Jag Pratap Singh Yadav, 2014, 2:8
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

The Geometry of Data: A Conceptual Framework for Dimensionality Reduction in High-Dimensional Spaces

Jag Pratap Singh Yadav
Department of Mathematics
Government Degree College, NadhaBhoor, Sahaswan, Badaun, Uttar Pradesh

Abstract- The ever-expanding high dimensional data in both scientific and technology domains has posed fundamental difficulty in analysis, interpretation, and visualization, described collectively as the "complexity crisis." In this paper, we argue for a geometric perspective of how dimensionality reduction is a principled response to these challenges. It opens with the discussion on the pathological effects of high dimensionality such as distance concentration, volume distortion, and exponential sampling complexity, which undermine traditional statistical and distance analysis methods. Based on this premise, this method contributes to providing the ground for the manifold hypothesis, which states that high-dimensional data are usually present on or close to a low-dimensional, smoothly embedded manifold. And it moves the emphasis from the ambient Euclidean structure into the geometry of relationships; this means that the difference that can accurately interpret meaningful data structure lies between Euclidean and geodesic distances. The paper also provides a general comparison of global and local dimensionality reduction methods, highlighting an inherent trade-off in obtaining large-scale variance without losing local neighborhood structure. It shows that reducing the dimension is inherently limited by the information loss — which is formally described in the Johnson–Lindenstrauss lemma and rate-distortion theory as limiting the fidelity of low-dimensional representations. One of the main contributions of this paper is the description and critique of geometric hallucinations— artificial structures (for example those created by dimensionality reduction algorithms) that may misrepresent the true high-dimensional geometry. This is a warning of 'over-interpreting' good looking embeddings without sound validation. More generally, we recast dimensionality reduction in the paper as a geometric and information-theoretic compromise between interpretability and fidelity. It also underscores the importance of methodological selection, critique of visualizations and approaches that focus on distortion, and designing processes that address distortion (to provide more robust and theory-derived applications of high-dimensional data analytics), providing a more trustworthy and theoretically informed approach to the application of high-dimensional data analysis.

Keywords— Dimensionality Reduction, High-Dimensional Data, Data Geometry, Manifold Learning, Feature Extraction, Data Representation

I. INTRODUCTION THE COMPLEXITY CRISIS:

1. The First Paragraph — The Data Explosion

The sheer scale of the high-dimensional data explosion in scientific, technological, and socioeconomic domains has transformed the modern analytic method greatly. Modern data generating systems - from next-gen sequencing in genomics, to distributed sensor networks in financial markets, graph-theoretic representations of ecosystems in social media - increasingly generate observation spaces that take up thousands and sometimes millions of factors for each sample unit.

One RNA sequencing experiment could generate expression profiles for 20,000–30,000 genes per individual cell and thousands of cells at one time (Luecken & Theis, 2019), a social network interaction matrix may encode behaviors over hundreds of thousands of latent features, high frequency financial trading systems could generate multivariate time series at millisecond resolution in thousands of inter-correlated instruments (Dixon et al., 2020). The character of this data epoch is not simply volume; rather it's geometric complexity: the data does not merely inhabit space, it happens to inhabit space in ways that do not lend themselves to the limits prescribed by conventional instruments to handle lower-dimensional regimes.

2. The Curse of Dimensionality: A Formal Problem Statement

The basic difficulty of high-dimensional data analysis was first clearly declared by Bellman (1957) in the context of dynamic programming and later formalized within the statistical learning literature as the curse of dimensionality – that the geometric and statistical properties of data deteriorate in counterintuitive and computationally catastrophic ways as the number of dimensions d grows large. To grasp the geometrical weight of this curse we can take the following formal notes:

Volume Concentration Phenomenon

For a d-dimensional hypersphere of radius r, the fraction of volume contained within a thin shell of thickness ε approaches unity as d → ∞:

$$\frac{V(r)-V(r-\varepsilon)}{V(r)} = 1 - \left(1 - \frac{\varepsilon}{r}\right)^d \xrightarrow{d\to\infty} 1$$

This shows that nearly all of the volume of a high-dimensional ball is just right over its surface, so estimating density — a foundational work of classical statistical inference — is fundamentally suspect for high dimensions.

Distance Concentration

Aggarwal et al. (2001) demonstrated that in high-dimensional spaces, the ratio of maximum and minimum pairwise distances for randomly sampled points converges towards unity:

$$\lim_{d\to\infty} \frac{\text{dist}_{\max} - \text{dist}_{\min}}{\text{dist}_{\min}} \to 0$$

The practical outcome is extremely destabilizing: near neighbor relationships — the geometric architecture of clustering, classification and manifold learning methods — become meaningless once all the points are approximately equidistant from each other. Distance-based methodologies like k-nearest neighbors, k-means clustering, and kernel density estimation are suffering from the devastatingly catastrophic degradation in this regime as well (Beyer et al., 1999).

Sampling Complexity

The quantity required to sample uniformly at a fixed resolution grows exponentially with dimensionality. To achieve a constant sampling density in every unit hypercube, the required sample size scales as N ~ r^{-d}, where r is the resolution required. d = 100 makes this computationally and practically prohibitive by orders of magnitude, making empirical coverage of the feature space an impossible task (Hastie et al., 2009).

Together, these geometric pathologies form what could be described as the Complexity Crisis of contemporary data science: the data available is at an all-time high-dimensional one for conventional analysis, while being also too sparse regarding the dimensionality of the space it resides in.

