
Measure Before You Look: Grounding Embeddings Through Manifold Metrics

Anonymous Author(s)

Affiliation

Address

email

Abstract

Dimensionality reduction methods are routinely employed across scientific disciplines to make high dimensional data amenable to analysis. Despite their widespread use, we often lack tools to assess whether their resulting embeddings are faithful to the underlying manifold structure. Without a rigorous quantitative assessment of an embedding’s structural properties, it is difficult to quantify their degree of preservation or distortion of the underlying manifold structure of the data. We introduce a complementary suite of geometric metrics to quantitatively audit embedding fidelity across neighborhood sizes: Tangent Space Approximation (TSA), Local Intrinsic Dimensionality (LID), and Participation Ratio (PR). We compare the dimensionality of each sample before and after embedding, where points that preserve similar values across transformations are deemed to be geometrically faithful and thus, representative of true manifold structure in the data.

Across synthetic and biological datasets, we show that these metrics expose distinct embedding failure modes: TSA is most sensitive to small-scale geometric distortions, LID captures heterogeneity in mixed-density regions, and PR diagnoses global variance structure. Finally, we demonstrate that applying Jacobian Frobenius penalties during autoencoder refinement of embeddings contracts tangent spaces, reduces disagreement between metrics, and improves alignment with intrinsic manifold geometry. We motivate moving beyond visual heuristics and making principled, geometry-based choices to inform method selection, improve representations and motivate geometry-aware objectives for representation learning.

1 Introduction

Dimensionality reduction is used for visualization, clustering, and downstream learning across numerous domains ranging from single-cell transcriptomics to population genetics [1–3]. When projecting high-dimensional data into two or three dimensions, we implicitly assume that the embedding faithfully represents the underlying manifold [4]. This reliability is rarely questioned. However, embeddings that appear to be well-separated visually can distort local neighborhoods, collapse directions of variance, or warp tangent spaces in ways that aren’t reliably visualizable in 2 or 3 dimensions [5, 6]. Common quantitative measures, such as explained variance, only provide partial guarantees. Embeddings can appear similar at a global scale but differ locally. This can lead to modeling inconsistencies. When dealing with high dimensional data, we assume it is generated from a lower-dimensional manifold. Modeling with representations that are inconsistent to this manifold geometrically but are otherwise visually consistent may result in misguided analyses.

In this work, we assess the structural fidelity of embedding spaces by integrating a complementary set of metrics spanning local and global scales, through neighborhood-based methods. We

36 measure an embedding space’s Local Intrinsic Dimensionality [7], Tangent Space Approximation [8],
 37 Participation Ratio [9], Trustworthiness and Continuity [10].

38 We contrast embeddings obtained from Principal Components Analysis (PCA) [11], which captures
 39 linear variation, against Archetypal Analysis (AA) [12] which emphasizes extreme points on a convex
 40 hull to understand how initial representation spaces influence subsequent embeddings. We then test
 41 whether embeddings can be actively refined to keep local geometry faithful across transformations
 42 via autoencoders, including variants penalizing the Frobenius norm of the Jacobian to smoothen
 43 latent geometry [13]. This introduces an explicit inductive bias: we trade off some global embedding
 44 fidelity, lost in reconstruction and through subsequent transformations, to obtain locally coherent
 45 latent spaces that better align with the intrinsic manifold structure of the embedding’s source.

46 2 Methods

47 Our metric suite is designed to evaluate embedding spaces with a complementary set of geometric
 48 metrics that account for both local and global structure. This enables a comprehensive audit of
 49 embedding fidelity and overall manifold structure. For local metrics, we estimate the Tangent Space
 50 Approximation (TSA) by performing PCA locally within a point’s neighborhood to infer the local
 51 intrinsic dimensionality required to explain a fixed fraction of variance. Local Intrinsic Dimensionality
 52 (LID) measures the effective number of degrees of freedom that are required to describe local structure,
 53 as determined by a maximum likelihood estimate of a point’s neighbors’ distances. Additionally, we
 54 employ Participation Ratio (PR) to estimate the effective number of variance directions, or eigenvalue
 55 spread, for a given local embedding patch. All of these are computed across neighborhood sizes
 56 $k \in \{5, 15, 25, 50, 100\}$. Trustworthiness and Continuity quantify how well local neighborhoods
 57 are preserved globally between the original ambient space and transformed embedded spaces. To
 58 ensure paired comparisons, we fix a random subset of 500 evaluation points per dataset and reuse it
 59 consistently across all methods and seeds.

