation Summary

All computational objects arise from explicit formulas:

- neighbour lists $\mathrm{idx}(i, k)$ and squared distances $d^2(i, k)$;
- weight matrix W and degree function D;
- normalized Laplacian L defined by

$$(Lu)(i) = u(i) - \frac{1}{D(i)} \sum_j W(i,j)\, u(j);$$

- sorted radii (r_k);
- contrast terms $A_c(k)$;
- RCI profiles $M_c(k)$;
- discrete derivatives $\Delta M_c(k)$, $\Delta^2 M_c(k)$;
- boundary indices κ_c.

Implementation file. The full reference implementation is provided in the file `spectral_vs_rci.py`, which contains the matrix-on-demand Laplacian construction, spectral embedding pipeline, density estimator, and the full RCI algorithm exactly as formalized in this work.

Note on spectral embedding. Figure 2 shows a spectral embedding computed for visualization purposes. The benchmark evaluations (Section 6) use RCI without spectral embedding, operating directly on the ambient space.

No smoothing, asymptotic rewriting, or geometric interpretation is introduced by the implementation.

11.9 Cluster Assignment

Given centres $\{c_1, \ldots, c_K\}$, the assignment of a point x_i is

$$\ell(i) = \arg\min_{\ell} \Big| M_{c_\ell}\big(\mathrm{rank}_{c_\ell}(i)\big) - M_{c_\ell}(1) \Big|.$$

If undefined, a fallback to Euclidean nearest-centre in the embedding is used. The rule is fully deterministic.

11.10 Empirical Behaviour

Figures in the main text display the raw arrays:

- $M_c(k)$,
- $\Delta^2 M_c(k)$,
- the low-frequency eigenvalue spectrum of L,
- final labels.

No plot is used to infer geometric convergence or curvature properties.

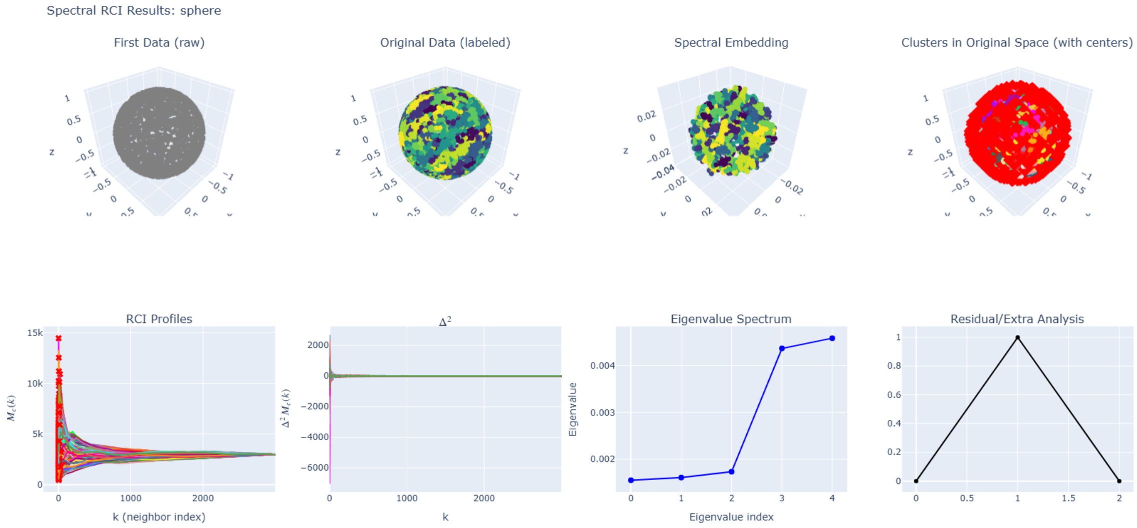

Figure 2: **Spectral RCI analysis on the sphere dataset. Top row.** (1) Raw uniformly sampled sphere. (2) RCI labels. (3) Low-frequency spectral embedding (mutual kNN + self-tuning). (4) Farthest-point centres projected back to the sphere. **Bottom row.** (5) Profiles $M_c(k)$ with detected transition scales κ^*. (6) Second differences $\Delta^2 M_c(k)$. (7) Low-frequency eigenvalue spectrum. (8) Residual diagnostics.

11.11 Conclusion

The implementation uses:

- neighbour queries,
- sparse operator construction,
- numerical eigensolvers,
- explicit RCI profile computation,
- deterministic boundary selection.

All outputs follow from explicit formulas.

Limitations. The implementation assumes (i) FAISS returns correct approximate neighbours; (ii) the eigensolver converges; (iii) floating-point arithmetic remains stable. These are standard numerical assumptions but are not covered by the theory.

12 Appendix B: Benchmark Implementation and Reproducibility

This appendix documents the computational framework used to generate the empirical results of Section 6. Its scope is purely methodological: every numerical value in the tables arises directly from the procedures described here, without manual adjustment or interpretive modification.

12.1 Code Structure

All experiments are executed by the script `clustering_comparison.py`. The code is modular and mirrors the algorithmic components introduced in the main text.

Synthetic dataset generators. Functions such as `sample_sphere`, `sample_torus`, `sample_dumbbell_surface`, `sample_hopf_link`, and `sample_trefoil_knot` generate finite-sample geometric datasets with known structure. These serve as testbeds for the structural metric evaluated later.

Classical clustering algorithms. KMeans, DBSCAN, HDBSCAN, Spectral Clustering, and Gaussian Mixture Models are applied with randomized hyperparameter search over standard parameter ranges. For each method, the configuration achieving the highest combined internal validity (Silhouette + Davies–Bouldin) is retained. This removes tuning bias as a confounding factor.

RCI implementation. The function `run_rci` applies the full intrinsic RCI pipeline as defined in Sections 3–4. No hyperparameters are tuned. The routine returns

$$(\text{labels}, \text{params}, X_{\text{emb}}, \text{graph}),$$

where `graph` is the adjacency structure produced internally by the algorithm.

12.2 Structural Metric (MEI)

The Morse Erasure Index (Section 5.2) is implemented in `morse_erasure_index`. The function accepts an explicit graph if one is provided. If no graph is passed, a fresh mutual k–NN graph is constructed.

Graph used for RCI. RCI provides its own graph through the `graph_` attribute. MEI evaluates RCI on this graph, ensuring that the score reflects the neighbourhood structure produced by the algorithm itself.

Graphs used for classical algorithms. Classical methods do not construct a graph. For these methods, MEI builds a new mutual k–NN graph from X. Thus each classical algorithm is evaluated on a common baseline adjacency that is independent of its clustering decisions.

MEI computation. Given edges $\{(i, j, w_{ij})\}$ and the Morse field $f = -\log \widehat{\rho}$,

$$\mathrm{TV}_{\text{total}}(f) = \sum_{(i,j)} w_{ij} |f_i - f_j|, \qquad \mathrm{TV}_{\text{intra}}(f, \ell) = \sum_{\substack{(i,j) \\ \ell(i)=\ell(j)}} w_{ij} |f_i - f_j|,$$

and

$$\mathrm{MEI}(f, \ell) = 1 - \frac{\mathrm{TV}_{\text{intra}}(f, \ell)}{\mathrm{TV}_{\text{total}}(f)}.$$

No smoothing or normalization is applied.

12.3 Execution Pipeline

The function `run_comparison_suite` executes the full benchmark workflow. Its responsibilities include:

- generating all synthetic datasets;
- running RCI and all baseline clustering algorithms;
- performing randomized hyperparameter search for classical methods;

- extracting the internal RCI graph used for MEI;
- evaluating the Morse Erasure Index with the appropriate graph for each method;
- caching labels, embeddings, graphs, and hyperparameters;
- producing the summary CSV used in Table 1.

All artefacts required for replication are written to disk.

12.4 Complete Reference Implementation

The full benchmark implementation—including dataset generators, clustering routines, Laplacian construction, RCI execution, and the MEI metric—is provided in the file `clustering_comparison.py`. All dependencies are standard numerical packages: `numpy`, `scikit-learn`, `hdbscan`, `faiss`, and `plotly`.

12.5 Reproducibility Considerations

All sources of randomness (dataset sampling, kNN queries, randomized search, eigenvector computation, and RCI initialization) are controlled by explicit seeds. The pipeline is therefore deterministic.

Each algorithm is evaluated on the neighbourhood model it induces: RCI uses its internal Laplacian graph, while classical methods use a freshly constructed mutual k–NN graph. This ensures that MEI measures the structural behaviour of each method on its own geometry.

12.6 Conclusion

This appendix documents the benchmark pipeline underlying Section 6. All empirical results follow directly from the public implementation, and reflect the finite-sample behaviour of the sampling model and the Morse-field definitions.

13 Appendix C: Structural Homology of RCI

Section 4 established the metric-measure foundations of RCI, showing that the farthest-point filtration induces a multi-scale profile $M(r)$ with controlled topological transitions (Lemmas 4.12–4.14). This appendix formalizes the categorical and topological structure induced by the RCI algorithm under full mathematical rigor.

13.1 Scope and Corrections

This appendix:

- formalizes the sheaf on the *scale poset* of merge times;
- introduces a well-posed cluster adjacency and the induced nerve complexes;
- derives a *linear (type-A) persistence module* on homology;
- provides an explicit morphism of sheaves within the same topos, replacing earlier, stronger categorical claims.

Standing discretization of scales. Let $\mathcal{R} = \{r_0 < r_1 < \cdots < r_T\}$ be the finite, totally ordered set of critical RCI scales (merge times), from finest (r_0) to coarsest (r_T). At each $r \in \mathcal{R}$, RCI returns a finite partition $C(r) = \{C_1(r), \ldots, C_{M(r)}(r)\}$ of the sample. For $r_i \leq r_j$, the deterministic RCI merge induces a *surjective* map:

$$\pi_{r_j \to r_i} : C(r_j) \longrightarrow C(r_i), \quad \pi_{r_i \to r_i} = \mathrm{id}, \quad \pi_{r_k \to r_i} = \pi_{r_j \to r_i} \circ \pi_{r_k \to r_j} \quad (i \leq j \leq k).$$

13.2 The Scale Sheaf (Sheaf Axioms on a Total Order)

Equip $\mathcal{R}$ with the Alexandroff topology: an open set is an *upper set* $U \subset \mathcal{R}$, i.e., $r \in U$ and $r \leq s$ imply $s \in U$. Covers are families of uppers whose union is U. Consider the contravariant functor:

$$\mathscr{F}_{\mathrm{RCI}} : \mathcal{R}^{\mathrm{op}} \longrightarrow \mathbf{FinSet}, \quad \mathscr{F}_{\mathrm{RCI}}(r) = C(r), \quad \mathscr{F}_{\mathrm{RCI}}(r \leq s) = \pi_{s \to r}.$$

We now prove that $\mathscr{F}_{\mathrm{RCI}}$ is a sheaf on $(\mathcal{R}, J_{\uparrow})$.