The Failure of Human Cognition and Classical Visual Models

In addition to mathematical formalism, there is another epistemological limit: the basic incompatibility of high-dimensional data structures with the cognitive-perceptual architecture of human intelligence. Humans are the evolutionary products of a three-dimensional physical world, and also possess strong perceptual and intuitive devices for reasoning about spatial relationships in at most three dimensions. It is also a constraint that goes beyond simple preference for visualization design: The constraint reflects a hard boundary on the capacity for human-interpretable scientific insight. This limitation is underpinned by extensive neuroscientific evidence. The hierarchical processing

of spatial, chromatic, and motion information organized under the human visual cortex creates the highest fidelity representations for configurations of objects in 2D and 3D Euclidean space (Mishkin et al., 1983). There are perceptual interference effects on graphical representations that encode four or more simultaneous dimensions through aesthetic channels (color, shape, size, opacity, texture), which ultimately yield significantly reduced interpretability (Ware, 2004). But beyond five or six dimensions, even expert practitioners fall short in the ability to make firm geometric inferences about how data were structured. This limitation has a clear impact on the scientific enterprise:

Exploratory data analysis is a process that is a critical part of the scientific process in and of itself: if a researcher creates a number of hypotheses about structure, clusters, and relationships in the data, their hypothesis is essentially meaningless without dimensionality reduction.

Model interpretability—now a growing regulatory and ethical issue in areas such as clinical medicine and financial risk analysis—means that high-dimensional behaviors with complex models must be understood as using reduced dimension in projections.

II. THE GEOMETRY OF DATA: THE MANIFOLD HYPOTHESIS

1. Opening Paragraph — The Hidden Simplicity of Complex Data

One of the most deep-seated and contradictory revelations to come about in the convergence of differential geometry and contemporary machine learning is the realization that complexity, in high-dimensional data, is frequently more apparent than real. But while the "ambient dimensionality" of modern datasets — the number of coordinates required before a single observation can be specified formally — might span thousands to millions, the fundamental geometric difficulty of real data is often orders of magnitude lower. This disjuncture between ambient and intrinsic dimensionality is not a statistical oddity; it's an inherent structural condition of the physical, biological, and social processes which produce real-world data. These processes are dictated by few underlying degrees of freedom — the laws of physics, biochemical constraints, social dynamics, evolutionary pressures — that impose strong geometric regularities on the data they produce. The Manifold Hypothesis formalizes this insight mathematically: it suggests that high-dimensional observational data, despite occupying a nominally large ambient space (Bengio et al., 2013; Fefferman et al., 2016), concentrates with overwhelming probability in proximity to a low-dimensional, smoothly curved geometric object (a manifold) embedded within that space. This study is primarily about understanding this hypothesis, its mathematical foundations, empirical evidence and dimensionality reduction implications; that is the major geometric work of this paper.

2. Manifolds: A Rigorous Introduction

Manifolds: A Detailed Explanation

Before the Manifold Hypothesis can be accurately examined, the central mathematical object of reference that it addresses is known as a manifold. In some sense, a manifold is a geometrical space which at a local level looks like Euclidean space despite the fact that its global structure is a far more complex space. It is precisely this characteristic of local Euclidean property that makes manifolds analytically tractable and geometrically rich.

Topological Manifold: Definition

A topological space M is known as a d-dimensional manifold if every point p ∈ M has an open neighborhood U ⊂ M that is homeomorphic to an open subset of $\mathbb{R}^d$. The integer d is referred to as the intrinsic dimension or topological dimension of M.

Smooth Riemannian Manifold (for definitions)

A d-dimensional Riemannian manifold (M, g) is a smooth manifold with a Riemannian metric tensor g — a smoothly varying, positive-definite inner product on each tangent spaceT_{pM} — that generates notions of distance, angle, curvature, and geodesic paths that are integral to the manifold and can be independent of any ambient embedding space. The crucial distinction for the purposes of dimensionality reduction is between two

fundamentally different notions of distance on a manifold embedded in ambient Euclidean space $\mathbb{R}^D$:

$$d_E(x_i, x_j) = \|x_i - x_j\|_2 = \sqrt{\sum_{k=1}^{D}(x_{ik} - x_{jk})^2}$$

Geodesic (intrinsic) distance: The length of the shortest path between two points *constrained to lie on the manifold M*, formally defined as:

$$d_G(p, q) = \inf_\gamma \int_0^1 \sqrt{g_{\gamma(t)}(\dot\gamma(t), \dot\gamma(t))}\, dt$$

In fact, this divergence between Euclidean and geodesic distance is our geometric signature for manifold curvature and is the primary reason why linear dimensionality reduction methods fail on curved manifolds: they preserve Euclidean distances through the ambient space while ignoring the intrinsic geodesic structure that encodes the true relationships between data points. This difference, conceptually illustrated in Figure 1, is the reason behind the entire program of nonlinear manifold learning.

3. The Manifold Hypothesis: Formal Statement and Intuition:

The Manifold Hypothesis: Now we have $X = \{x^1, x^2, \ldots, x_N\} \subset \mathbb{R}^D$, and it is a dataset extracted i.i.d. from an unknown probability distribution P over $\mathbb{R}^D$. The Manifold Hypothesis asserts that P is supported on — or located in a thin neighborhood of — a compact, smooth Riemannian manifold M of intrinsic dimension d ≪ D, smoothly nested in $\mathbb{R}^D$:

$$\text{supp}(P) \subseteq \mathcal{M} \subset \mathbb{R}^D, \ \dim(\mathcal{M}) = d \ll D$$

where the intrinsic dimension d is the actual value of the independent degrees of freedom that mediate the relevant variation in the data.

