IsoScore: Measuring the Uniformity of Embedding Space Utilization

William Rudman[†], **Nate Gillman**[‡], **Taylor Rayne**[*], **Carsten Eickhoff**[†]
Department of Computer Science, Brown University[†]
Department of Mathematics, Brown University[‡]
Quest University[*]
{william_rudman, ngillman, carsten}@brown.edu
taylor.rayne@questu.ca

Abstract

The recent success of distributed word representations has led to an increased interest in analyzing the properties of their spatial distribution. Several studies have suggested that contextualized word embedding models do not isotropically project tokens into vector space. However, current methods designed to measure isotropy, such as average random cosine similarity and the partition score, have not been thoroughly analyzed and are not appropriate for measuring isotropy. We propose IsoScore: a novel tool that quantifies the degree to which a point cloud uniformly utilizes the ambient vector space. Using rigorously designed tests, we demonstrate that IsoScore is the only tool available in the literature that accurately measures how uniformly distributed variance is across dimensions in vector space. Additionally, we use IsoScore to challenge a number of recent conclusions in the NLP literature that have been derived using brittle metrics of isotropy. We caution future studies from using existing tools to measure isotropy in contextualized embedding space as resulting conclusions will be misleading or altogether inaccurate.

1 Introduction & Background

The first step in any natural language processing pipeline is to represent text in a vector space. Understanding how contextualized word embedding models project tokens into vector space is crucial for advancing the field of natural language processing. Several recent studies analyzing the spatial distribution of contextualized word embeddings claim that the point clouds induced by models such as BERT or GPT-2 do not uniformly utilize all dimensions of the vector space they occupy (Ethayarajh, 2019; Mickus et al., 2019; Cai et al., 2021; Coenen et al., 2019b; Gao et al., 2019).

Figure 1 illustrates a two-dimensional disk that uniformly utilizes the x and y axes in two-dimensional space, but does not uniformly utilize all dimensions when embedded into three dimensions.

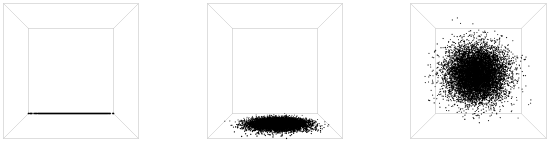

Figure 1: From left to right, a line, disk, and ball embedded in 3D space.

A distribution is *isotropic* when variance is uniformly distributed across all dimensions. Namely, a distribution is fully isotropic when the covariance matrix is proportional to the identity matrix. Several authors suggest that isotropy correlates with improved performance of embedding models (Biś et al., 2021; Wang et al., 2019; Coenen et al., 2019a; Gong et al., 2018; Hasan and Curry, 2017; Hewitt and Manning, 2019; Liang et al., 2021; Zhou et al., 2019, 2021). However, current methods of measuring the spatial utilization of contextualized embedding models do not truly measure isotropy. The most commonly used methods for measuring spatial distribution in embedding spaces include average random cosine similarity, the partition score, variance explained and intrinsic dimensionality estimation. In Section 5 we argue that all current methods of measuring isotropy have fundamental shortcomings that render them inadequate measures of spatial distribution.

To overcome these limitations, we introduce *IsoScore*: a novel tool for measuring the extent to which the variance of a point cloud is uniformly distributed across all dimensions in vector space. In contrast to previous attempts of measuring isotropy, IsoScore is the first score that incorporates the *mathematical definition* of isotropy into its formulation. As a result, IsoScore has the following desirable properties that surpass the capabilities of existing metrics: (i) It is a global measure of how

Findings of the Association for Computational Linguistics: ACL 2022, pages 3325 - 3339

uniformly distributed points are in vector space that is robust to changes in the distribution mean and scalar changes in covariance; (ii) It is rotation invariant; (iii) It increases linearly as more dimensions are utilized; and (iv) It is not skewed by highly isotropic subspaces within the data. This paper makes the following novel contributions.

1. This paper outlines essential conditions for measuring isotropy and uses a testing suite to empirically verify if a given method meets these conditions.

2. We highlight fundamental shortcomings of state-of-the-art tools and demonstrate that none of the existing methods accurately measure isotropy.

3. We present IsoScore, the first rigorously defined method for measuring isotropy in point clouds of data.

4. We share an efficient Python implementation of IsoScore with the community.[1]

The remainder of this paper is structured as follows: Section 2 reviews previous works attempting to study isotropy in contextualized word embeddings. Section 3 formally defines isotropy and describes existing tools in detail. The formal definition of IsoScore is presented in Section 4 and in Section 5, we report empirical results from experiments on contextualized word embeddings. Finally, Section 6 concludes with an outlook on future directions of work.