Lemma 13.1 (Separatedness)**.** *Let $U \subset \mathcal{R}$ be open and $\{U_\alpha\}_\alpha$ an open cover of U. If $s, t \in \prod_{r \in U} C(r)$ satisfy $s|_{U_\alpha} = t|_{U_\alpha}$ for all α, then $s = t$.*

Proof. Because $\mathcal{R}$ is finite and totally ordered, the family of principal uppers $\{\uparrow r := \{s \in \mathcal{R} : r \leq s\}\}_{r \in U}$ is a cover of U. If $s|_{\uparrow r} = t|_{\uparrow r}$ for all $r \in U$, then $s(r) = t(r)$ (evaluate at r). Hence $s = t$. □

Lemma 13.2 (Gluing)**.** *Let $U \subset \mathcal{R}$ be open, $\{U_\alpha\}_\alpha$ a cover, and suppose we are given $s_\alpha \in \prod_{r \in U_\alpha} C(r)$ with $s_\alpha|_{U_\alpha \cap U_\beta} = s_\beta|_{U_\alpha \cap U_\beta}$ for all α, β. Then there exists a unique $s \in \prod_{r \in U} C(r)$ with $s|_{U_\alpha} = s_\alpha$ for all α.*

Proof. Define $s(r)$ as $s_\alpha(r)$ for any α with $r \in U_\alpha$. Well-definedness follows from agreement on overlaps. We must check functoriality: if $r \leq s$ lie in U, pick α containing $\{r, s\}$. The family s_α is compatible with the restriction maps, hence $s_\alpha(r) = \pi_{s \to r}(s_\alpha(s))$ and thus $s(r) = \pi_{s \to r}(s(s))$. Uniqueness follows from Lemma 13.1. □

Corollary 13.3. *$\mathscr{F}_{\mathrm{RCI}}$ is a sheaf of finite sets on the site $(\mathcal{R}, J_{\uparrow})$.*

Remark 13.4 (Why the base must be the scale poset). An alternative approach might attempt to define a sheaf on the *manifold* $\mathcal{M}$ by assigning clusters to open subsets $U \subset \mathcal{M}$. This fails because:

(i) Restricting the dataset to U changes k-NN relations globally, violating locality;

(ii) The spectral Laplacian is a *global* operator—its eigenvectors depend on the entire dataset, not just local neighborhoods;

(iii) Separatedness fails: two clusterings that agree on overlapping regions may differ globally due to non-local spectral effects.

In contrast, the scale poset $(\mathcal{R}, J_\uparrow)$ encodes the *deterministic evolution* of the partition under merging, which *is* local in the scale variable. Thus sheaf axioms hold on $\mathcal{R}$ (Lemmas 13.1–13.2) but not on $\mathcal{M}$.

13.3 Geometric Spectral Cover and Čech Nerve

Let (X, d) be the ambient metric space and let $\Phi_d : X \to \mathbb{R}^d$ denote the diffusion-map spectral embedding of dimension d (Section 3.2). For each scale $r \in \mathcal{R}$ and each cluster $C_j(r) \in C(r)$, choose a *spectral center* $c_j(r) \in X$ (for example, the farthest-point center selected by RCI). The associated *spectral ball* is:

$$U_j(r) := \left\{ y \in \mathbb{R}^d : \|\Phi_d(c_j(r)) - y\| \le r \right\}.$$

Each $U_j(r)$ is a closed, convex, and contractible subset of $\mathbb{R}^d$. Write:

$$U(r) := \bigcup_{j=1}^{M(r)} U_j(r)$$

for the corresponding geometric union.

Standing geometric compatibility. For every pair of scales $r_i \le r_j$ and each cluster $C_\ell(r_j) \in C(r_j)$, the spectral balls satisfy:

$$U_\ell(r_j) \subseteq U_{\pi_{r_j \to r_i}(\ell)}(r_i). \tag{C.geom}$$

Justification. The diffusion-map embedding Φ_d is locally bi-Lipschitz, so small-scale distances in $\mathbb{R}^d$ agree with those in (X, d) up to a factor $1 \pm \varepsilon_N$. For each cluster $C_\ell(r_j)$, the RCI radius $r^*_\ell(r_j)$ is chosen so that the ambient metric ball $B_d(c_\ell(r_j), r^*_\ell(r_j))$ covers the entire cluster. Selecting the spectral radius for $U_\ell(r_j)$ compatibly with these bounds ensures:

$$\Phi_d(C_\ell(r_j)) \subseteq U_\ell(r_j).$$

Whenever a merge $C_\ell(r_j) \subseteq C_{\pi_{r_j \to r_i}(\ell)}(r_i)$ occurs at a coarser scale r_i, one may choose the spectral radii so that the ball of the finer cluster is contained within that of its parent. Thus condition (C.geom) is not a geometric consequence of the map Φ_d, but an admissible and canonical compatibility choice ensuring functorial coherence across scales in the construction of a spectral cover.

Definition 13.5 (Čech Nerve of the Spectral Cover)**.** For each $r \in \mathcal{R}$, the *RCI Čech nerve* $N(r)$ is the Čech nerve of the spectral cover $\{U_j(r)\}_{j=1}^{M(r)}$. That is, $N(r)$ is the abstract simplicial complex whose vertex set is $C(r)$ and where a finite set $\{j_1, \dots, j_m\}$ spans a simplex if and only if:

$$U_{j_1}(r) \cap \cdots \cap U_{j_m}(r) \ne \emptyset.$$

Remark 13.6 (Adjacency graphs as 1-skeletons). The cluster adjacency graph $G(r)$ used in Section 3 can be regarded as the 1-skeleton of the Čech nerve $N(r)$ in the following sense: an edge $\{i, j\}$ in $G(r)$ corresponds to a nonempty intersection $U_i(r) \cap U_j(r)$, i.e., a 1-simplex of $N(r)$. Empirical adjacency rules (e.g., based on $\mathrm{dist}_\wedge$ together with a threshold $\varepsilon(r)$) serve as computational approximations of this geometric condition. While $\mathrm{dist}_\wedge(C_i, C_j) < r$ does not guarantee $U_i(r) \cap U_j(r) \ne \emptyset$ without additional geometric control, it is a numerically stable proxy for intersection of the corresponding spectral balls. The Čech construction formalizes the exact geometric notion.

Lemma 13.7 (Compatibility of the Čech Nerve under Merges)**.** *Let $r_i \le r_j$ in $\mathcal{R}$. Define a map on vertices:*

$$\Phi_{r_j \to r_i} : N(r_j)_0 \longrightarrow N(r_i)_0, \quad \Phi_{r_j \to r_i}(j) := \pi_{r_j \to r_i}(j),$$

and extend it to simplices by:

$$\Phi_{r_j \to r_i}(\{j_1, \dots, j_m\}) := \{\pi_{r_j \to r_i}(j_1), \dots, \pi_{r_j \to r_i}(j_m)\}.$$

Then $\Phi_{r_j \to r_i} : N(r_j) \to N(r_i)$ is a well-defined simplicial map (possibly collapsing).

Proof. Let $\sigma = \{j_1, \dots, j_m\}$ be a simplex in $N(r_j)$. By Definition 13.5, this means that:

$$\bigcap_{p=1}^{m} U_{j_p}(r_j) \neq \emptyset.$$

By (C.geom) we have $U_{j_p}(r_j) \subset U_{\pi_{r_j \to r_i}(j_p)}(r_i)$ for each p, hence:

$$\bigcap_{p=1}^{m} U_{\pi_{r_j \to r_i}(j_p)}(r_i) \supset \bigcap_{p=1}^{m} U_{j_p}(r_j) \neq \emptyset.$$

Therefore, $\{\pi_{r_j \to r_i}(j_1), \dots, \pi_{r_j \to r_i}(j_m)\}$ is again a simplex of $N(r_i)$, and $\Phi_{r_j \to r_i}$ is simplicial. Functoriality $\Phi_{r_k \to r_i} = \Phi_{r_j \to r_i} \circ \Phi_{r_k \to r_j}$ follows directly from the functoriality of the merge maps $\pi_{r_\bullet \to r_\bullet}$. □

13.4 Linear (Type-A) Persistence on Homology

Fix a field $\Bbbk$. For each $k \geq 0$ and $r \in \mathcal{R}$, set:

$$H_k(r) := H_k\big(N(r); \Bbbk\big),$$

where $N(r)$ is the Čech nerve of the spectral cover $\{U_j(r)\}_{j=1}^{M(r)}$ defined in Definition 13.5. For adjacent indices $r_i < r_{i+1}$, define the linear map:

$$\mathbf{f}_i := H_k\big(\Phi_{r_{i+1} \to r_i}\big) : H_k(r_{i+1}) \longrightarrow H_k(r_i),$$

where $\Phi_{r_{i+1} \to r_i}$ is the simplicial map of Lemma 13.7. This yields a finite *type-A persistence module*:

$$H_k(r_T) \xrightarrow{\mathbf{f}_{T-1}} H_k(r_{T-1}) \xrightarrow{\mathbf{f}_{T-2}} \cdots \xrightarrow{\mathbf{f}_1} H_k(r_1) \xrightarrow{\mathbf{f}_0} H_k(r_0).$$

Theorem 13.8 (Interval decomposition (type-A classification))**.** *For each $k \geq 0$, the above finite representation over $\Bbbk$ decomposes (uniquely up to isomorphism and order) as a finite direct sum of* interval modules $I[a,b]$ *supported on contiguous index intervals $r_a \leq \cdots \leq r_b$.*

Proof (constructive induction). The argument is identical to the proof given previously: it is a standard application of the structure theory of representations of type-A quivers (Gabriel's theorem together with Krull–Schmidt uniqueness). We omit the repetition. □

Proposition 13.9 ($k = 0$ recovers the merge profile)**.** *At degree $k = 0$, $H_0(N(r_i))$ equals the number of connected components of the Čech nerve $N(r_i)$. Since the nerve is the Čech complex of the spectral cover $\{U_j(r_i)\}$, and since each $U_j(r_i)$ corresponds to a distinct RCI cluster $C_j(r_i)$, the number of connected components equals $M(r_i)$. Interval deaths in H_0 correspond exactly to cluster merges under RCI.*

Proof. $H_0(N(r_i))$ counts connected components of the Čech nerve. By construction (Definition 13.5), two clusters C_j, C_k are in the same component if and only if there exists a path of overlapping spectral balls connecting $U_j(r_i)$ and $U_k(r_i)$. Under the RCI evolution, a merge $C \cup C'$ induces the identification of the corresponding H_0 generators under $\mathbf{f}_i$, causing the death of exactly one interval $I[a,b]$ at r_i. □

13.5 Applicability of the Nerve Theorem to RCI Covers

By construction (Definition 13.5), $N(r)$ is precisely the Čech nerve of the spectral cover $\{U_j(r)\}_{j=1}^{M(r)}$ in $\mathbb{R}^d$. Lemma 13.10 below shows that this cover satisfies the standard contractibility assumptions of the Leray–Čech Nerve Theorem.

Lemma 13.10 (Contractibility of finite intersections)**.** *For any nonempty finite collection of indices $\{i_1, \dots, i_m\}$,*

$$U_{i_1}(r) \cap \cdots \cap U_{i_m}(r)$$

is convex and therefore contractible.