The physical intuition underpinning this hypothesis, of course, is best expressed through what we shall call the degrees-of-freedom argument. Think of any complicated real-world data generation process at all: the pose and illumination of a human face in a photo session, the profile of gene expression of a cell on a developmental trajectory, the price behavior (of a financial instrument) and macroeconomic dynamics behind the price. In either scenario, the nominal dimensionality of the measurement millions of pixels, tens of thousands of gene transcripts, thousands of correlated price series — far exceeds the number of independent physical factors that account for variation in the data. The manifold is effectively the geometric shadow cast by the underlying causal structure of the data-generating process: it has dimension equal to the number of independent latent variables constituting that process, and curvature encodes nonlinear relationships between those latent variables and the measured data.

4. The Crumpled Paper: A Geometric Archetype

To put the abstract geometry of the structure into concrete terms and to set up the conceptual architecture for the remainder of this paper, we present a canonical analogy that accurately represents the most important geometric relationships in manifold-structured data and its reduction.

The Analogy Constructed

Consider a sheet of paper. It is, in its flat, undeformed form, an object of unambiguous geometric identity: a two-dimensional Euclidean plane, parameterized by two coordinates (u, v) ∈ $\mathbb{R}^2$, with a flat Riemannian metric, zero Gaussian curvature everywhere, and geodesic distances equal to Euclidean distances. The intrinsic geometry of the flat sheet itself is the simplest of all: straight lines are geodesics, circles are round, and the Pythagorean theorem holds exactly.

Now crumple the paper and place it in a three-dimensional room. The physical object has been transformed in its relationship to the ambient three-dimensional space, $\mathbb{R}^3$: now it takes on a complex, folded, curved geometric configuration, in which all three spatial coordinates are needed to specify the position of any point on its surface. A naive observer encountering this crumpled object without prior knowledge of its origin would conclude that understanding its geometry requires the full machinery of three-dimensional analysis.

A fundamental geometric fact, however, remains unchanged by the crumpling: the intrinsic geometry

of the paper — the distances measured along its surface, the angles between curves drawn on it, the areas of regions — is identical to that of the original flat sheet. The reason is that bending deformations are isometric, preserving the Riemannian metric and hence all intrinsic geometric quantities, in accordance with Gauss's Theorema Egregium (Gauss, 1828): the Gaussian curvature of a surface is an intrinsic invariant, unchanged by isometric embeddings.

Thus, this crumpled paper is a 2-dimensional manifold isometrically embedded in 3-dimensional ambient space. It is a geometric object of intrinsic dimension 2 ($d = 2$) embedded in an ambient space of dimension 3 ($D = 3$), with $d \ll D$ in the relative sense relevant to the manifold hypothesis.

5. The Biggest Geometric Lesson: Ambient vs. Intrinsic Distance

The crumpled paper analogy distills the most critical geometric learning in dimensionality reduction; the yawning potential distance away from ambient Euclidean distance and intrinsic geodesic distance as proxies for the real-world relationship between points in a graph. We consider two points A and B on the surface of the crumpled paper as a result of a folding geometry that, as predicted, occurs within a very near spatial proximity in the three-dimensional room — $\mathbb{R}^3$ distance by which the two points are separated. But the shortest distance from A to B that stays on the paper's surface — their geodesic distance — could be a whole lot larger, requiring a long walk across the paper's surface to walk around the folds and creases separating them. In contrast, two points C and D that lie far apart in the three-dimensional room can be linked by a short surface trajectory through a simple unfolded structure of the paper. This geometric case is not only applicable in high-dimensional data — it is the generic case. For data with manifold structure, the Euclidean distance between two observations (in ambient feature space) is a systematically misleading measure where the true functional or semantic similarity is better reflected by their geodesic distance along the data manifold. A dimensionality reduction algorithm which is based solely on Euclidean distance — as also happens with any linear method—is, in the crumpled-paper metaphor, not measuring distances through the surface, but through air: an error that becomes more and more serious as the manifold becomes more folded, more and more curved, and more and more part of a high-dimensional ambient space.

III. COMPARATIVE APPROACHES: GLOBAL VS. LOCAL GEOMETRY

The movement of high-dimensional manifold-structured data into low-dimensional representations is no isolated mathematical operation, but rather a set of operations that are based on a number of distinct geometric concepts about what is meaningful structure that deserves to be preserved and what is merely useless noise that does not deserve to be preserved. Central to this methodological diversity is an underlying tension in the field of dimensionality reduction — it's the tension between the global and local geometric fidelity. The questions at the heart of the global approach are: what is the best single perspective from which to view the entire dataset, such that the overall spread and orientation of the data is maximally preserved? Local methods raise a question: for each data point, which other points constitute its immediate neighborhood, and how can that neighborhood's structure be faithfully reproduced in a lower-dimensional space? These are not only different questions, they are, in a mathematical sense, geometrically antagonistic problems: algorithms that are optimized to answer one question well are systematically called upon to answer the other poorly. A critical appreciation of this tension (from the mathematical roots, from the practical ramifications and from philosophical implications of each approach) is essential in selecting the principled methodology in high-dimensional data analysis. This section creates an ambitious comparative approach to global and local dimensionality reduction, structured in terms of two key conceptual metaphors: global methods as the change of perspective, and local methods in neighborhood preservation and the geometric and epistemological underpinnings that distinguish them.

1. The Fundamental Dichotomy: A Geometric Framework

Before examining specific algorithms, it is instructive to establish a unified mathematical framework within which the global-local distinction can be precisely characterized. All dimensionality reduction methods seek a mapping:

$$\phi : \mathbb{R}^D \to \mathbb{R}^d, \ d \ll D$$

that minimizes some notion of **geometric distortion** — the degree to which the mapping fails to preserve the relationships between data points. The critical parameter distinguishing global from local methods is the **scale at which geometric relationships are measured and preserved**.