60 2.1 Datasets

61 We evaluate dimensionality reduction methods on two datasets, chosen to highlight distinct manifold
 62 geometries, across synthetic and real world settings. We use the Swiss Roll as a controlled benchmark:
 63 a synthetic 2 dimensional manifold projected into 200 dimensions, with known geodesic distances.
 64 Second, we assess high-dimensional population genomic data from the Human Genome Diversity
 65 Project (HGDP) [14] and Thousand Genomes Project (1KGP) [15] with mixed-mode distributions,
 66 known population admixture and multi-scale variation, providing a realistic test to assess embedding
 67 fidelity. See appendix B.1 and B.2 for further details about swissroll and HGDP+1KGP respectively.

68 2.2 Experimental Pipeline

69 Our goal is to evaluate how different initial representations and refinement steps affect the preservation
 70 of manifold structure across datasets. We designed a three-stage pipeline beginning with **initial**
 71 **PCA or AA representations**. We then obtain **refined** embeddings through both unregularized
 72 and regularized autoencoder reconstructions of the input spaces. Finally, we evaluated all obtained
 73 **downstream embeddings**, at both the reconstruction and further transformation stages with our
 74 comprehensive suite of **manifold-aware metrics** and their correlations against the ambient space
 75 they were generated from.

76 **Stage 1: Initial representations.** For each dataset, we first construct two 50-dimensional baseline
 77 representations from both of our source datasets:

- 78 • **PCA₅₀**: captures directions of maximal variance via principal components.
- 79 • **AA₅₀**: emphasizes extreme points on the convex hull as anchors, points as mixtures of these.

80 **Stage 2: Representation refinement with autoencoders.** We train two types of autoencoders to
 81 reconstruct the PCA₅₀ and AA₅₀ representations:

- 82 • **AE(PCA₅₀)** and **AE(AA₅₀)** Standard or "Vanilla" autoencoders trained by minimizing
 83 reconstruction error L_{recon} on the baseline embeddings.

- $\mathbf{AE_F(PCA_{50})}$ and $\mathbf{AE_F(AA_{50})}$. Autoencoders trained with an additional penalty on the Frobenius norm of the encoder Jacobian:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_J \|J_f(x)\|_F^2$$

where $J_f(x)$ is the Jacobian of the encoder mapping f with respect to its inputs, and λ_J controls the strength of regularization.

This Frobenius penalty encourages the encoder to produce locally smooth, low-curvature latent spaces, making the refined embeddings less sensitive to noise and better aligned with intrinsic manifold geometry [13].

Stage 3: Downstream embeddings. We further reduce each of the three representations from baseline and refinement stages into 2-dimensional spaces using two common projection methods: UMAP [16] and PHATE [17], where we sweep over their respective neighborhood parameters `n_neighbors` or `knn` $\in \{5, 15, 25, 50, 100\}$ to probe method and metric sensitivity to local or global structure. We then compute the Spearman’s rank correlation from both the "refined" autoencoder and six downstream embeddings with respect to their original representation spaces, AA or PCA. See appendix C for full architectural details.

3 Results

3.1 Our metric suite reliably estimates known intrinsic dimension on controlled data

We tested if these metrics captured the dimensionality of a dataset that is known to be 2-dimensional. The Local Intrinsic Dimensionality (LID), Tangent Space Approximation (TSA) and Participation Ratio (PR) reliably capture the intrinsic dimension of a Swiss Roll in 2 dimensions. The values for TSA diverge as the number of neighbors gets larger, starting at $k = 25$ with a more pronounced effect for the AA than the PCA initial embedding space. Participation Ratio stabilizes at $k = 25$ in AA, and right after in PCA. LID shows largely consistent values with a consistent divergence margin from 2 after $k = 5$. We found that regularizing the autoencoder with a Frobenius Jacobian penalty produces embeddings that are locally smoother and consistently lower-dimensional. Full definitions can be found in Appendix A.1. The full tables for PCA and AA baselines can be found in Appendix Table 2.

Table 1: Average metric performance of baseline methods on the Swiss Roll dataset across varying numbers of neighbors (k) for both PCA and AA ambient spaces. For PR, LID, and TSA, values closer to the intrinsic dimension of the dataset (2) are better.