2 Related Work

2.1 Word Embeddings

In recent years, there has been an increased interest in analyzing the spatial organization of point clouds induced by word embeddings (Biś et al., 2021; Mickus et al., 2019; Ethayarajh, 2019; Coenen et al., 2019b; Cai et al., 2021; Mu et al., 2017; Liang et al., 2021). Several studies have concluded that contextualized embeddings form highly anisotropic, "narrow cones" in vector space (Ethayarajh, 2019; Cai et al., 2021; Gao et al., 2019; Gong et al., 2018). The most prevalent tools used to quantify the geometry of word embedding models calculate the average cosine similarity of a small number of randomly sampled pairs of points in embedding space. Ethayarajh (2019) claims that in some cases, contextualized embedding models have an average random cosine similarity that approaches 1.0, meaning all points are oriented in the same direction in space irrespective of their syntactic or semantic function.

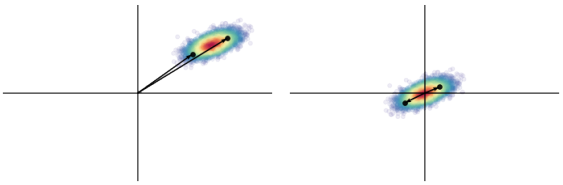

Figure 2: *Left:* Point cloud $X \subset \mathbb{R}^2$. *Right:* Result of applying a zero-mean transform to X.

In Section 5, we demonstrate that both average random cosine similarity and the partition score are significantly influenced by the mean of the data irrespective of how data points are distributed in vector space. Namely, if we normalize data to have zero-mean, average random cosine similarity and the partition score will artificially produce a score that reflects maximal isotropy. Figure 2 demonstrates that a applying a zero-mean transform to a point cloud increases the angle of randomly sampled points. Accordingly, the average random cosine of the left point cloud in Figure 2 approaches 1 while the average random cosine similarity of the right point cloud approaches 0. It is well known that word embedding models have non-zero mean vectors (Yonghe et al., 2019; Liang et al., 2021). In the case of GPT-2 embeddings obtained from the WikiText-2 corpus (Merity et al., 2016), we find values in the mean vector range from -32.36 to 198.19. Although cosine similarity has long been used to capture the "semantic" differences between words in static embeddings, adapting any cosine similarity-based methods to measure isotropy obscures the true distribution of contextualized word embeddings.

2.2 Existing Methods

We briefly review the most commonly used tools to measure the spatial distribution of point clouds $X \subseteq \mathbb{R}^n$. A mathematical exposition of these tools can be found in Appendix B.

Average Random Cosine Similarity: We define the *Average Random Cosine Similarity Score* as 1 minus the average cosine similarity of $N = 100,000$ randomly sampled pairs of points from X. **Note:** for ease of comparison to other methods, we

[1] *https://github.com/bcbi-edu/p_eickhoff_isoscore*. Alternatively: `pip install IsoScore`.

Test	IsoScore	AvgRandCosSim	Partition	ID Score	VarEx
1. Mean Agnostic	✓	✗	✗	✓	✓
2. Scalar Covariance	✓	✗	✗	✓	✓
3. Maximum Variance	✓	✗	✓	✗	✗
4. Rotation Invariance	✓	✓	✗	✓	✓
5. Dimensions Used	✓	✗	✗	✗	✗
6. Global Stability	✓	✗	✓	✓	✗

Table 1: Performance of current methods for measuring spatial utilization.

calculate 1 minus the absolute value of the average random cosine similarity so that 0 would indicate minimal isotropy and 1 would indicate maximal isotropy. We demonstrate in Section 5 that average random cosine similarity is not a measure of isotropy.

Partition Isotropy Score: Mu et al. (2017) define this score to be a particular quotient involving the partition function first proposed by Arora et al. (2015): $Z(c) := \sum_{x \in X} \exp(c^{\mathrm{T}} x)$, where c is carefully chosen from the eigenspectrum of XX^{T}. It is believed that a score closer to 0 indicates an anisotropic space, while a score near 1 indicates an isotropic space. We refer to this as the *Partition Score*.

Intrinsic Dimensionality: Algorithms for estimating intrinsic dimensionality aim to compute the true dimension of a given manifold from which we assume a point cloud has been sampled. Intrinsic dimensionality has been used to argue word embedding models are anisotropic (Cai et al., 2021). We use the MLE method to calculate intrinsic dimensionality (Levina and Bickel, 2004). Dividing the intrinsic dimensionality of $X \subseteq \mathbb{R}^n$ by n provides us with a normalized score of isotropy, which we refer to as the *ID Score*.