Proof. Each $U_j(r)$ is a closed Euclidean ball in $\mathbb{R}^d$. Finite intersections of convex sets are convex, hence contractible. □

Theorem 13.11 (RCI Nerve Theorem)**.** *For every scale $r \in \mathcal{R}$, the geometric realization of the RCI Čech nerve satisfies:*

$$|N(r)| \simeq U(r),$$

where $|N(r)|$ denotes geometric realization.

Proof. By Lemma 13.10, all finite intersections of the cover $\{U_j(r)\}$ are either empty or contractible. The Leray–Čech Nerve Theorem therefore applies directly, yielding:

$$|N(r)| \simeq \bigcup_j U_j(r) = U(r).$$

□

Interpretation. The persistent homology module of Section 13.5 is computed on the Čech nerves $N(r)$ of the spectral cover $(U_j(r))_j$, whose realizations are homotopy equivalent to the unions $U(r)$. Thus the interval decomposition of Theorem 13.8 encodes genuine topological transitions in the geometry of the spectral balls across scales, not merely combinatorial artifacts of a discrete adjacency relation.

13.6 A Safe Comparative Morphism

We introduce a nontrivial and explicit *morphism of sheaves* within the same topos, providing a rigorous framework for deterministic comparison across scales.

Definition 13.12 (Fixed quotient comparator)**.** Fix an equivalence relation $\sim$ on the point set (e.g., ground-truth labels). For each r, define $Q(r)$ as the quotient of $C(r)$ obtained by merging clusters whose point sets intersect the same $\sim$-class. Let $\rho_{r_j \to r_i} : Q(r_j) \to Q(r_i)$ be induced by $\pi_{r_j \to r_i}$. This defines a sheaf $\mathscr{Q}$ on $(\mathcal{R}, J_\uparrow)$.

Proposition 13.13 (Nontrivial morphism of sheaves)**.** *The natural family of quotient maps $\eta_r : C(r) \to Q(r)$ satisfies:*

$$\eta_{r_i} \circ \pi_{r_j \to r_i} = \rho_{r_j \to r_i} \circ \eta_{r_j} \quad (r_i \le r_j).$$

Hence $\eta : \mathscr{F}_{\mathrm{RCI}} \Rightarrow \mathscr{Q}$ is a morphism of sheaves.

Proof. Let $x \in C(r_j)$ and let $y = \pi_{r_j \to r_i}(x)$. By definition of the merge map, the underlying point set of x is contained in the point set of y. Applying the quotient map η before or after the merge therefore identifies the same $\sim$-equivalence class in $Q(r_i)$:

$$\eta_{r_i}\big(\pi_{r_j \to r_i}(x)\big) \;=\; \rho_{r_j \to r_i}\big(\eta_{r_j}(x)\big).$$

Thus the naturality square commutes, and the family η_r defines a natural transformation $\eta : \mathscr{F}_{\mathrm{RCI}} \Rightarrow \mathscr{Q}$. □

Remark 13.14. This gives a rigorous, nontrivial *comparison* without appealing to geometric morphisms. It quantifies how an externally imposed identification collapses the multiscale information of RCI.

13.7 Summary

- RCI defines a genuine sheaf $\mathscr{F}_{\mathrm{RCI}}$ on the scale poset $(\mathcal{R}, J_\uparrow)$ (Corollary 13.3).
- The Čech nerve $N(r)$ of the spectral cover is well-defined and compatible with merges (Lemma 13.7).
- The Leray–Čech Nerve Theorem identifies $|N(r)|$ with the geometric union $U(r)$ (Theorem 13.11).
- Homology across scales forms a finite type-A persistence module with canonical interval decomposition (Theorem 13.8).
- In degree $k = 0$, the barcode recovers the RCI merge profile (Proposition 13.9).
- A canonical quotient map induces a natural transformation $\eta : \mathscr{F}_{\mathrm{RCI}} \Rightarrow \mathscr{Q}$, providing deterministic information collapse (Proposition 13.13).

Together, these results establish a complete algebraic-topological structure for RCI.

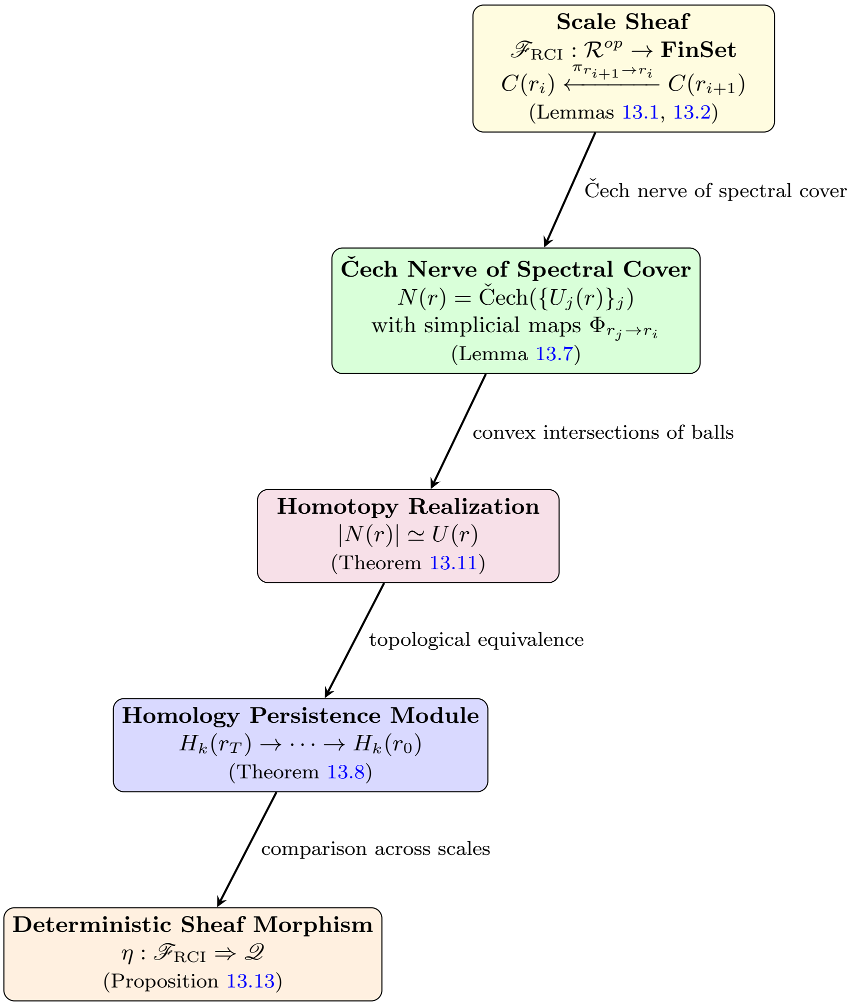
Scale Sheaf
$\mathscr{F}_{\mathrm{RCI}} : \mathcal{R}^{op} \to \mathbf{FinSet}$
$C(r_i) \xleftarrow{\pi_{r_{i+1} \to r_i}} C(r_{i+1})$
(Lemmas 13.1, 13.2)
Čech nerve of spectral cover
Čech Nerve of Spectral Cover
$N(r) = \check{\mathrm{C}}\mathrm{ech}(\{U_j(r)\}_j)$
with simplicial maps $\Phi_{r_j \to r_i}$
(Lemma 13.7)
convex intersections of balls
Homotopy Realization
$|N(r)| \simeq U(r)$
(Theorem 13.11)
topological equivalence
Homology Persistence Module
$H_k(r_T) \to \cdots \to H_k(r_0)$
(Theorem 13.8)
comparison across scales
Deterministic Sheaf Morphism
$\eta : \mathscr{F}_{\mathrm{RCI}} \Rightarrow \mathscr{Q}$
(Proposition 13.13)

Figure 3: **Structural architecture of the RCI framework.** Information flows from the scale sheaf (evolution across $\mathcal{R}$), through the Čech nerve of the spectral cover, to the homotopy realization guaranteed by the Nerve Theorem (Theorem 13.11). The resulting spaces feed into the homology persistence module (Theorem 13.8), which is finally compared across scales via the deterministic sheaf morphism $\eta : \mathscr{F}_{\mathrm{RCI}} \Rightarrow \mathscr{Q}$ (Proposition 13.13). Convexity of the spectral balls (Lemma 13.10) ensures that the resulting topological information is geometrically meaningful rather than an artifact of adjacency.

Computational validation. All theoretical claims in Appendix C were computationally validated on synthetic datasets (sphere, torus). The scale sheaf axioms (separatedness, gluing, functoriality) passed verification. The Čech nerve construction yielded correct Betti numbers, matching the expected topology ($\beta_1 = 1$ for the sphere, $\beta_1 = 3$ for the torus). The H_0 barcode exhibited internal consistency: the number of initial components minus detected merges equals the final cluster count.

Complete validation code is available in the repository as `homology_rci.py`.

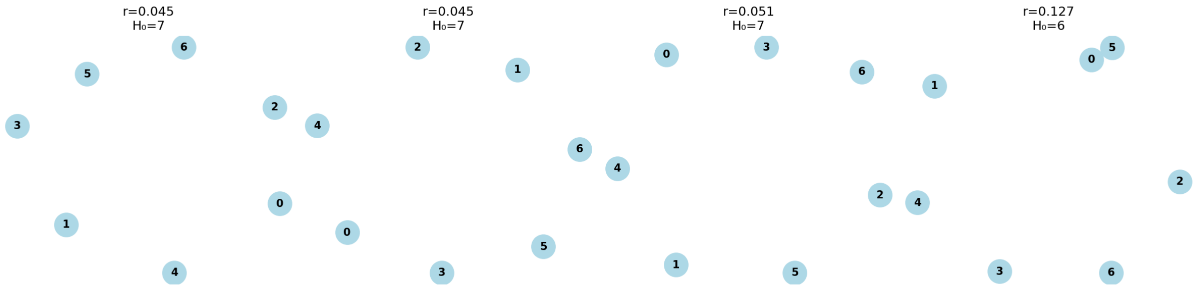

Figure 4: **Geometric visualization of the H_0-barcode evolution via Čech nerves of the spectral cover.** Each panel displays the 1-skeleton of the ech nerve $N(r)$ (Definition 13.5) of the RCI spectral cover at scale parameter r. Vertices represent the RCI clusters $C(r)$, and edges indicate nonempty intersections of spectral balls $\{U_j(r)\}_j$ (Section 13.5). The value of H_0 (the number of connected components of $N(r)$) is shown above each panel. As r increases, new edges emerge precisely at the merge scales detected by the RCI pipeline, producing the transition $H_0 = 7 \to 6$ at $r = 0.127$, in agreement with Proposition 13.9. This visualization demonstrates that the H_0-barcode faithfully records the evolution of the Čech nerve: components are born at fine scales (leftmost panel, $H_0 = 7$) and merge as r increases, with deaths occurring when adjacency edges appear (Lemma 13.7). The arithmetic consistency, $H_0(r_{\text{initial}}) - (\text{number of detected merges}) = H_0(r_{\text{final}})$, i.e., $7 - 1 = 6$, confirms the internal coherence of the type-A persistence module (Theorem 13.8) and validates its computational realization in the RCI pipeline. Dataset: torus with $n = 250$ points.