Method	Metric	Number of Neighbors (k)				
		5	15	25	50	100
PCA	Local Intrinsic Dim. (LID)	3.33	2.04	1.82	1.70	1.85
	Tangent Space Approx. (TSA)	1.93	2.01	2.06	2.39	2.77
	Participation Ratio (PR)	1.50	1.68	1.76	1.95	2.21
	Trustworthiness	1.00	1.00	1.00	1.00	1.00
	Continuity	0.80	0.84	0.85	0.87	0.90
AA	Local Intrinsic Dim. (LID)	3.82	2.40	2.21	2.25	2.54
	Tangent Space Approx. (TSA)	2.14	2.77	3.11	3.81	4.96
	Participation Ratio (PR)	1.59	1.94	2.12	2.45	2.90
	Trustworthiness	1.00	1.00	0.99	0.99	0.99
	Continuity	0.47	0.55	0.60	0.68	0.74

3.2 Improved local geometry propagates to downstream embeddings

To gauge consistency across embedding spaces, we use the Spearman rank correlation to quantify how well local neighborhood structure is preserved relative to the baseline embedding. See Figure 1 for LID, TSA, PR computed at $k \in \{5, 15, 25, 50, 100\}$ for all obtained embeddings. Each method’s `n_neighbors` or `knn` parameter included in the figure matches the k value at which it is evaluated.

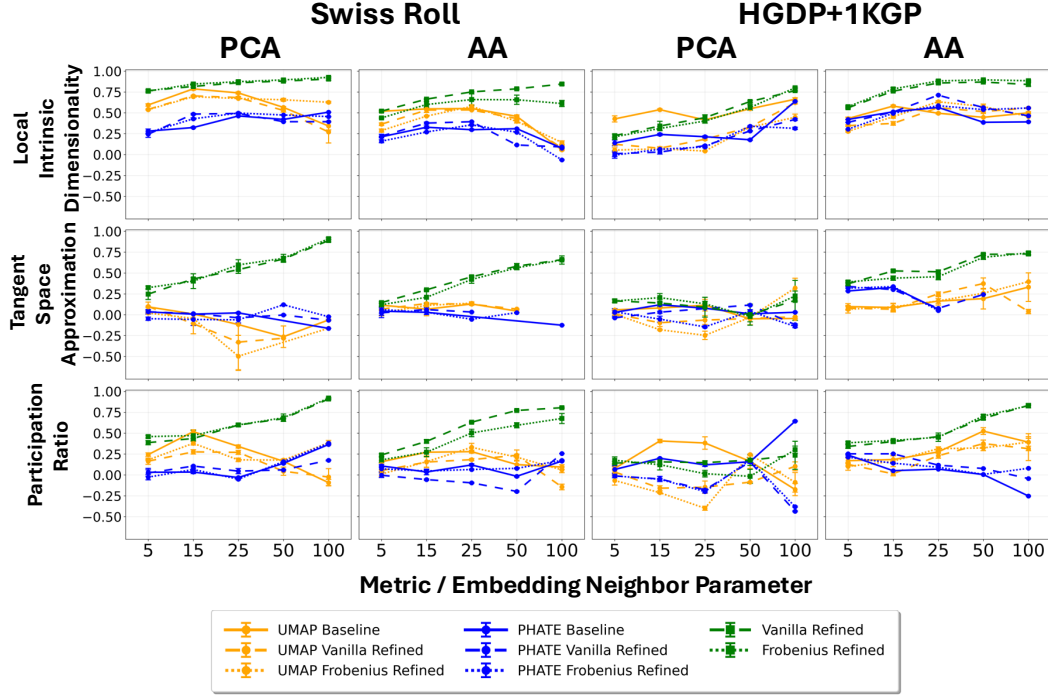


Figure 1: Correlations of average metric performance of reconstructions and downstream methods on the Swiss Roll and HGDP+1KGP datasets across varying numbers of neighbors k against both of the original PCA and AA ambient spaces.

114 Frobenius-regularized autoencoders show a higher correlation to the original space for both the Swiss
 115 Roll and heterogenous HGDP+1KGP data. Exceptions are observed for PR and LID in the HGDP
 116 PCA cases, where a UMAP computed on the $50d$ baseline outperforms it.