Define the **distortion function** of a dimensionality reduction mapping φ at scale *r* as:

$$\mathcal{D}(\phi, r) = \mathbb{E}\left[\left| d_{\mathbb{R}^D}(x_i, x_j) - d_{\mathbb{R}^d}(\phi(x_i), \phi(x_j)) \right| \ \middle| \ d_{\mathbb{R}^D}(x_i, x_j) \leq r \right]$$

Global methods minimize $\mathcal{D}(\varphi, r)$ for $r \to \infty$ — they search for mappings such that the distance between each pair of points, including those separated by large distances in the ambient space, is preserved. Local methods minimize $\mathcal{D}(\varphi, r)$ for small r – they search to get mappings whose distances are preserved between points only in local neighborhoods of radius r, even if at a larger scale fidelity is necessarily sacrificed for much higher resolution local structures. This distortion framework, dependent on scale, shows not a binary local-global dichotomy but a continuous one: methods can be parameterized across a range of locality, from the classic global (classical MDS, PCA) to intermediate scales (Isomap, LargeVis) to highly local (t-SNE with small perplexity, UMAP with small k). Part of the practical art of method selection is, in this case, tuning the locality parameter to the scale of the structure relevant to the specific scientific question.

2. Global Approaches: Dimensionality Reduction as Perspective Shift

The Conceptual Metaphor—Finding the Optimal Viewpoint

The governing metaphor for global dimensionality reduction methods is that of perspective shift—that is, the solution to find from where an optimal vantage point is located for looking at complicated high-dimensional geometric objects such that the most information about its structural characteristic is captured in their resultant two- or three-dimensional view. This metaphor is not an image in the purest sense, though; it is mathematically exact. Most of them are global methodologies, most prominently Principal Component Analysis (PCA), in which the linear subspace of $\mathbb{R}$^D on which the data projection maximizes the preserved variance (ie: the extent to which the represented vector maintains the distributional, orientation, and relational structure of our original high-dimensional setup) is sought. Picture confronting a multi-dimensional sculpture and trying to encapsulate its essence in a single image. Choosing a camera angle — the perspective — is critical: perspectives will tell you where the sculpture's axes of extension are and what are its major lines of relationship between the parts that exist or others will make the camera make a projection in which the different parts of the sculpture looked like overlaying or important detail is hiding some sort of shape in the structure. The ideal angle is the angle at which the most information is transmitted out of the resultant 2D projection — the one that tells us the most about the true 3D structure of the sculpture. This problem is exactly the one addressed by solutions based on global dimensionality reduction techniques, generalized to arbitrary dimension space.

Principal Component Analysis: Maximum Variance Geometry

Principal Component Analysis (PCA) is an approach to dimensionality reduction popularized by Pearson (1901) and further developed by Hotelling (1933) so far that it is the most popular dimensionality reduction method among many fields of study and the original exemplar of the global geometric approach. The mathematical formulation makes the

point quite explicitly of the geometric concepts behind the perspective-shift metaphor.

Problem Formulation

Given a centered data matrix with X ∈ ℝ^{N×D} (having zero empirical mean, obtained by subtracting the column means), PCA aims to obtain an orthonormal basis $W = [w^1, w^2, \dots, w_d] \in \mathbb{R}^{\{D\times d\}}$— the principal components — such that the projection $Z = XW \in \mathbb{R}^{\{N\times d\}}$maximizes the sum of the total preserved variance:

$$W^* = \arg\max_{W\in\mathbb{R}^{D\times d},\, W^TW=I} \operatorname{tr}(W^T S W)$$

Geometric Solution

The solution to this constrained optimization problem is given by the **spectral decomposition** of the covariance matrix:

$$S = V\Lambda V^T$$

where $V = [v^1, v^2, \dots, v_D] \in \mathbb{R}^{\{D\times D\}}$is the orthogonal matrix of eigenvectors; and $\Lambda = \operatorname{diag}(\lambda^1, \lambda^2, \dots, \lambda_D)$ is the diagonal matrix of eigenvalues whose order is such that $\lambda^1 \geq \lambda^2 \geq \dots \geq \lambda\ D \geq 0$. The optimal directions of the projection— for example the principal components—are the top d eigenvectors $W* = [v^1, v^2, \dots, v_d]$ and λ_k is the explained variance of the k-th principal component. The proportion of PCA projection as represented by the overall variance being explained by the d-dimensional projection is:

$$\rho_d = \frac{\sum_{k=1}^{d} \lambda_k}{\sum_{k=1}^{D} \lambda_k}$$

This number, called the scree ratio, sets out a principled criterion for choosing target dimensionality d: that is enough principal components to represent up to a set proportion (a typical 85–95 percent) of total variance.

Geometric Interpretation: Ellipsoids, Axes, and Orientation

The most apparent geometric content of PCA is found in the language of quadratic forms and ellipsoids. The covariance matrix S tells us about a quadratic form, $x^{TS^{\{-1\}x}} = 1$, with a level set that is the D-dimensional ellipsoid of the data ellipsoid, with principal axes corresponding to eigenvectors of S and semi-axis lengths proportional to the square roots corresponding to the eigenvalues$\sqrt{\lambda_k}$. In a geometric sense, PCA is literally the thing that you do in order to identify the principal axes of this data ellipsoid (i.e., directions of maximum extension); use these axes for the reduced-dimensional representation. The first principal component v_1 sets the direction in which the data ellipsoid is elongated by most—the axis of maximum variance—and projects the data in this direction, giving a one-dimensional reflection but still covering the largest amount of variation possible for the total data distribution. Additional principal components can be depicted in orthogonal directions with lower variance, leading to a nested sequence of projection subspaces.