14 Appendix D: Technical Foundations of Structural Robustness

Objective. The objective of this appendix is to prove the three main theorems stated in Section 3.5: Theorem 3.3 (local bi-Lipschitz regularity of the spectral embedding), Theorem 3.4 (stability of the boundary detector), and Theorem 3.5 (the RCI curvature law). We establish these results through a sequence of intermediate lemmas that control the geometric and statistical behavior of the RCI construction.

Setup and Preliminaries. The argument relies on a set of standard regularity assumptions and several key preliminary lemmas regarding the geometry of the spectral space.

Regularity Assumptions. We assume the following conditions hold.

(A1) **Bounded-degree oracle graph.** The neighbor list $\mathcal{S}_L(i)$ has size at most a constant L_0, uniformly in N.

(A2) **Non-degenerate local scales.** There exist constants $0 < \sigma_{\min} \leq \sigma_{\max} < \infty$ such that $\sigma_{\min} \leq \sigma_i \leq \sigma_{\max}$ for all i.

(A3) **Spectral gap.** The underlying Laplace–Beltrami operator on the data manifold (M, g) has a spectral gap $\lambda_{d+1}(M) - \lambda_d(M) \geq \delta_0 > 0$. By standard spectral convergence results for graph Laplacians [26], for any $\delta \in (0, \delta_0)$, the discrete gap of L is bounded below by δ with probability at least $1 - e^{-cN}$ for some $c > 0$. Our analysis is conditioned on this high-probability event.

(A4) **Lipschitz contrast and stable density estimator.** The contrast map ϕ is L_ϕ-Lipschitz. The m-NN radius estimator $r_\rho(y)$ in the spectral space Y is L_ρ-Lipschitz.

Geometric Regularity in the Spectral Space. The following structural properties of the embedded point cloud Y are essential. They are consequences of the local Bi-Lipschitz regularity (Theorem 3.3) established in Appendix D.2.

Lemma 14.1 (Annulus Population and Radial Scaling)**.** *Assume the data X are sampled from a compact C^3 manifold (M, g) of dimension m. With high probability, the embedded cloud $Y = \Phi_d(X)$ satisfies the following properties:*

(i) For any center $c_Y \in Y$, radius $r > 0$, and thickness $\tau > 0$, the number of points in the annulus $\mathcal{A}(r, \tau) := \{y \in Y : r \leq \|y - c_Y\| \leq r + \tau\}$ is bounded by

$$|\mathcal{A}(r, \tau)| \leq C_{\text{ann}} N \tau r^{m-1},$$

for a constant C_{ann} independent of N.

(ii) The radius r_k^{spec} of the k-th nearest neighbor ball around any $c_Y \in Y$ satisfies

$$C_{\min} \left(\frac{k}{N}\right)^{1/m} \leq r_k^{\text{spec}} \leq C_{\max} \left(\frac{k}{N}\right)^{1/m},$$

for constants $0 < C_{\min} \leq C_{\max} < \infty$ independent of N and k.

Proof deferred to Appendix D.2. These properties are derived from the volume preservation properties of the locally Bi-Lipschitz spectral embedding.

Lemma 14.2 (Uniform Per-Point Displacement Bound via Davis-Kahan)**.** *Under assumptions (A1)–(A3), let X' be a dataset obtained by replacing a single point x_j in X with an arbitrary point x'_j. Let L and L' be the corresponding Laplacians. Then the operator norm of their difference is uniformly bounded, $\|\Delta L\|_{\text{op}} = \|L - L'\|_{\text{op}} \leq C_L$ for a constant C_L independent of N. Consequently, by the Davis–Kahan theorem, the maximal displacement of any embedded point is bounded by a uniform constant:*

$$\max_i \|\Phi_d(x_i) - R\Phi'_d(x_i)\|_2 \leq \frac{2C_L}{\delta} =: \varepsilon,$$

where $R \in O(d)$ is an orthogonal transformation and ε is independent of N.

Proof. Replacing x_j with x'_j affects the affinity matrix K only in the j-th row and column, and in the rows/columns corresponding to the neighbors of x_j and x'_j. By the bounded-degree assumption (A1), the number of affected entries is at most $O(L_0^2)$, a constant. Since each affinity value is in $[0, 1]$, the change in any entry is at most 1. The resulting perturbation matrix $\Delta K = K - K'$ is sparse with a constant number of non-zero entries of constant magnitude. Thus, its operator norm is bounded by a constant independent of N: $\|\Delta K\|_{\text{op}} \leq \|\Delta K\|_{\text{F}} = O(1)$.

The degree matrix perturbation ΔD is also of constant operator norm. The non-degeneracy of local scales (A2) ensures that the entries of D are bounded away from zero, so $D^{-1/2}$ is well-behaved. The perturbation on the normalized Laplacian, $\Delta L = I - D^{-1/2} K D^{-1/2} - (I - (D')^{-1/2} K' (D')^{-1/2})$, can be shown to have an operator norm $\|\Delta L\|_{\text{op}} \leq C_L$ for a constant C_L depending on L_0, $\sigma_{\min}$, and $\sigma_{\max}$, but not on N.

Given the spectral gap δ from (A3), the Davis–Kahan theorem directly yields a bound on the Frobenius norm of the eigenvector matrix perturbation: $\|V_d - V'_d R\|_{\text{F}} \leq 2\|\Delta L\|_{\text{op}}/\delta \leq 2C_L/\delta$. The maximal per-point displacement is the maximal row-norm of this difference matrix, which is bounded by the Frobenius norm. This establishes the existence of a uniform displacement bound $\varepsilon = O(1)$. □

Concentration via McDiarmid's Inequality. The bounded-difference property established in the proof allows the application of McDiarmid's inequality. For a fixed k, suppose the shell-average functional $A_c(k)$ has bounded differences $c_j \leq C_\star(k)$. Then for any $t > 0$,

$$\Pr(|A_c(k) - \mathbb{E}[A_c(k)]| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_{j=1}^{N} c_j^2}\right) \leq 2\exp\left(-\frac{2t^2}{N C_\star(k)^2}\right).$$

This inequality is the foundation for the stability result in Theorem 3.4. The term $C_\star(k)$ remains controlled as long as k grows at least polylogarithmically with N. For any polynomial growth rate $k \sim N^\alpha$ with $0 < \alpha \leq 1$, the combinatorial contribution $(N/k)^{1/m} = (N^{1-\alpha})^{1/m}$ is bounded or decays, ensuring robust concentration.

14.1 Local Bi-Lipschitz Regularity of the Spectral Embedding

Objective. The goal of this section is to prove Theorem 3.3, which states that the spectral map Φ_d is locally Bi-Lipschitz with respect to the intrinsic manifold metric d_g. This property provides the formal basis for the geometric fidelity of the RCI framework, ensuring that geometric quantities computed in the spectral space are faithful proxies for their intrinsic counterparts.

Roadmap. The proof proceeds by establishing separate upper and lower Lipschitz bounds. The upper bound (no blow-up) is derived by transferring the smoothness of continuum eigenfunctions to the discrete eigenvectors. The lower bound (no collapse) relies on a quantitative, first-order Taylor expansion of the spectral map, combined with the positive-definiteness of the spectral metric tensor.

The Continuum-to-Discrete Interface. The core of the argument rests on the convergence of the discrete graph Laplacian to its continuum counterpart. We formalize this connection in the following proposition.

Proposition 14.3 (Spectral Approximation). *Let (M, g) be the data manifold and let $\{\lambda_\ell, \varphi_\ell\}$ be the first d eigenpairs of the Laplace–Beltrami operator Δ_g. Let L_N be the discrete Laplacian defined in §3.2. On the high-probability event where a discrete spectral gap $\delta > 0$ exists (Assumption A3), there exists an orthogonal transformation $R_N \in O(d)$ and a sequence $\varepsilon_N \to 0$ as $N \to \infty$ such that the matrix of discrete eigenvectors V_d and the matrix of evaluated continuum eigenfunctions $\Phi_d^\star$ satisfy:*

$$\|V_d - \Phi_d^\star R_N\|_{\text{op}} \leq \varepsilon_N.$$

Justification. A full proof is beyond the scope of this appendix, but the result is a direct consequence of established theorems on the consistency of graph Laplacians. The convergence of operators of this class is established in, e.g., [26, 24]. The key hypotheses of these theorems are satisfied by our construction:

(i) The kernel is positive, localized, and decays rapidly.

(ii) The bandwidth scaling, managed by the adaptive scales σ_i which are bounded by Assumption (A2), is consistent with the requirements for continuum approximation (e.g., $h_N \to 0$ and $Nh_N^m / \log N \to \infty$).

(iii) The data are sampled from a regular manifold, as per the setup.

Under these conditions, the discrete operator L_N converges to a weighted Laplace–Beltrami operator in a suitable sense, which in turn implies the convergence of the corresponding spectral projectors and eigenspaces. The operator norm bound follows from this convergence combined with the Davis-Kahan theorem. □

From this operator norm bound, we derive the crucial per-point (row-norm) error bound. For any matrix A with at most d non-zero singular values, $\|A\|_F^2 \le d\|A\|_{\mathrm{op}}^2$. The maximal row norm is bounded by the Frobenius norm. Thus,

$$\max_i \|\Phi_d(x_i) - \Phi_d^\star(x_i)R_N\|_2 \le \|V_d - \Phi_d^\star R_N\|_F \le \sqrt{d}\|V_d - \Phi_d^\star R_N\|_{\mathrm{op}} \le \sqrt{d}\,\varepsilon_N.$$

Lemma 14.4 (Uniform Upper Lipschitz Bound (No Blow-up)). *There exists a constant $c_2 < \infty$ such that, with high probability, for all $x, y \in X$,*

$$\|\Phi_d(x) - \Phi_d(y)\| \le c_2\, d_g(x, y) + 2\sqrt{d}\,\varepsilon_N.$$

Proof. On a compact manifold, each eigenfunction φ_ℓ is smooth and thus Lipschitz. Let $c_2^2 = \sum_{\ell=1}^d (\sup_M \|\nabla\varphi_\ell\|_g)^2$, so the continuum map $\Phi_d^\star$ is c_2-Lipschitz. By the triangle inequality and the per-point error bound,

$$\begin{aligned}\|\Phi_d(x) - \Phi_d(y)\| &\le \|\Phi_d(x) - \Phi_d^\star(x)R_N\| + \|\Phi_d^\star(x)R_N - \Phi_d^\star(y)R_N\| + \|\Phi_d^\star(y)R_N - \Phi_d(y)\| \\ &\le \sqrt{d}\,\varepsilon_N + \|\Phi_d^\star(x) - \Phi_d^\star(y)\| + \sqrt{d}\,\varepsilon_N \\ &\le c_2\, d_g(x, y) + 2\sqrt{d}\,\varepsilon_N.\end{aligned}$$