Variance Explained Ratio: The variance explained ratio, which we refer to as the *VarEx Score*, measures how much total variance is explained by the first k principal components of the data. We compute this by dividing the variance explained by the first k principal components by k/n. The VarEx Score requires us to specify *a priori* the number of principal components we wish to examine, which makes comparisons between vector spaces with different dimensions difficult and results in undesirable behavior, particularly when the dimension of the vector space is large.

Section 5 demonstrates that all existing methods have fundamental shortcomings that make them unreliable measures of spatial distribution. Using any of the above existing tools to make claims about isotropy will be misleading as none of the described methods truly measure isotropy.

3 Measuring Embedding Space Utilization

3.1 Definition of Isotropy

A distribution is *isotropic* if its variance is uniformly distributed across all dimensions. Namely, the covariance matrix of an isotropic distribution is proportional to the identity matrix. Conversely, an *anisotropic* distribution of data is one where the variance is dominated by a single dimension. For example, a line in n-dimensional vector space is maximally anisotropic. Robust isotropy metrics should return maximally isotropic scores for balls and minimally isotropic (i.e. anisotropic) scores for lines. Appendix D provides a geometric interpretation of "medium isotropy". We interpret a medium isotropic space in $\mathbb{R}^n$ to be one where the data uniformly utilizes approximately $n/2$ dimensions in space as defined below. Note that we exclude two edge cases for measuring isotropy. Firstly, since isotropy is a property of the covariance matrix of a distribution, the dimensionality of the space needs to be greater than 1. Secondly, we do not consider the extreme case where the data consists of a single point.

3.2 Dimensions utilized

Given a point cloud $X \subseteq \mathbb{R}^n$, we measure how many dimensions of $\mathbb{R}^n$ are truly utilized by X. For example, we denote by $I_n^{(k)}$ the $n \times n$ covariance matrix where $a_{i,i} = 1$ for $i \in \{1, 2, ..., k\}$ and all other elements are 0. Note that when $k = n$, we recover the identity matrix. Thus, $I_n^{(k)}$ represents a covariance matrix where the first k dimensions are being uniformly utilized. Figure 1 illustrates point clouds in $\mathbb{R}^3$ that have covariance matrix $I_3^{(1)}$, $I_3^{(2)}$, and $I_3^{(3)}$. These utilize 1, 2, and 3 dimensions in $\mathbb{R}^3$. To make this discussion rigorous and general, we make the following definition:

$\iota(I_9^{(1)})$	$\iota(I_9^{(2)})$	$\iota(I_9^{(3)})$	$\iota(I_9^{(4)})$	$\iota(I_9^{(5)})$	$\iota(I_9^{(6)})$	$\iota(I_9^{(7)})$	$\iota(I_9^{(8)})$	$\iota(I_9^{(9)})$
0.000	0.125	0.250	0.375	0.500	0.625	0.750	0.875	1.000

Table 2: Linearly increasing dimensions utilized in $\mathbb{R}^9$ linearly increases IsoScore. We prove in Appendix D that IsoScore satisfies the formula $\iota(I_n^{(k)}) = \frac{k-1}{n-1}$.

Definition 3.1. *Consider a point cloud $X \subseteq \mathbb{R}^n$. Let Σ be the covariance matrix of X and assume all the off-diagonal entries of Σ are zero. Let $\Sigma_D \in \mathbb{R}^n$ denote the diagonal of Σ.*

1. *We say X* utilizes k dimensions *in $\mathbb{R}^n$ if the first k entries of Σ_D are non-zero and the remaining $n-k$ entries are zero.*
2. *We say X* uniformly utilizes k dimensions *in $\mathbb{R}^n$ if X utilizes k dimensions in $\mathbb{R}^n$ and if all the non-zero entries in Σ_D are equal.*

Having a diagonal sample covariance matrix Σ implies there are no correlations between any coordinates of X. In Section 4, we reduce the general case of X to the case where the covariance matrix of X is diagonal. Figure 3 illustrates three point clouds in $\mathbb{R}^2$ that each utilize 2 dimensions. We argue that it is of practical importance to differentiate between the cases in Figure 3. The leftmost panel uniformly utilizes all dimensions of $\mathbb{R}^2$, while the rightmost panel does not uniformly utilize two dimensions of space. Note that average random cosine similarity returns maximal isotropy scores for each point cloud pictured in Figure 3.

Our proposed IsoScore reflects the dimensions utilized by a point cloud in a linear fashion. See Table 2 for a concrete example of how IsoScore reflects dimensions utilized in $\mathbb{R}^9$.