IV. CRITICAL TRADE-OFFS: INTERPRETABILITY VS. FIDELITY: OPENING STATEMENT--THE FAUSTIAN DEAL OF DIMENSIONALITY REDUCTION

Each case of dimensionality reduction can therefore easily be imagined as a Faustian bargain, if one makes this bargain with mathematics: the exchange for geometric completeness is made through reduced cognitive accessibility, and not only with physical data, but also with the very nature of the structure for the interpretation. When a researcher transforms a dataset consisting of one hundred, one thousand or one hundred thousand dimensions into the two-dimensional space of a printed figure or computer screen, they don't just reduce data, they fundamentally transform it, dumping the geometrical information that may be scientifically irreplaceable, imposing distortions that are visually indistinguishable from genuine structure and creating visualizations that look clear but aren't. This part addresses arguably the most impactful and least debated dimensionality reduction approach: the systematic, mathematically inevitable costs of that process at work – the information lost forever by way of projection, the structural deformation and distortion wreaked by flattening and the devastating potentials of dimensionality reduction algorithms to produce visual artifacts in the form of biological,

physical or social reality that act as images only they cannot be reconstructed and/or be manipulated. These costs cannot be considered to be a peripheral technical issue but rather a serious question of scientific integrity - with far-reaching implications for the relevance, credibility of published findings across genomics, neuroscience and social science, and in every other area where dimensionality reduction has become a tried-and-true analytic tool. The field has, we submit, developed a problematic tendency to treat outputs of dimensionality reduction as transparent windows onto high-dimensional reality when rather it is more accurately described as distorting mirrors — mirrors that show real structure but also add systematic artifacts that must be discerned and interpreted only through analytic geometric literacy.

1. Projection and Missing Information: The Mathematics of Loss

Projection as Irreversible Geometric Surgery

Dimensionality reduction comes at an irreconcilable cost — data loss, i.e., permanent information loss of geometric information whenever a high-dimensional geometric structure is projected onto a lower-dimensional space. This loss isn't some artificial artifact or algorithm failure; it's a mathematical theorem, an inescapable consequence of the geometry of projections that no algorithm, no matter how sophisticated, can evade. To nail down this need specifically, let's turn our attention to the dimension-counting argument. In the broadest case, N points in $\mathbb{R}^D$ correspond to $\frac{N(N-1)}{2}$ pairwise distances — and this number grows quadratically with the number of data points. With N points in $\mathbb{R}^D$ (with $d \ll D$), there are only Nd degrees of freedom - which scales linearly. Since N must be sufficiently large with respect to d, pairwise distance constraints should be greater than the degrees of freedom granted for the lower-dimensional regime, and therefore, the precise retention of all pairwise distances would be an impossible mathematical task. There are distances which must be distorted; all that is needed is to know which, and how. This observation can be formalized in the Johnson-Lindenstrauss Lemma (Johnson & Lindenstrauss, 1984), which specifies the least dimensionality required to approximately preserve pairwise distances:

Theorem (Johnson-Lindenstrauss): For any set of N points $X \subset \mathbb{R}^D$ and any distortion parameter $\varepsilon \in (0, 1)$, there is a linear mapping $\varphi: \mathbb{R}^D \to \mathbb{R}^d$ with:

$$d \geq \frac{4 \ln N}{\varepsilon^2/2 - \varepsilon^3/3}$$

such that for all pairs *i, j*:

$$(1-\varepsilon)\|x_i - x_j\|_2 \leq \|\varphi(x_i) - \varphi(x_j)\|_2 \leq (1+\varepsilon)\|x_i - x_j\|_2$$

The key finding of this theorem is that the minimum dimensionality taken to realize a pairwise distance within a factor of (1 ± ε) increases logarithmically with number of data points and inversely with the square of the allowed distortion. For N = 10,000 points and a distortion demand of ε = 0.1 (10%) or less, the minimum dimensionality is around d ≥ 1,842 – much greater than the typical 2 or 3 dimensions needed for viewing. Therefore, scaling down to two dimensions ensures distortions of the pairwise distance configuration which far exceed a reasonable accuracy limit, whatever the algorithm that is used.

2. The Spectral Decomposition of Information Loss

For linear dimensionality reduction—exemplified by PCA—the information loss can be quantified with mathematical precision through the spectral decomposition of the data covariance matrix. The total information content of the data (in terms of total variance) is:

$$\mathcal{I}_{\text{total}} = \text{tr}(S) = \sum_{k=1}^{D} \lambda_k$$

The information retained by a *d*-dimensional PCA projection is:

$$\mathcal{I}_{\text{retained}}(d) = \sum_{k=1}^{d} \lambda_k$$

And the **information loss** — the structural clarity sacrificed for visual interpretability — is:

$$\mathcal{I}_{\text{lost}}(d) = \sum_{k=d+1}^{D} \lambda_k = \text{tr}(S) - \sum_{k=1}^{d} \lambda_k$$

This lost information is not random noise — it's structured geometric data which corresponds to the directional variation captured by the discarded eigenvectors($v_{\{d+1\}} \ldots, v_D$). In datasets where the eigenvalue spectrum decays slowly — where variance is distributed relatively uniformly across many dimensions — the information loss from reducing to d = 2 may be catastrophic, with the two retained dimensions capturing only a small fraction of total variance. For example, we can look at single-cell genomics: a typical scRNA-seq dataset will show an eigenvalue spectrum with only the top 2 principal components accounting for 15–25% of the overall variance (Luecken & Theis, 2019). Accordingly, every PCA visualization of such data discards 75–85% of the total geometric information — much of it might contain vital information related to cell-type differences, regulatory state heterogeneity, and developmental trajectory structure that shows up as lower-variance dimensions.