117 4 Discussion

118 TSA excels for small, near-linear neighborhoods, making it well-suited to measure spaces with
 119 near linear variation. LID is strongest at medium values of k in heterogeneous regions, and PR is
 120 informative at high values of k but loses resolution when variance is concentrated along few directions.
 121 This is especially apparent as noted by a divergence in metric agreement when measuring high
 122 dimensional AA spaces. As AA captures global extremal structure at the expense of local coherence,
 123 methods like TSA fail to capture local neighborhoods while LID and PR remain comparatively
 124 robust in successfully capturing intrinsic dimension. This is the center of our argument: internal
 125 geometrical coherence across embeddings isn't something that can be reliably seen with global
 126 or isolated metrics. Visual intuition builds on the former, therefore it is insufficient as a measure
 127 of embedding consistency; embeddings must be thoroughly audited before they're trusted to be
 128 representative of their data.

129 Embedding spaces can be actively steered via inductive biases like encoder-Jacobian Frobenius
 130 penalties. This geometric regularization contracts tangent spaces, reduces metric disagreement,
 131 and improves alignment between representations. The benefit starts to become apparent when
 132 the neighborhoods grow in heterogeneity, beyond tight local spaces at $k = 25$. Systematically
 133 quantifying these trade-offs establishes a foundation for principled, geometry-aware embedding
 134 refinement. This is essential for auditing representations produced by modern foundation models
 135 in a principled, geometry-aware way. Will implementing TSA/PR/LID penalties in the loss terms
 136 of existing encoders when fine-tuning allow us to reconstruct faithful, easily traversable manifolds?
 137 Can measuring geometric embedding quality be a proxy to establish favorable or "well-behaved"
 138 training dynamics? Knowing if methods favor consistent learned representations will allow us to be
 139 intentional about what type of geometry we propagate through in our representations.

References

- [1] The geometry of admixture in population genetics: The blessing of dimensionality. *Journal Name Unknown*, 2024. Accessed: 2025-04-14.
- [2] Cyril de Bodt, Alex Diaz-Papkovich, Michael Bleher, and et al. Low-dimensional embeddings of high-dimensional data, 2025.
- [3] Alex Diaz-Papkovich, Luke Anderson-Trocmé, and Simon Gravel. A review of umap in population genetics. *Journal of Human Genetics*, 66(1):85–91, October 2020.
- [4] Bradley C. A. Brown, Anthony L. Caterini, Brendan Leigh Ross, Jesse C. Cresswell, and Gabriel Loaiza-Ganem. Verifying the union of manifolds hypothesis for image data, July 2022.
- [5] Tara Chari and Lior Pachter. The specious art of single-cell genomics. *PLoS Comput. Biol.*, 19(8):e1011288, August 2023.
- [6] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 2016.
- [7] Laurent Amsaleg, Oussama Chelly, Michael E. Houle, Ken-Ichi Kawarabayashi, Miloš Radovanović, and Weeris Treeratanajaru. Intrinsic dimensionality estimation within tight localities: A theoretical and experimental analysis. *arXiv preprint arXiv:2209.14475*, September 2022.
- [8] Uzu Lim, Harald Oberhauser, and Vidit Nanda. Tangent space and dimension estimation with the wasserstein distance. <https://people.maths.ox.ac.uk/nanda/source/DimDet.pdf>. Accessed: 2025-04-14.
- [9] Stefano Recanatesi, Serena Bradde, Vijay Balasubramanian, Nicholas A. Steinmetz, and Eric Shea-Brown. A scale-dependent measure of system dimensionality. *Patterns (N.Y.)*, 3(8):100555, August 2022.
- [10] Jarkko Venna and Samuel Kaski. Neighborhood preservation in nonlinear projection methods: An experimental study. In *Artificial Neural Networks — ICANN 2001*, pages 485–491. Springer, 2001.
- [11] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572, 1901.
- [12] Finding archetypal spaces using neural networks. <https://arxiv.org/abs/1901.09078>, 2019. arXiv preprint.
- [13] Judy Hoffman, Daniel A. Roberts, and Sho Yaida. Robust learning with jacobian regularization. *arXiv preprint arXiv:1908.02729*, August 2019.
- [14] Siwei Chen, Laurent C. Francioli, Julia K. Goodrich, Ryan L. Collins, Masahiro Kanai, and et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature*, 625(7993):92–100, December 2023.
- [15] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 2015.
- [16] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, February 2018.
- [17] Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 2019.

A Details of Measurements Used

A.1 Common notation and metric definitions.