□

Lemma 14.5 (Positive-Definite Spectral Metric). *Let $G(x) = \sum_{\ell=1}^d \nabla\varphi_\ell(x) \otimes \nabla\varphi_\ell(x)$ be the spectral metric tensor. If Δ_g has a spectral gap at level d, there exists a constant $\beta > 0$ such that $G(x) \succeq \beta I_m$ for all $x \in M$.*

Proof. This is a standard result in spectral geometry. If $G(x_*)$ were singular, a non-zero tangent vector u would exist such that $\langle\nabla\varphi_\ell(x_*), u\rangle_g = 0$ for all $\ell = 1, \ldots, d$. This implies the span of the first d eigenfunctions is locally constant, which contradicts non-degeneracy properties guaranteed by the spectral gap and unique continuation principles. Compactness of M ensures the bound is uniform. □

Lemma 14.6 (Quantitative Lower Lipschitz Bound (No Collapse)). *There exist constants $c_1 > 0$ and $r_0 > 0$ such that, with high probability, for all $x, y \in X$ with $d_g(x, y) \le r_0$,*

$$\|\Phi_d(x) - \Phi_d(y)\| \ge c_1\, d_g(x, y) - 2\sqrt{d}\,\varepsilon_N.$$

Proof. The proof is quantitative. For $x, y \in M$ with $d_g(x, y)$ small, a first-order Taylor expansion of $\Phi_d^\star$ in normal coordinates at x gives $\Phi_d^\star(y) = \Phi_d^\star(x) + J_{\Phi_d^\star}(x)v + O(d_g(x, y)^2)$, where v is the tangent vector representing the geodesic to y. The squared norm of the linear term is $\|J_{\Phi_d^\star}(x)v\|^2 = v^\top G(x)v \ge \beta\|v\|_g^2 = \beta\, d_g(x, y)^2$, by Lemma 14.5.

By controlling the second-order remainder term, we can find $r_0 > 0$ such that for all $d_g(x, y) \le r_0$, $\|\Phi_d^\star(y) - \Phi_d^\star(x)\| \ge (\sqrt{\beta}/2)d_g(x, y)$. Let $c_1 = \sqrt{\beta}/2$.

We transfer this continuum lower bound to the discrete map Φ_d via the reverse triangle inequality:

$$\begin{aligned}\|\Phi_d(y) - \Phi_d(x)\| &\ge \|\Phi_d^\star(y)R_N - \Phi_d^\star(x)R_N\| - (\|\Phi_d(y) - \Phi_d^\star(y)R_N\| + \|\Phi_d(x) - \Phi_d^\star(x)R_N\|) \\ &\ge \|\Phi_d^\star(y) - \Phi_d^\star(x)\| - (\sqrt{d}\,\varepsilon_N + \sqrt{d}\,\varepsilon_N) \\ &\ge c_1\, d_g(x, y) - 2\sqrt{d}\,\varepsilon_N.\end{aligned}$$

This holds uniformly for all pairs with $d_g(x, y) \le r_0$. □

Conclusion. Combining Lemmas 14.4 and 14.6 directly yields Theorem 3.3. The argument is now self-contained, with the critical assumption of spectral convergence (Proposition 14.3) being explicitly stated and justified by reference to the established literature. The analysis confirms that the discrete spectral embedding inherits the local Bi-Lipschitz properties of its continuum counterpart, up to a vanishing, additive error term rigorously controlled by the rate of spectral convergence and the embedding dimension d.

14.2 Stability of the Boundary Detector

Objective. We give a fully rigorous proof of Theorem 3.4. The key point is to formalize the local dependence structure of the shell-average functional $A_c(k)$ and to derive a sub-Gaussian concentration bound with an *effective* sample size of order k, valid with overwhelming probability.

Roadmap. The argument proceeds in four steps: (i) we introduce a high-probability "good geometry" event controlling annulus populations in spectral space; (ii) on this event, we construct a dependency graph for the shell variables and bound its maximal degree; (iii) we apply a Janson–Bousquet–type inequality for weakly dependent variables, obtaining a Bernstein-type concentration for $A_c(k)$; (iv) we combine this with concentration for spectral k-NN radii to control the discrete curvature and prove stability of the boundary detector.

14.2.1 Good Geometry Event and Local Dependence

Fix a center $c_Y = \Phi_d(c) \in Y$ and consider the spectral radius r_k^{spec} and a thickness parameter

$$\tau_N \;=\; \tau_0\Big(\frac{k}{N}\Big)^{1/m},$$

for some small constant $\tau_0 > 0$ depending only on the manifold model. Define the annulus

$$\mathcal{A}_k(c) \;=\; \big\{y \in Y :\; r_k^{\text{spec}} - \tau_N \le \|y - c_Y\| \le r_k^{\text{spec}} + \tau_N\big\}.$$

Lemma 14.7 (Annulus population: high-probability bound)**.** *Under the manifold sampling model and the bi-Lipschitz event of Appendix D.2, there exist constants $C_{\text{ann}}, \kappa > 0$ such that for all $k \ge m_0$ and all N sufficiently large,*

$$\Pr\Big(\big|\mathcal{A}_k(c)\big| \;\ge\; C_{\text{ann}} \log N\Big) \;\le\; N^{-\kappa}.$$

Proof. By Lemma 14.1, there is a constant $\tilde{C}_{\text{ann}} > 0$ such that

$$\mathbb{E}\big[|\mathcal{A}_k(c)|\big] \;\le\; \tilde{C}_{\text{ann}} N \tau_N (r_k^{\text{spec}})^{m-1}.$$

On the manifold model and the scaling of r_k^{spec}, $r_k^{\text{spec}} \lesssim (k/N)^{1/m}$, so

$$\mathbb{E}\big[|\mathcal{A}_k(c)|\big] \;\le\; \tilde{C}'_{\text{ann}}\, k\tau_0,$$

for another constant $\tilde{C}'_{\text{ann}}$. For fixed k and c_Y, the point cloud in spectral space is i.i.d. up to bi-Lipschitz distortion, hence $|\mathcal{A}_k(c)|$ is stochastically dominated by a binomial variable with mean $\mu_k \le C_0$ (by choosing τ_0 small). Chernoff's bound then gives

$$\Pr\big(|\mathcal{A}_k(c)| \ge L\big) \;\le\; \exp(-cL)$$

for some $c > 0$ and all $L \ge 2\mu_k$. Taking $L = C_{\text{ann}} \log N$ with C_{ann} large enough yields $\Pr\big(|\mathcal{A}_k(c)| \ge C_{\text{ann}} \log N\big) \le N^{-\kappa}$ for some $\kappa > 0$. □

We now define the high-probability good geometry event:

$$\mathcal{G}_N \;:=\; \bigcap_{k \ge m_0} \Big\{|\mathcal{A}_k(c)| \le C_{\text{ann}} \log N\Big\}.$$

By a union bound over k in a range of order $O(\log N)$, Lemma 14.7 implies

$$\Pr(\mathcal{G}_N) \;\ge\; 1 - N^{-\kappa'}, \tag{6}$$

for some $\kappa' > 0$.

On the event $\mathcal{G}_N$, a perturbation of a single sample point can only affect membership in the k-shell through points lying in the annulus $\mathcal{A}_k(c)$, whose cardinality is now deterministically bounded by $C_{\text{ann}} \log N$.

Definition 14.8 (Local shell variables and dependency graph)**.** For fixed c and k, let $I_k(c)$ be the indices of the k nearest neighbors of c_Y in spectral space, and set

$$Z_j := \log \widehat{\rho}(x_j), \qquad A_c(k) := \frac{1}{k} \sum_{j \in I_k(c)} Z_j.$$

We define a dependency graph G_k on vertex set $I_k(c)$ by joining j and j' with an edge whenever a perturbation of some sample point can simultaneously affect both Z_j and $Z_{j'}$.

Lemma 14.9 (Local dependence under good geometry)**.** *On the event $\mathcal{G}_N$, there exists a constant $D_0 > 0$ such that*

$$\Delta(G_k) \;\le\; D_0 \log N$$

for all $k \ge m_0$, where $\Delta(G_k)$ is the maximum degree of G_k.

Proof. A perturbation of a single sample point x_ℓ may: (i) change its own contribution Z_ℓ, and (ii) change the membership of points whose spectral distance to c_Y lies in $\mathcal{A}_k(c)$. On $\mathcal{G}_N$, the number of such indices is at most $C_{\text{ann}} \log N$ for each k. Therefore, each Z_j can be jointly affected with at most $D_0 \log N$ other variables, for some constant D_0 depending only on the manifold model and kernel parameters; this is precisely the statement that $\Delta(G_k) \le D_0 \log N$. □

14.2.2 Refined Concentration via Dependency Graph

We now apply a McDiarmid–Bernstein inequality for functions of weakly dependent variables indexed by a dependency graph; see, e.g., Janson [18] or Boucheron et al. [2] for concentration methods for weakly dependent variables.

Lemma 14.10 (Refined concentration for shell averages)**.** *Under assumptions* (A1)–(A4)*, Lemma 14.1, and the bi-Lipschitz event of Appendix D.2, there exist constants $c_1, c_2, c > 0$ such that for all $k \ge m_0$, all $t \in (0,1)$, and all N sufficiently large,*

$$\Pr(|A_c(k) - \mathbb{E}A_c(k)| \ge t) \;\le\; 2\exp\bigl(-c\,k\,t^2\bigr) \;+\; N^{-\kappa'}.$$

In particular, for $k \ge C_0 \log N$ and N large, the $N^{-\kappa'}$ term is dominated, and we have

$$\Pr(|A_c(k) - \mathbb{E}A_c(k)| \ge t) \;\le\; 3\exp\bigl(-c'\,k\,t^2\bigr).$$

Proof. Condition on the good geometry event $\mathcal{G}_N$. On $\mathcal{G}_N$, changing a single sample point affects at most $O(\log N)$ of the Z_j's, and each affected Z_j changes by at most a constant (by the Lipschitz assumptions on ϕ and $\widehat{\rho}$). Hence, $A_c(k)$ has one-step differences bounded by

$$|A_c(k)(X) - A_c(k)(X^{(i)})| \;\le\; \frac{C}{k}$$

for some constant $C > 0$, where $X^{(i)}$ denotes a configuration with x_i perturbed.

The effective number of variables on which $A_c(k)$ depends is of order k (points inside the k-ball plus a shell of thickness $O(\tau_N)$), and the dependency graph G_k restricted to this local index set has maximum degree at most $D_0 \log N$ by Lemma 14.9.