3. Rate-Distortion Theory: The Information-Theoretic Framework

Shannon's rate-distortion theory (Shannon, 1959), initially a lossy source coding framework but applicable to the dimensionality reduction problem of geometric compression, provides the most rigorous framework to analyze the tradeoff between model dimensionality (rate) and information loss (distortion). Define the rate-distortion function R(Δ) as the minimum number of bits needed to encode the geometric structure of the dataset such that the expected distortion — measured by the mean squared error between original and reconstructed pairwise distances — does not exceed Δ.

$$R(\Delta) = \min_{p(\hat{x}|x):\, \mathbb{E}[d(x,\hat{x})] \le \Delta} I(X;\hat{X})$$

where I(X; X̂) denotes the mutual information between original and reconstructed data representations. For a Gaussian source with covariance S having eigenvalues$\lambda^1 \geq \lambda^2 \geq \ldots \geq \lambda_D$, the rate-distortion function is achieved by the reverse waterfilling solution:

$$R(\Delta) = \frac{1}{2} \sum_{k=1}^{D} \max\left(0,\ \log_2 \frac{\lambda_k}{\theta}\right)$$

where θ is a Lagrange multiplier ("water level") chosen to satisfy the distortion constraint $\sum_k min(\lambda_k, \theta) = \Delta$.

This framework establishes a fundamental lower bound on the distortion achievable at any given dimensionality d: no algorithm can produce a d-dimensional representation with distortion below the rate-distortion bound. The practical implication is unambiguous — there exists a hard, algorithm-independent limit on the fidelity achievable in any low-dimensional representation, and this limit becomes increasingly severe as the target dimensionality d is reduced toward 2 or 3.

4. Structural Clarity vs. Visual Interpretability: The Core Tradeoff

The rate-distortion framework formalizes what might be called the fundamental tradeoff of dimensionality reduction: the inverse relationship between structural clarity (the fidelity with which the reduced representation captures the true high-dimensional geometry) and visual interpretability (the cognitive accessibility of the representation to human perception and analysis).

This tradeoff can be visualized as a Pareto frontier in the space of (interpretability, fidelity) — a curve along which improvements in one dimension necessarily come at the cost of the other. Methods that produce clean, visually compelling cluster structures — t-SNE and UMAP with aggressive locality parameters — achieve high visual interpretability by sacrificing geometric fidelity, compressing and distorting the data in ways that make it visually appealing but geometrically unreliable. Methods that maintain high geometric fidelity — classical MDS with large target dimensionality, PCA retaining many components preserve structural clarity at the cost of visual interpretability, producing representations that are geometrically faithful but cognitively opaque.

The tragedy of current practice in many scientific fields is the systematic selection of methods that maximize visual interpretability at the expense of geometric fidelity — producing figures that are visually compelling and easy to interpret, but that

misrepresent the true structure of the underlying data in ways that are difficult to detect without specialized geometric knowledge.

V. DIMENSIONALITY REDUCTION HALLUCINATIONS: THE CREATION OF FAKE STRUCTURE:

1. Defining the Hallucination Problem

The most alarming failure mode of dimensionality reduction — and the one with the gravest implications for scientific validity — is the phenomenon we term geometric hallucination: the systematic generation of apparent structures in low-dimensional representations that have no counterpart in the true high-dimensional data geometry. These hallucinations take multiple forms: phantom clusters that appear as distinct populations in the visualization but correspond to continuous, unstructured distributions in high-dimensional space; false trajectories that suggest developmental or temporal ordering where none exists; spurious separations between populations that are genuinely overlapping in the ambient space; and artificial compression of genuinely distinct populations into apparent proximity.

The critical and deeply unsettling property of geometric hallucinations is that they are, in general, visually indistinguishable from genuine structure. A phantom cluster produced by t-SNE's repulsive force dynamics appears in the visualization with the same visual characteristics — compact shape, sharp boundaries, internal cohesion — as a genuine biological cell population, a genuine social community, or a genuine physical phase. Without independent validation from the high-dimensional data or domain-specific prior knowledge, a naive interpreter has no visual basis for distinguishing the two.

2. The Mechanical Origin of Hallucinations in t-SNE

The mechanical origin of geometric hallucinations in t-SNE can be traced with mathematical precision to the asymmetry of the KL divergence objective and the dynamics of the attractive-repulsive force system that governs the optimization.

Recall that the t-SNE objective minimizes:

$$L_{\text{t-SNE}} = \text{KL}(P \parallel Q) = \sum_{i \neq j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right)$$

The asymmetry of KL divergence — specifically the fact that $KL(P \parallel Q) \neq KL(Q \parallel P)$ — has a profound and underappreciated geometric consequence. The forward KL divergence $KL(P \parallel Q)$ is infinite whenever Q(x) = 0 and P(x) > 0, creating an overwhelming penalty for *failing to represent pairs that are neighbors in high-dimensional space as neighbors in low-dimensional space*. By contrast, the forward KL divergence imposes no penalty for representing non-neighboring pairs as close in low-dimensional space — for *false proximities* that have no basis in the high-dimensional geometry.

The asymmetry of KL divergence — in particular, that $KL(P \parallel Q) \neq KL(Q \parallel P)$ — has a significant geometric implication but is not well-appreciated. The forward KL divergence $KL(P \parallel Q)$ is infinite whenever Q(x) = 0 and P(x) > 0, making it a staggering penalty for failing to represent pairs that are neighbors in high-dimensional space as neighbors in low-dimensional space. In contrast, the forward KL divergence applies no penalty for describing non-neighboring pairs as being close to each other in low-dimensional space — for false proximities that do not arise in high-dimensional geometry. This asymmetry leads to an inherent attractive bias — we are very biased toward imposing a stronger penalty for pushing genuine neighbors apart (false separations, at least locally) but less on the other hand to promote the pull of non-neighbors together (false proximities) — at least locally. While the global repulsive term in the gradient offsets the loss of the neighborhood, it is likely to be quite sensitive to the perplexity parameter and unique geometry of the dataset; hence, we cannot account for all contributions. The hallucination mechanism can be broken down into the following steps, which can be analysed in order to use the force-based way in order to interpret the t-SNE gradient.