For each point x_i , let $N_k(x_i)$ be its k nearest neighbors (excluding x_i itself) in the ambient or embedded space. Form the local data matrix $X^{(i)} \in \mathbb{R}^{k \times D}$ by centering those neighbors, then compute its covariance

$$C^{(i)} = \frac{1}{k} (X^{(i)})^\top X^{(i)}.$$

Let $\lambda_1^{(i)} \geq \lambda_2^{(i)} \geq \dots \geq \lambda_D^{(i)}$ be the eigenvalues of $C^{(i)}$. We will build each metric from these $\{\lambda_j^{(i)}\}$.

Participation Ratio (PR). The local participation ratio measures the effective number of dimensions utilized in a neighborhood. It is defined by

$$\text{PR}(x_i) = \frac{(\sum_{j=1}^D \lambda_j^{(i)})^2}{\sum_{j=1}^D (\lambda_j^{(i)})^2}.$$

Tangent Space Approximation (TSA). TSA estimates local intrinsic dimension by asking ‘‘How many eigenvalues are needed to capture at least a fraction q of the total variance?’’ Concretely:

$$d_{\text{TSA}}(x_i; q) = \min \left\{ d : \frac{\sum_{j=1}^d \lambda_j^{(i)}}{\sum_{j=1}^D \lambda_j^{(i)}} \geq q \right\},$$

where in our experiments $q = 0.95$.

Local Intrinsic Dimensionality (LID). Unlike variance-based measures, LID estimates local dimensionality via a Maximum Likelihood framework based on neighbor distances. Let $d_j(x_i)$ be the distance to the j -th nearest neighbor of x_i , and let $d_k(x_i)$ be the distance to the k -th (i.e., furthest) neighbor in the set $N_k(x_i)$. The LID at point x_i is the maximum likelihood estimate, given by:

$$\hat{d}_{\text{LID}}(x_i) = \left(-\frac{1}{k} \sum_{j=1}^k \log \frac{d_j(x_i)}{d_k(x_i)} \right)^{-1}.$$

Continuity. Continuity measures the proportion of original high-dimensional neighbors preserved in the embedding:

$$\text{Continuity}(k) = \frac{1}{n} \sum_{i=1}^n \frac{|N_k(x_i) \cap N'_k(x_i)|}{k},$$

where $N'_k(x_i)$ is the set of k -nearest neighbors of x_i in the *embedded* space.

Trustworthiness. Trustworthiness penalizes neighbors in the embedding that weren’t true neighbors in the original space:

$$\text{Trustworthiness}(k) = 1 - \frac{2}{n k (2n - 3k - 1)} \sum_{i=1}^n \sum_{j \in N'_k(x_i) \setminus N_k(x_i)} (r(i, j) - k),$$

where $r(i, j)$ is the rank of j among the original high-dimensional neighbors of i .

A.2 Rank-based Comparison

We evaluate consistency between embeddings and their original high-dimensional spaces through rank-based correlation analyses. Given corresponding metrics M_i and M'_i for each point i in the original and embedded spaces respectively, Spearman’s correlation ρ is:

$$\rho = 1 - \frac{6 \sum_{i=1}^n (r_i - r'_i)^2}{n(n^2 - 1)}, \quad (1)$$

where r_i and r'_i denote the ranks of metric values M_i and M'_i respectively.

207 **B Dataset details and pre-processing**

208 **B.1 Swiss roll**

209 We constructed a synthetic Swiss roll dataset to evaluate the capability of dimensionality reduction
210 methods. The Swiss roll is designed to represent a two-dimensional plane (y, t) that is smoothly
211 embedded into a three-dimensional space (x, y, z) through a spiral transformation. By default, the
212 dataset consists of 100 distributions, each with 50 points, resulting in a total of 5000 samples. For each
213 distribution, we sample random means for t and y , and add Gaussian noise to introduce variability
214 around these means. These (y, t) coordinates encode the intrinsic geometry of the data manifold. We
215 embed the 2D manifold into 3D space using:

$$x = t \cdot \cos t, \quad z = t \cdot \sin t,$$

216 while the y coordinate retains its noisy values. This transforms the flat plane into a 3D Swiss
217 roll. After this transformation, additional Gaussian noise is added to all 3D coordinates to simulate
218 observation or measurement noise.

219 To increase the complexity and evaluate robustness to ambient dimensionality, we further embed the
220 dataset into a 200-dimensional space by applying a random orthogonal transformation. This projection
221 preserves pairwise distances while eliminating any axis alignment, simulating high-dimensional
222 real-world scenarios.