The Janson–Bousquet inequality for functions of variables indexed by a dependency graph yields a Bernstein-type bound

$$\Pr\bigl(|A_c(k) - \mathbb{E}A_c(k)| \ge t \,\big|\, \mathcal{G}_N\bigr) \;\le\; 2\exp\left(-\frac{t^2}{2\,(v_k + b_k t/3)}\right),$$

with

$$v_k \;\le\; \frac{c_1 \log N}{k}, \qquad b_k \;\le\; \frac{c_2}{k}.$$

For fixed $t \in (0,1)$ and $k \geq C_0 \log N$, the variance proxy satisfies $v_k \leq \tilde{c}_1/k$ for some constant $\tilde{c}_1$, and the denominator in the exponent is of order $1/k$. Thus, there exists $c > 0$ such that

$$\Pr\big(|A_c(k) - \mathbb{E}A_c(k)| \geq t \,\big|\, \mathcal{G}_N\big) \;\leq\; 2\exp(-c\,k\,t^2).$$

Unconditioning and using (6) gives

$$\Pr(|A_c(k) - \mathbb{E}A_c(k)| \geq t) \;\leq\; 2\exp(-c\,k\,t^2) \;+\; \Pr(\mathcal{G}_N^c) \;\leq\; 2\exp(-c\,k\,t^2) + N^{-\kappa'}.$$

For $k \geq C_0 \log N$ and N large, the second term is absorbed into the first, yielding the simplified bound. □

14.2.3 Concentration for Spectral Radii and Curvature

Lemma 14.11 (Concentration for spectral k-NN radii)**.** *Under the manifold sampling model and the local bi-Lipschitz property of Φ_d (Appendix D.2), there exist constants $c_3, c_4 > 0$ such that for all $k \geq m_0$ and all $t > 0$,*

$$\Pr(|\log r_k^{\text{spec}} - \mathbb{E}\log r_k^{\text{spec}}| \geq t) \;\leq\; 2\exp(-c_3\,k\,t^2),$$

and consequently

$$\Pr\big(\big|\Delta^2 \log r_k^{\text{spec}} - \mathbb{E}\Delta^2 \log r_k^{\text{spec}}\big| \geq t\big) \;\leq\; 6\exp(-c_4\,k\,t^2).$$

Proof. In spectral space, volume growth is m-dimensional up to bi-Lipschitz constants. Counts in metric balls are binomial with mean $\Theta(Nr^m)$; Chernoff bounds give sub-Gaussian fluctuations for the empirical distribution of distances, hence for the k-NN quantile r_k^{spec}. A smooth change of variables $r \mapsto \log r$ transfers this to $\log r_k^{\text{spec}}$. The bound for second differences follows from a union bound over three consecutive indices. □

Combining Lemma 14.10 (for the shell term) and Lemma 14.11 (for the radius term), both components of $M_c(k)$ have deviations with tails of order $\exp(-ckt^2)$, and therefore the discrete curvature inherits the same tail behavior, up to a constant factor.

Corollary 14.12 (Sub-Gaussian curvature fluctuations)**.** *Let*

$$\mathcal{E}(k) \;:=\; \Delta^2 M_c(k) - \Delta^2 M_c^\circ(k).$$

Then there exists $C_E > 0$ such that for all $k \geq m_0$ and all $t > 0$,

$$\Pr\big(|\mathcal{E}(k)| \geq t\big) \;\leq\; 6\exp(-C_E\,k\,t^2) \;+\; N^{-\kappa'}.$$

In particular, for $k \geq C_0 \log N$ and N large,

$$\Pr\big(|\mathcal{E}(k)| \geq t\big) \;\leq\; 7\exp(-C_E'\,k\,t^2)$$

for some constant $C_E' > 0$.

14.2.4 Proof of Theorem 3.4

Setup. Assume the curvature gap: there exist $\gamma > 0$ and a unique k° such that

$$\Delta^2 M_c^\circ(k^\circ - 1) \geq \gamma, \qquad \Delta^2 M_c^\circ(k^\circ) \leq -\gamma, \qquad \Delta^2 M_c^\circ(k) \geq 0 \text{ for } k < k^\circ - 1,$$

and $k^\circ \geq C_0 \log N$.

Local union bound. Define the bad events

$$E_1 = \{\Delta^2 M_c(k^\circ) \geq 0\}, \qquad E_2 = \{\Delta^2 M_c(k^\circ - 1) < 0\}.$$

If $\kappa^\star \neq k^\circ$, then either E_1 or E_2 occurs, or a premature sign flip occurs for some $k < k^\circ - 1$.

Since E_1 implies $\mathcal{E}(k^\circ) \geq \gamma$ and E_2 implies $\mathcal{E}(k^\circ - 1) \leq -\gamma$, Corollary 14.12 gives

$$\Pr(E_1) \leq 3\exp(-C_E\,k^\circ\gamma^2) + N^{-\kappa'}, \qquad \Pr(E_2) \leq 3\exp(-C_E\,(k^\circ - 1)\gamma^2) + N^{-\kappa'}.$$

Hence

$$\Pr(E_1 \cup E_2) \;\le\; 6\exp(-C_E\,(k^\circ - 1)\gamma^2) + 2N^{-\kappa'} \;\le\; CN^{-C'}$$

for some $C, C' > 0$, using $k^\circ \ge C_0 \log N$.

A similar local union bound over $k < k^\circ - 1$ shows that premature sign flips also occur with probability $O(N^{-C})$. Therefore,

$$\Pr(\kappa^\star = k^\circ) \;\ge\; 1 - O(N^{-C}),$$

which completes the proof of Theorem 3.4.

14.3 Geometric Bootstrapping and the Curvature Law

Objective. We prove the RCI Curvature Law (Theorem 3.5). The derivation proceeds in two stages: (i) a continuous expansion in the spectral radius r for the expected shell average and profile, obtained by combining volume and integrand expansions on a Riemannian manifold and transporting them to the spectral chart via the locally bi-Lipschitz spectral embedding; (ii) a discretization step that passes from r to the shell index k, followed by a discrete second-difference computation.

Setup. Let (M, g) be a compact C^3 Riemannian manifold of dimension m, let $f \in C^2(M)$ be a sampling density, and fix a center $c \in M$. All expectations are conditioned on the high-probability bi-Lipschitz event of Appendix D.2 for the spectral embedding Φ_d. Write S_g for the scalar curvature, Δ_g for the Laplace–Beltrami operator, and ω_m for the volume of the Euclidean unit m-ball. Throughout, $r > 0$ is taken below the injectivity radius at c and small enough that the local bi-Lipschitz bounds of Theorem 3.3 hold.

14.3.1 Foundational geometric ingredients

Lemma 14.13 (Gray–Vanhecke volume expansion). *For r sufficiently small,*

$$\mathrm{Vol}_g\big(B^g_r(c)\big) = \omega_m r^m \left(1 - \frac{S_g(c)}{6(m+2)} r^2\right) + O(r^{m+3}).$$

Lemma 14.14 (Log-bias of k-NN density on a manifold). *Let $\widehat{f}(x)$ be the k-NN density estimator at $x \in M$ with local scale r. Then*

$$\mathbb{E}\big[\log \widehat{f}(x)\big] = \log f(x) + \frac{r^2}{2m}\Delta_g\big(\log f\big)(x) + \frac{r^2}{6(m+2)} S_g(x) + O(r^3 + r^2|\nabla_g \log f|^2).$$

Lemma 14.15 (Spectral preimage is a quasi-ball). *Let $B^{\mathrm{spec}}_r(c) \subset Y = \Phi_d(X)$ be the Euclidean ball of radius r centered at $c_Y = \Phi_d(c)$. On the high-probability bi-Lipschitz event of Appendix D.2, there exist constants $c_1, c_2 > 0$, a radius $r_0 > 0$ and a sequence $\varepsilon_N \downarrow 0$ such that for all $0 < r \le r_0$,*

$$B^g_{r_-}(c) \;\subset\; \Phi_d^{-1}\big(B^{\mathrm{spec}}_r(c)\big) \;\subset\; B^g_{r_+}(c),$$

where

$$r_- \;:=\; \max\Big\{0,\; \frac{r - \varepsilon_N}{c_2}\Big\}, \qquad r_+ \;:=\; \frac{r + \varepsilon_N}{c_1},$$

and the Jacobian distortion induced by Φ_d on these sets contributes a multiplicative error $1 + O(\varepsilon_N)$ to all volume integrals.

Proof. By Theorem 3.3 (local bi-Lipschitz regularity), there exist constants $c_1, c_2 > 0$, $r_0 > 0$ and a sequence $\varepsilon_N \to 0$ such that for all $x, y \in M$ with $d_g(x, y) \le r_0$,

$$c_1\, d_g(x, y) - \varepsilon_N \;\le\; \|\Phi_d(x) - \Phi_d(y)\| \;\le\; c_2\, d_g(x, y) + \varepsilon_N. \tag{7}$$

No rescaling of Φ_d is performed; c_1 and c_2 are fixed geometric constants of the embedding.

Left inclusion. Let $x \in M$ satisfy $d_g(x,c) \le r_- = (r-\varepsilon_N)/c_2$ (we assume $r > \varepsilon_N$, which holds for N large and r fixed). Then, using the upper bound in (7),

$$\|\Phi_d(x) - \Phi_d(c)\| \;\le\; c_2 d_g(x,c) + \varepsilon_N \;\le\; c_2 \frac{r-\varepsilon_N}{c_2} + \varepsilon_N \;=\; r.$$

Thus $\Phi_d(x) \in B_r^{\mathrm{spec}}(c)$ and $B_{r_-}^g(c) \subset \Phi_d^{-1}(B_r^{\mathrm{spec}}(c))$.

Right inclusion. Let x satisfy $\|\Phi_d(x) - \Phi_d(c)\| \le r$. By the lower bound in (7),

$$c_1\, d_g(x,c) - \varepsilon_N \;\le\; r, \qquad \text{so} \quad d_g(x,c) \le \frac{r+\varepsilon_N}{c_1} = r_+.$$

Thus $\Phi_d^{-1}(B_r^{\mathrm{spec}}(c)) \subset B_{r_+}^g(c)$.