Stage 1 — Random Initialization: The optimization process starts with a random low-dimensional setting, where data points are uniformly arranged. The high-dimensional neighborhood probabilities p_{ij} around genuine neighbors lie primarily on the gradient; the low-dimensional probabilities $q_{ij}$are almost equally distributed. The strong attractive forces pulling the genuine neighbors are the dominant force acting on the gradient.

Stage 2 — Cluster Formation: As the neighbors pull closer and closer to each other, the attractive forces cause data points to aggregate into local clusters that roughly correspond to high-density parts of the neighborhood graph. This step faithfully reflects actual local structure in the high-dimensional data, whereas the repulsive forces between forming clusters induced by the q_ij terms inside the gradient begin to push the clusters apart, creating the characteristic white space separation between clusters in t-SNE visualizations.

Stage 3 — Amplifying Hallucination: And as we optimize, the repulsive forces between clusters do become more and more dominant. This pushes the clusters into the periphery of the embedding space and artificially intensifies the apparent separation between them. Importantly, the size of inter-cluster repulsion depends on the low-dimensional geometry, not on the high-dimensional data structure — meaning that a cluster that is located nearby to the middle of the embedding will experience stronger net repulsion than one positioned closer to the periphery, resulting in position-dependent distortions with no basis in the original data.

Stage 4: Fragmentation — In hallucination's critical failure mode, the repulsive forces between the sub-clusters of a truly continuous high-dimensional population become strong enough to rend the population apart into seemingly distinct clusters separated by regions of empty embedding space that do not correspond to a real gap in the high-dimensional distribution. This gives rise to a visualization in which one continuous population is presented as two or more discrete clusters — a phantom structure that is a complete artifact of the optimization dynamics.

VI GLOBAL VARIANCE MAXIMIZATION VIA EIGEN-DECOMPOSITION (THE PCA FRAMEWORK):

Let

$$X = \{x_1, x_2, \dots, x_n\} \subset \mathbb{R}^{D`}$$

with sample mean

$$\overline{\{x\}} = \left(\frac{1}{n}\right)\sum_{i=1}^{n} x_i$$

Define the centered data matrix

$$\tilde{X} = [x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}]^T \in R^{n\times D}$$

The empirical covariance matrix is

$$C = \frac{1}{n-1}\widetilde{X^T}\tilde{X} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$$

Dimensionality reduction in PCA is formulated as the search for a $rank - (k)$ orthogonal projection

$$W \in \mathbb{R}^{\{D\times k\}}, \qquad W^T W = I_k,$$

such that the variance of the projected data is maximized.

The low-dimensional representation $of\ (x_{i)}$ is

$$`y_i = W^T(x_i - \bar{x}) \in R^k$$

The total variance in the projected space is

$$\sum_{j=1}^{k} \text{Var}(y^j) = \text{Tr!}\,(W^T C W)$$

Hence PCA solves

$$\max_{W\in R^{D\times k}} Tr(W^T C W)$$
$$subject\ to \quad W^T W = I_k$$

Introducing the Lagrangian

$$\mathcal{L}(W, \Lambda) = (W^T C W) - Tr\left(\Lambda(W^T W - I_k)\right)$$

where $`\Lambda \in Ð\text{o}\{R\}^{\{k\ \backslash times\ k\}`}$ is symmetric, the stationary condition gives

$$\frac{\partial \mathcal{L}}{\partial W} = 2CW - 2W\Lambda = 0$$

hence

$$CW = W\Lambda$$

Thus, the columns of W are eigenvectors of C If

$$Cv_j = \lambda_j v_j, \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D \geq 0,$$

then the optimal projection matrix is

$$W_k = [v_1, v_2, \ldots, v_k]$$

The reduced coordinates are therefore

$$`y_i = W_k^T(x_i - \bar{x})$$

The variance retained by the \(k\)-dimensional representation is

$$\sum_{j=1}^{k} \lambda_j$$

and the proportion of explained variance is

$$\eta_k = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{D} \lambda_j} \quad (11)$$

Equivalently, PCA may be expressed as a minimum reconstruction error problem:

$$\min_{W \in R^{D \times k}} \sum_{i=1}^{n} |(x_i - \bar{x}) - WW^T(x_i - \bar{x})|_2^2$$

subject to

$$W^T W = I_k$$

Using matrix notation, this becomes

$$\min_{W^T W = I_k} |\tilde{X} - \tilde{X}WW^T|_F^2$$

By the Eckart–Young theorem, the minimizer is again given by the subspace spanned by the top \(k\) eigenvectors of C, or equivalently by the k right singular vectors of $\tilde{X}$

If the singular value decomposition of the centered data matrix is

$$\tilde{X} = U\Sigma V^T,$$

then

$$C = \frac{1}{n-1} V\Sigma^2 V^T$$

so that the principal directions are the columns of V, and

$$\lambda_j = \frac{\sigma_j^2}{n-1}$$

The k-dimensional PCA embedding may thus also be written as

$$Y = \tilde{X}W_k = \tilde{X}V_k$$

For any unit vector $w \in R^D$, the variance of the scalar projection$w^T(x_i - \bar{x})$

$$Var(w^T x) = w^T C\, w$$

and the first principal component solves

$$v_1 = \arg \max_{||w||_2 = 1} w^T C\, w$$

Recursively, the mth component solves

$$v_m = \arg \max_{||w||_2 = 1, w^T v_j = 0, j < m} w^T C\, w \quad (20)$$

The orthogonal projector onto the principal subspace is

$$P_k = W_k W_k^T$$

Thus, the PCA reconstruction of x_i is

$$\hat{x_i} = \bar{x} + P_k(x_i - \bar{x}) = \bar{x} + W_k W_k^T(x_i - \bar{x})$$