223 **B.2 HGDP + 1KGP**

224 We combined HGDP and 1KGP whole-genome sequencing (WGS) 30X release [15]. To limit
225 the genetics markers to a set of good quality and informative polymorphisms, we intersected this
226 dataset positions contained on a largely used genotyping array (Illumina’s GSAMD-24v3). This set
227 of positions was then extracted from the 1000G dataset and additional filters were applied on this
228 resulting intersection, including a maximum missing rate of 10% and minor allele frequency (MAF)
229 5% thresholds. We performed linkage disequilibrium (LD) pruning on the filtered 1000G dataset
230 (plink v1.9 –indep-pairwise 50 5 0.5 parameters), followed by the exclusion of the human leukocyte
231 antigen (HLA) region. The remaining missing data was imputed using ShapeIT5.

232 **C Model architecture and tuning**

233 **C.1 Vanilla Autoencoder (AE) Configuration**

234 For baseline comparison, a standard autoencoder was trained for 150 epochs to reconstruct the
235 50-dimensional PCA and AA embeddings. The network’s encoder mapped the 50-dimensional input
236 to a 10-dimensional latent space through hidden layers of 32 and 16 neurons. The decoder then
237 reconstructed the original 50-dimensional vector from this latent representation. The model was
238 optimized using a standard Mean Squared Error (MSE) reconstruction loss, without any additional
239 regularization on the latent space geometry. An early stopping validation loss was set with a patience
240 of 10 epochs.

241 **C.2 Autoencoder with Frobenius Penalty (AE_F) Configuration**

242 To investigate the effect of smoothing regularization, a second autoencoder (AE_F) was trained for 150
243 epochs on the 50-dimensional PCA embeddings. This model used an identical network architecture,
244 with an encoder compressing the input to a 10-dimensional bottleneck via hidden layers of 32 and
245 16 neurons. Its objective function was augmented with a Jacobian penalty. The model was trained
246 to minimize a composite loss consisting of the MSE reconstruction error and a regularization term
247 based on the squared Frobenius norm of the encoder’s Jacobian, with the penalty weighted by a factor
248 of $\lambda = 0.01$.

Table 2: Comparison of PCA and AA Baselines with Autoencoder Refinement and Frobenius Regularization. Metrics reported for varying neighborhood sizes (k). Note: values for LID $k = 15$ for both autoencoder cases were interpolated due to a mathematical error in computation.

Method	Metric	k	Original	AE Refined	AE Frobenius
AA	Local Intrinsic Dim.	5	3.82	3.63	3.75
		15	2.20	2.16	2.21
		25	2.20	2.16	2.03
		50	2.19	2.16	1.98
		100	2.41	2.40	2.15
	Participation Ratio	5	1.52	1.49	1.51
		15	1.88	1.85	1.84
		25	2.10	2.07	2.03
		50	2.41	2.36	2.32
		100	2.90	2.81	2.75
	Tangent Space Approx.	5	2.08	2.01	2.06
		15	2.82	2.70	2.71
		25	3.23	3.10	3.08
		50	4.01	3.82	3.71
		100	5.35	5.07	4.77
	Trustworthiness	5	0.99	0.99	0.99
		15	0.98	0.97	0.98
		25	0.97	0.96	0.97
		50	0.96	0.95	0.96
		100	0.94	0.94	0.94
	Continuity	5	0.47	0.48	0.46
		15	0.54	0.55	0.54
		25	0.58	0.59	0.57
		50	0.64	0.64	0.63
		100	0.68	0.69	0.67
PCA	Local Intrinsic Dim.	5	4.87	3.89	3.85
		15	2.88	2.41	2.39
		25	2.46	2.13	2.13
		50	2.07	1.90	1.88
		100	1.81	1.67	1.65
	Participation Ratio	5	1.65	1.57	1.55
		15	2.03	1.77	1.76
		25	2.06	1.78	1.79
		50	2.11	1.92	1.88
		100	2.53	1.92	1.88
	Tangent Space Approx.	5	2.17	2.07	2.05
		15	2.79	2.29	2.28
		25	3.04	2.30	2.31
		50	3.40	2.49	2.47
		100	3.53	2.69	2.58
	Trustworthiness	5	0.98	0.98	0.98
		15	0.98	0.98	0.98
		25	0.98	0.98	0.98
		50	0.98	0.97	0.97
		100	0.97	0.97	0.97
	Continuity	5	0.50	0.46	0.47
		15	0.55	0.51	0.52
		25	0.58	0.54	0.55
		50	0.63	0.60	0.60
		100	0.69	0.66	0.67