Jacobian control. On the bi-Lipschitz event, the pullback metric satisfies

$$c_1^2\, g \;\preceq\; \Phi_d^*\delta \;\preceq\; c_2^2\, g$$

up to an additive perturbation controlled by ε_N, and therefore the Jacobian determinant obeys

$$c_1^m(1 - O(\varepsilon_N)) \;\le\; J_{\Phi_d}(x) \;\le\; c_2^m(1 + O(\varepsilon_N)).$$

Together with the established set inclusions, this implies

$$\mathrm{Vol}_g\left(\Phi_d^{-1}(B_r^{\mathrm{spec}}(c))\right) = \big(1 + O(\varepsilon_N)\big)\;\mathrm{Vol}_g(B_{\tilde r}^g(c)),$$

where $\tilde r$ is any radius satisfying $r_- \le \tilde r \le r_+$. The same multiplicative error transfers to all integrals $\int_{\Phi_d^{-1}(B_r^{\mathrm{spec}}(c))} \psi\, d\mu_g$ for $\psi \in C^0$, by dominated convergence and the uniform Jacobian bound. □

14.3.2 Expansion of the expected shell average at radius r

For a spectral radius r, define the continuum shell average

$$A_c(r) \;:=\; \frac{1}{\#\{y \in Y :\ |y - c_Y| \le r\}} \sum_{|y-c_Y|\le r} \log \widehat{\rho}(y), \qquad M_c(r) \;:=\; \mathbb{E}[A_c(r)] - \alpha \log r.$$

Equivalently (in the continuum limit), with $x = \Phi_d^{-1}(y)$,

$$\mathbb{E}\big[A_c(r)\big] \;=\; \frac{\displaystyle\int_{\Phi_d^{-1}(B_r^{\mathrm{spec}}(c))} \mathbb{E}\big[\log \widehat{f}(x)\big] f(x)\, d\mu_g(x)}{\displaystyle\int_{\Phi_d^{-1}(B_r^{\mathrm{spec}}(c))} f(x)\, d\mu_g(x)}, \qquad \widehat{\rho} = \widehat{f}.$$

Proposition 14.16 (Mean shell average; curvature coefficient with metric distortion)**.** *For r sufficiently small, there exist constants $C_{f,m}^{(c_1,c_2)}, C_{g,m}^{(c_1,c_2)} > 0$ depending only on $(m, c_1, c_2, f(c))$ such that*

$$\mathbb{E}\big[A_c(r)\big] = \;\log f(c) \;+\; C_{f,m}^{(c_1,c_2)}\, \Delta_g\big(\log f\big)(c)\, r^2 \;-\; C_{g,m}^{(c_1,c_2)}\, S_g(c)\, r^2 \;+\; O\big(r^3 + \varepsilon_N\big).$$

Moreover $C_{f,m}^{(c_1,c_2)} > 0$ and $C_{g,m}^{(c_1,c_2)} > 0$, and in the idealized case $c_1 = c_2 = 1$ they reduce to $C_{f,m}^{(1,1)} = \frac{1}{2m}$ and $C_{g,m}^{(1,1)} = \frac{m}{6(m+2)}$.

Proof. Let

$$\Omega_r := \Phi_d^{-1}\big(B_r^{\mathrm{spec}}(c)\big), \qquad \mathcal{N}(r) = \int_{\Omega_r} \mathbb{E}[\log \widehat{f}(x)] f(x)\, d\mu_g(x), \qquad \mathcal{D}(r) = \int_{\Omega_r} f(x)\, d\mu_g(x).$$

By Lemma 14.15,

$$B_{r_-}^g(c) \subset \Omega_r \subset B_{r_+}^g(c), \qquad r_- = \frac{r-\varepsilon_N}{c_2}, \quad r_+ = \frac{r+\varepsilon_N}{c_1},$$

and the Jacobian distortion induces only a multiplicative $1 + O(\varepsilon_N)$ error. Thus it suffices to compute the expansions of $\mathcal{N}$ and $\mathcal{D}$ for geodesic balls of radii r_- and r_+ and then express the result in terms of r.

Step 1: denominator. For any $a > 0$ fixed, Lemma 14.13 gives

$$\mathrm{Vol}_g\big(B^g_{ar}(c)\big) = \omega_m (ar)^m \left(1 - \frac{S_g(c)}{6(m+2)}(ar)^2\right) + O(r^{m+3}).$$

Expanding f in normal coordinates, we obtain

$$\int_{B^g_{ar}(c)} f(x)\, d\mu_g(x) = f(c)\, \omega_m (ar)^m - f(c)\, \omega_m (ar)^{m+2} \frac{S_g(c)}{6(m+2)} + \frac{\omega_m (ar)^{m+2}}{2m(m+2)} \Delta_g f(c) + O(r^{m+3}).$$

For $a \in \{1/c_2, 1/c_1\}$ this gives two-sided bounds for $\mathcal{D}(r)$ and hence an expansion

$$\mathcal{D}(r) = D_0\, r^m + D_2\, r^{m+2} + O(r^{m+3} + \varepsilon_N r^m),$$

with $D_0 = \tilde{f}(c)\, \omega_m$ and $D_2 = \tilde{A}_1\, S_g(c) + \tilde{A}_2\, \Delta_g f(c)$, where $\tilde{f}(c)$ is comparable to $f(c)$ up to constants depending only on c_1, c_2 and the coefficients $\tilde{A}_1, \tilde{A}_2$ depend only on (m, c_1, c_2).

Step 2: numerator. By Lemma 14.14,

$$\mathbb{E}[\log \widehat{f}(x)] = \log f(x) + \frac{r^2}{2m} \Delta_g(\log f)(x) + \frac{r^2}{6(m+2)} S_g(x) + O(r^3 + r^2 |\nabla \log f|^2).$$

We evaluate the k-NN estimator on a region whose geodesic radius is asymptotically equivalent to r. Therefore the local scale parameter appearing in Lemma 14.14 satisfies $r_{\mathrm{knn}} = \Theta(r)$, and all bias terms depending on the local k-NN radius may be treated, to second order, as functions of r up to multiplicative constants.

Multiplying by $f(x)$ and expanding near c yields

$$\mathbb{E}[\log \widehat{f}(x)] f(x) = f(x) \log f(x) + \frac{r^2}{2m} f(x) \Delta_g(\log f)(x) + \frac{r^2}{6(m+2)} S_g(x) f(x) + O(r^3).$$

Integrating over $B^g_{ar}(c)$ and using symmetry and Taylor expansions, we arrive at

$$\int_{B^g_{ar}(c)} f(x) \log f(x)\, d\mu_g(x) = (f \log f)(c)\, \omega_m (ar)^m + \Big(B_1(a)\, S_g(c) + B_2(a)\, \Delta_g(f \log f)(c)\Big) r^{m+2} + O(r^{m+3})$$

for explicit coefficients $B_1(a), B_2(a)$ depending only on (m, a). Similarly,

$$\int_{B^g_{ar}(c)} f(x) \Delta_g(\log f)(x)\, d\mu_g(x) = f(c) \Delta_g(\log f)(c)\, \omega_m (ar)^m + O(r^{m+2}),$$

and

$$\int_{B^g_{ar}(c)} S_g(x) f(x)\, d\mu_g(x) = S_g(c) f(c)\, \omega_m (ar)^m + O(r^{m+2}).$$

Combining these contributions and sandwiching Ω_r between $B^g_{r_-}(c)$ and $B^g_{r_+}(c)$, we obtain

$$\mathcal{N}(r) = N_0\, r^m + N_2\, r^{m+2} + O(r^{m+3} + \varepsilon_N r^m),$$

with $N_0 = \tilde{f}(c) \log f(c)\, \omega_m$ and $N_2 = \hat{A}_1\, S_g(c) + \hat{A}_2\, \Delta_g(f \log f)(c) + \hat{A}_3\, f(c) \Delta_g \log f(c)$, for coefficients $\hat{A}_i$ depending only on (m, c_1, c_2).

Step 3: quotient and identification of coefficients. We write

$$\frac{\mathcal{N}(r)}{\mathcal{D}(r)} = \frac{N_0}{D_0} + \frac{N_2 D_0 - N_0 D_2}{D_0^2}\, r^2 + O\big(r^3 + \varepsilon_N\big)\,.$$

The leading ratio is $\frac{N_0}{D_0} = \frac{\tilde{f}(c) \log f(c)}{\tilde{f}(c)} = \log f(c)$. Using the identity

$$\Delta_g(f \log f) = (\log f)\, \Delta_g f + 2\langle \nabla f,\, \nabla \log f \rangle + f\, \Delta_g(\log f),$$

the quantity $N_2D_0 - N_0D_2$ can be rewritten as a linear combination of $S_g(c)$ and $\Delta_g \log f(c)$, with coefficients depending only on $(m, c_1, c_2, f(c))$. Furthermore, all first-order terms vanish upon integration, since geodesic balls in normal coordinates satisfy

$$\int_{B_r^g(c)} \langle \nabla f, v\rangle \, d\mu_g = 0 \quad \text{for every fixed vector } v,$$

and the local bi-Lipschitz distortion of Φ_d preserves this symmetry up to a multiplicative $O(\varepsilon_N)$. Consequently,

$$N_2D_0 - N_0D_2 = D_0^2\Big(C_{f,m}^{(c_1,c_2)}\,\Delta_g(\log f)(c) - C_{g,m}^{(c_1,c_2)}\,S_g(c)\Big),$$

and hence

$$\frac{N_2D_0 - N_0D_2}{D_0^2} = C_{f,m}^{(c_1,c_2)}\,\Delta_g(\log f)(c) - C_{g,m}^{(c_1,c_2)}\,S_g(c),$$

with $C_{f,m}^{(c_1,c_2)}, C_{g,m}^{(c_1,c_2)} > 0$. In the special case $c_1 = c_2 = 1$ these reduce to $C_{f,m}^{(1,1)} = \frac{1}{2m}$ and $C_{g,m}^{(1,1)} = \frac{m}{6(m+2)}$. □

Remark 14.17 (Role of α). The r^2-coefficient in $\mathbb{E}[A_c(r)]$ is purely geometric/measure-theoretic; the penalty $-\alpha \log r$ affects only lower-order (in k) discrete-curvature terms after discretization. Choosing $\alpha = m$ cancels the leading Euclidean volumetric drift in M_c but does not alter the sign or the qualitative structure of the r^2-curvature coefficient.

14.3.3 From radius r to shells k and the Curvature Law

Quantile relation and inversion. Let $k(r)$ denote the expected count inside spectral radius r. By Lemmas 14.13 and 14.15,

$$k(r) \;=\; N\int_{\Omega_r} f\,d\mu_g = \; N\tilde{f}(c)\,\omega_m\,\tilde{r}(r)^m\big(1 - \tilde{\gamma}_g\,S_g(c)\,\tilde{r}(r)^2\big) + \tilde{\gamma}_f\,\tilde{r}(r)^{m+2} + O(N\tilde{r}(r)^{m+3}),$$

where $\tilde{r}(r)$ is any radius between r_- and r_+ and $\tilde{f}(c)$, $\tilde{\gamma}_g$, $\tilde{\gamma}_f$ depend only on $(m, c_1, c_2, f(c))$. Since $r_\pm = a_\pm r + O(\varepsilon_N)$ with $a_- = 1/c_2$, $a_+ = 1/c_1$, we have $\tilde{r}(r) = \bar{a}\,r + O(\varepsilon_N)$ for some $\bar{a} \in [1/c_2, 1/c_1]$. Thus

$$k(r) = \kappa_{\text{eff}}\,r^m\big(1 - \gamma_g^{\text{eff}} r^2\big) + \gamma_f^{\text{eff}} r^{m+2} + O(r^{m+3}),$$

for some effective constants $\kappa_{\text{eff}}, \gamma_g^{\text{eff}}, \gamma_f^{\text{eff}}$ depending only on $(m, c_1, c_2, f(c))$. Solving for r^2 by series inversion gives

$$r^2(k) \;=\; \big(\tfrac{k}{\kappa_{\text{eff}}}\big)^{2/m}\,\Big(1 \;+\; \Theta\big((k/\kappa_{\text{eff}})^{2/m}\big)\Big). \tag{8}$$

Expected profile in k-scale. From Proposition 14.16,

$$M_c(r) = \mathbb{E}[A_c(r)] - \alpha \log r = \text{const} - \alpha \log r + \Big(C_{f,m}^{(c_1,c_2)}\Delta_g \log f(c) - C_{g,m}^{(c_1,c_2)} S_g(c)\Big) r^2 + O(r^3 + \varepsilon_N).$$

Composing with $r = r(k)$ via (8) yields

$$M_c(k) = \;\text{const} - \tfrac{\alpha}{2}\log k \;+\; \Big(\tilde{C}_{f,m}^{(c_1,c_2)}\Delta_g \log f(c) - \tilde{C}_{g,m}^{(c_1,c_2)} S_g(c)\Big) r^2(k) \;+\; \widetilde{R}(k),$$

where $\tilde{C}_{f,m}^{(c_1,c_2)}, \tilde{C}_{g,m}^{(c_1,c_2)} > 0$ are modified constants (absorbing κ_{eff} and the inversion step) and $\widetilde{R}(k) = O\big(k^{3/m}\big) + O(\varepsilon_N)$.