The corresponding residual is

$$r_i = (I_D - W_k W_k^T)(x_i - \bar{x})$$

with total residual energy

$$`\sum_{i=1}^{n} ||r_i||_2^2 = (n-1) \sum_{j=k+1}^{D} \lambda_j$$

Hence PCA yields the optimal linear k-dimensional approximation in the sense that

$$\min_{\text{rank}(P)=k, P^2=P, P^T=P} \sum_{i=1}^{n} |(x_i - \bar{x}) - P(x_i - \bar{x})|^2 = (n-1) \sum_{j=k+1}^{D} \lambda_j$$

Geometrically, PCA assumes that the data are concentrated near an affine subspace

$$\mathcal{A}_k = \bar{\mathrm{x}} + \mathrm{span}\{\mathrm{v}_1, \ldots, \mathrm{v}_\mathrm{k}\}$$

and selects $\mathcal{A}_k$such that the orthogonal projection error is minimized.

Compact High-Impact Mathematical Interpretation

$$\text{PCA: } R^D \to R^k, x \mapsto W_k^T(x - \bar{x})$$

with

$$W_k = \arg \max_{W^T W = I_k} \mathrm{Tr}(W^T C W) = arg \min_{W^T W = I_k} |\tilde{X} - \tilde{X}WW^T|F^2$$

Thus, PCA preserves maximal global variance but only under the constraint of a linear orthogonal embedding, implying that nonlinear manifold curvature is not explicitly modeled:

$x_i \in \mathcal{M} \subset R^D,\ \mathcal{M} \subset R^k,$
PCA induces geometric distortion

VII. CONCLUSION

This article adopts a grave geometric approach to dimensionality reduction in high dimensionality spaces, not the problem of a computationally simple nature, but a fundamentally geometrically and epistemologically meaningful change. The exploration is framed around the concept of "complexity crisis", which has the potential to induce such phenomena that the classical statistics and visualisation are constrained, to an extreme extent, like in high dimensional spatial data due to concentration of distances, volume distortion, with exponential sampling needed. To do so, authors propose the manifold hypothesis as a unified paradigm to demonstrate that high dimensional data are typically observed in low-dimensional, smoothly embedded manifolds in the data world. This result helps lend theoretical justification for dimensionality reduction while rejecting analysis of ambient space and highlights the necessity of intrinsic geometric structure. It is separated from Euclidean and geodesic distances, which is a persistent thesis, suggesting that meaningful data relations depend less on ambient proximity than on intrinsic manifold geometry. Additionally, a comparison of the global and local dimensionality reduction methods demonstrates that none is able to do these processes more efficiently than the other. Global methods such as PCA can accommodate high variance, but are not able to take into account the nonlinear curvature, and local methods concentrate on the structure of the neighborhood and not fidelity at the whole region. This unavoidable tension is a larger geometric compromise between the preservation of a uniform local context and global structure. And, the paper too contributes significantly by taking a step towards dealing with the formality of information loss in dimensionality reduction. The framework shows that in its frameworks including the Johnson–Lindenstrauss lemma and rate-distortion theory, any low-dimensional embedding simply needs to involve distortion (i.e., one with fundamental constraints on fidelity can work). Therefore, dimensionality reduction is not a simple transformation either, it is an irreversible transformation of the geometry of data. Most prominently, it pulls back the curtain on geometric hallucination: in other words, dimensionality reduction techniques produce artificial structures phantom clusters or false separations, say — that do not exist in the original high-dimensional space of the object. One of the general principles to visualizations is to be cautious and respectful of the interpretation of visualizations since visually interesting structures may not mirror a structural property. Dimensionality reduction can be viewed as a tradeoff between interpretability and fidelity and is conditioned by the deep geometric and information-theoretic constraints of (and is held back by) modern mathematics. Compared to this high dimensionality view of reality, reduced representations are ordered approximations that we are required to read carefully. Future research will be fruitful in identifying methods that quantitatively quantify distortion, leverage domain knowledge, and involve effective methods for deriving the real structure and the artifacts created by algorithms. Ultimately, however, advancing the field will also depend on the increasing geometrical literacy underpinning the examination of larger and larger datasets — if we can hope it won't sit idle waiting for algorithmic breakthroughs.

REFERENCES

1. Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In T. Van den Bussche & V. Vianu (Eds.), Database theory—ICDT 2001 (Lecture Notes in Computer Science, Vol. 1973, pp. 420–434). Springer.
2. Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798–1828.
3. Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, 15(6), 1373–1396.

4. Coifman, R. R., & Lafon, S. (2006). Diffusion maps. Applied and Computational Harmonic Analysis, 21(1), 5–30.
5. Fefferman, C., Mitter, S., & Narayanan, H. (2013). Testing the manifold hypothesis. Journal of the American Mathematical Society. Advance online publication.
6. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.
7. Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. Science, 290(5500), 2323–2326. https://doi.org/10.1126/science.290.5500.2323
8. Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. Science, 290(5500), 2319–2323.
9. van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9, 2579–2605.
10. Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. Foundations and Trends in Machine Learning, 1(1–2), 1–305.
11. Ware, C. (2004). Information visualization: Perception for design (2nd ed.). Morgan Kaufmann.