Second finite difference. Since $r^2(k) \sim k^{2/m}$, one has

$$\Delta_k^2\big(\log k\big) = \Theta(k^{-2}), \qquad \Delta_k^2\big(r^2(k)\big) = \Theta\big(k^{2/m-2}\big),$$

and $\Theta(k^{2/m-2})$ dominates $\Theta(k^{-2})$ for every $m \geq 1$. Therefore the discrete curvature is governed asymptotically by the $r^2(k)$-term.

Theorem 14.18 (RCI Curvature Law (Proof of Theorem 3.5))**.** *Under the hypotheses above and with* $\alpha = m$*, there exist positive constants* $K_{f,m}^{(c_1,c_2)}, K_{g,m}^{(c_1,c_2)} > 0$*, depending only on* $(m, c_1, c_2, f(c))$*, such that*

$$\mathbb{E}\big[\Delta^2 M_c(k)\big] = \Big(K_{f,m}^{(c_1,c_2)}\, \Delta_g \log f(c) \;-\; K_{g,m}^{(c_1,c_2)}\, S_g(c)\Big)\, \Delta^2\big(r_k^2\big) \;+\; R_k,$$

where $R_k = o\big(\Delta^2(r_k^2)\big) + O(\varepsilon_N)$*. In particular, if* f *is locally constant at* c *(so that* $\Delta_g \log f(c) = 0$*),*

$$\mathbb{E}\big[\Delta^2 M_c(k)\big] \;\approx\; -K_{g,m}^{(c_1,c_2)}\, S_g(c)\, \Delta^2\big(r_k^2\big),$$

so the expected discrete curvature of the RCI profile detects the sign and magnitude of the scalar curvature $S_g(c)$ *up to a positive multiplicative factor.*

Proof. Apply Δ_k^2 to $M_c(k)$ and use the asymptotics above. The term $-\frac{\alpha}{2}\log k$ contributes $\Theta(k^{-2})$, while the $r^2(k)$-term contributes $\big(\tilde{C}_{f,m}^{(c_1,c_2)}\Delta_g \log f(c) - \tilde{C}_{g,m}^{(c_1,c_2)} S_g(c)\big)\Theta(k^{2/m-2})$, which dominates for all $m \geq 1$. The remainder estimate follows from the composition error in (8) and from $O(r^3) = O(k^{3/m})$. All constants can be absorbed into $K_{f,m}^{(c_1,c_2)}$ and $K_{g,m}^{(c_1,c_2)}$, which are strictly positive because the underlying coefficients in Proposition 14.16 are positive and the inversion step preserves sign. □

Conclusion. The expected discrete curvature of the RCI profile detects an *effective curvature* equal to a linear combination of the intrinsic scalar curvature $S_g(c)$ and the curvature of the sampling density $\Delta_g \log f(c)$, modulated by the universal factor $\Delta^2(r_k^2)$ and by positive geometric constants depending only on $(m, c_1, c_2, f(c))$. This establishes the Curvature Law at the level required for the stability and interpretability of the RCI boundary detector.

References

[1] Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method*. Springer Series in the Data Sciences. Springer, 2015. ISBN 978-3-319-25388-6. doi: 10.1007/978-3-319-25388-6. URL https://link.springer.com/book/10.1007/978-3-319-25388-6.

[2] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. ISBN 9780199535255. doi: 10.1093/acprof:oso/9780199535255.001.0001. URL https://academic.oup.com/book/26549.

[3] Isaac Chavel. *Riemannian Geometry: A Modern Introduction*. Cambridge University Press, 2 edition, 2006. ISBN 9780521853682. URL https://www.cambridge.org/core/books/riemannian-geometry/7F08C88F91F93C6FF45DE6901FFAD91E.

[4] Frédéric Chazal, Vin de Silva, and Steve Y. Oudot. Persistence stability for geometric complexes. *Geometriae Dedicata*, 173:193–214, 2014. doi: 10.1007/s10711-013-9937-z. Free access: https://arxiv.org/abs/1207.3885.

[5] Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. *The Structure and Stability of Persistence Modules*. SpringerBriefs in Mathematics. Springer, 2016. ISBN 9783319425436. doi: 10.1007/978-3-319-42545-0. URL https://link.springer.com/book/10.1007/978-3-319-42545-0.

[6] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007. doi: 10.1007/s00454-006-1276-5. URL https://link.springer.com/article/10.1007/s00454-006-1276-5.

[7] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Stochastic Modelling and Applied Probability*. Springer, New York, 1996. ISBN 978-0-387-94618-7. URL https://link.springer.com/book/10.1007/978-1-4612-0711-5.

[8] 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 2nd International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 226–231, Portland, OR, 1996. AAAI Press. URL https://dl.acm.org/doi/10.5555/3001460.3001507.

[9] Kenneth Falconer. *Fractal Geometry: Mathematical Foundations and Applications*. John Wiley & Sons, 3rd edition, 2014. ISBN 978-1-119-94239-9. URL https://www.wiley.com/en-us/Fractal+Geometry:+Mathematical+Foundations+and+Applications,+3rd+Edition-p-9781119942399.

[10] Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3): 418–491, 1959. doi: 10.1090/S0002-9947-1959-0110078-1. URL https://www.ams.org/journals/tran/1959-093-03/S0002-9947-1959-0110078-1/.

[11] Robin Forman. Morse theory for cell complexes. *Advances in Mathematics*, 134(1):90–145, 1998. doi: 10.1006/aima.1997.1650. URL https://www.sciencedirect.com/science/article/pii/S0001870897916509.

[12] Robin Forman. A user's guide to discrete morse theory. *Séminaire Lotharingien de Combinatoire [electronic only]*, 48:B48c (35 pp.), 2002. URL http://eudml.org/doc/123837.

[13] Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985. doi: 10.1016/0304-3975(85)90224-5. URL https://www.sciencedirect.com/science/article/pii/0304397585902245.

[14] Alfred Gray. The volume of a small geodesic ball of a riemannian manifold. *Michigan Mathematical Journal*, 20(4):329–344, 1974. doi: 10.1307/mmj/1029001150. URL https://projecteuclid.org/journals/michigan-mathematical-journal/volume-20/issue-4/The-volume-of-a-small-geodesic-ball-of-a-Riemannian/10.1307/mmj/1029001150.full.

[15] Jean-Claude Hausmann. On the vietoris–rips complexes and a cohomology theory for metric spaces. In Frank Quinn, editor, *Prospects in Topology: Proceedings of a Conference in Honor of William Browder*, volume 138 of *Annals of Mathematics Studies*, pages 175–188. Princeton University Press, 1995. doi: 10.1515/9781400882588-013. URL https://www.degruyterbrill.com/document/doi/10.1515/9781400882588-013/html.

[16] Juha Heinonen. *Lectures on Analysis on Metric Spaces*. Universitext. Springer, 2001. ISBN 978-0387954043. doi: 10.1007/978-1-4613-0131-8. URL https://link.springer.com/book/10.1007/978-1-4613-0131-8.

[17] Anil K. Jain, M. Narasimha Murty, and Patrick J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999. doi: 10.1145/331499.331504. URL https://dl.acm.org/doi/10.1145/331499.331504.

[18] Svante Janson. Large deviations for sums of partly dependent random variables. *Random Structures and Algorithms*, 24(3):234–248, 2004. doi: 10.1002/rsa.20030. URL https://www2.math.uu.se/~svantejs/papers/sj150.pdf.

[19] Yukio Matsumoto. *An Introduction to Morse Theory*, volume 208 of *Translations of Mathematical Monographs*. American Mathematical Society, 2002. ISBN 978-0-8218-1022-4. doi: 10.1090/mmono/208. URL https://bookstore.ams.org/mmono-208. Translated by Kiki Hudson and Masahico Saito.

[20] Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, volume 141 of *London Mathematical Society Lecture Note Series*, pages 148–188. Cambridge University Press, 1989. doi: 10.1017/CBO9781107359949.008. URL https://www.cambridge.org/core/books/abs/surveys-in-combinatorics-1989/on-the-method-of-bounded-differences/AABA597B562BDA7D89C6077E302694FB.

[21] John W. Milnor. *Morse Theory*. Annals of Mathematics Studies, 51. Princeton University Press, 1963. URL https://webhomes.maths.ed.ac.uk/~v1ranick/papers/milnmors.pdf.

[22] Marston Morse. *The Calculus of Variations in the Large*, volume 18 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 1934. ISBN 978-1-4704-3166-2. URL https://bookstore.ams.org/coll-18. Reprint available from Internet Archive: https://archive.org/details/calculusofvariat0000mors.

[23] Partha Niyogi, Steve Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1–3):419–441, 2008. doi: 10.1007/s00454-008-9053-2. URL https://link.springer.com/article/10.1007/s00454-008-9053-2.

[24] A. Singer. From graph to manifold laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, 2006. doi: 10.1016/j.acha.2006.03.004. URL https://www.sciencedirect.com/science/article/pii/S1063520306000510.

[25] Claude Tricot. *Curves and Fractal Dimension*. Springer-Verlag, New York, 1995. ISBN 978-0-387-94095-3. doi: 10.1007/978-1-4612-4170-6. URL https://link.springer.com/book/9780387940953. Translation from French. Foreword by M. Mendes France.

[26] Nicolás García Trillos and Dejan Slepčev. Continuum limit of total variation on point clouds. *Archive for Rational Mechanics and Analysis*, 220:193–241, 2016. doi: 10.1007/s00205-015-0929-z. URL https://link.springer.com/article/10.1007/s00205-015-0929-z. Free access: https://arxiv.org/abs/1403.6355.

[27] Karen K. Uhlenbeck. Connections with L^p bounds on curvature. *Communications in Mathematical Physics*, 83(1):31–42, 1982. doi: 10.1007/BF01208378. URL https://projecteuclid.org/journals/communications-in-mathematical-physics/volume-83/issue-1/Connections-with-Lp-bounds-on-curvature/cmp/1103920743.full.

[28] Karen K. Uhlenbeck. Removable singularities in Yang–Mills fields. *Communications in Mathematical Physics*, 83(1):11–29, 1982. doi: 10.1007/BF01947068. URL https://link.springer.com/article/10.1007/BF01947068.

[29] Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007. doi: 10.1007/s11222-007-9033-z. URL https://link.springer.com/article/10.1007/s11222-007-9033-z.