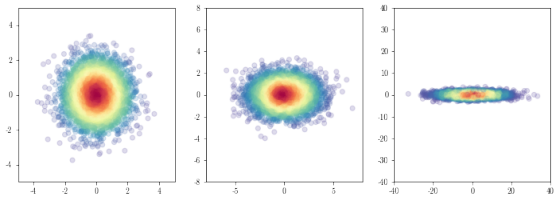

Figure 3: Points sampled from a 0 mean, 2D Gaussian with covariance $\left(\begin{smallmatrix} x & 0 \\ 0 & 1 \end{smallmatrix}\right)$ where $x = 1, 3, 75$.

3.3 Essential Properties of Isotropy

We now outline the essential properties that a measure of isotropy must possess.

1: Mean Agnostic. Recall that a distribution is isotropic if variance is uniform across all dimensions. It is essential to note that *isotropy is strictly a property of the covariance matrix* of a distribution. If changes to the mean of a distribution influence an isotropy score, then the given score does not measure isotropy.

2: Scalar Changes to the Covariance Matrix. Since isotropy is defined as the *uniformity* of variance across all dimensions, isotropy scores should not change when we multiply the covariance matrix of the underlying distribution of the data by a positive scalar value. If the covariance matrix of a distribution of data is equal to $\lambda \cdot I_n$ where $\lambda > 0$ is some scalar value and I_n is the $n \times n$ identity matrix, then a tool must return an isotropy score approaching 1.

3: Maximum Variance. As we increase the difference between the maximum variance value in our covariance matrix and the average variance value of the remaining dimensions, isotropy scores should monotonically decrease to zero. Figure 3 illustrates the effect of increasing the difference between the average variance value and the maximum value in the covariance matrix. Increasing the difference between the maximum variance value and the average variance value increases the amount of variance explained by the first principal component of the data. Namely, larger maximum variance values reduce the efficiency of spatial utilization.

4: Rotation Invariance. Given a point cloud $X \subset \mathbb{R}^n$, an ideal measure of spatial utilization should remain constant under rotations of X since the distribution of principal components remains constant under rotation. Accordingly, we consider the canonical distribution of the variance of X to be the variance after projecting X using principal component analysis. Figure 4 illustrates the process of PCA-reorientation.

5: Dimensions Used. As described in Subsection 3.2, there is a direct link between isotropy and the number of dimensions utilized by the data. Intuitively, increasing the number of dimensions uniformly utilized by the data expands the number of principal components it takes to explain all of the variance in the data. Accordingly, a good score of spatial utilization should increase linearly as we

increase the number of dimensions uniformly utilized by the data. Figure 1 depicts data utilizing one, two, and three out of three ambient dimensions, respectively.

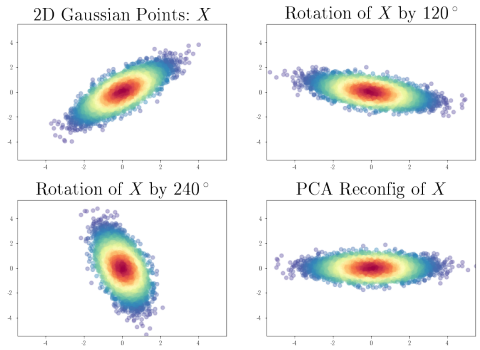

Figure 4: Left: 2D zero-mean Gaussian with covariance $\left(\begin{smallmatrix}1 & 0.8\\ 0.8 & 1\end{smallmatrix}\right)$. We rotate X by 120° and 240°, respectively. Right: Points after applying PCA reorientation.

6: Global stability. A metric of efficient spatial utilization should be a *global* reflection of the distribution. A robust method should be stable even when the data exhibits small subpopulations where a score would return an extreme value.

We test this by computing IsoScore for the union of a noisy sphere and a line; we provide a geometric rendering of this in Figure 5 in Appendix E. We refer to this test as the "skewered meatball" test. A good score of spatial distribution for a "skewered meatball" should reflect the ratio of the number of points sampled from the line and the number of points sampled from the sphere.

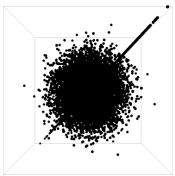

Figure 5: 2D rendering of a line in 3D space intersecting noisy sphere. AKA "skewered meatball."

In Table 1, we list which existing methods satisfy which essential conditions. Section 5 outlines the numerical experiments we execute to obtain this table. As each of the above properties have been derived from the mathematical definition of isotropy, an accurate tool for measuring isotropy needs to satisfy each essential condition.

4 IsoScore

This section introduces the proposed IsoScore metric of uniform spatial utilization.

4.1 Formal Definition of IsoScore

Algorithm 1 gives a high-level overview of the procedure. Afterwards, we discuss the individual steps in detail.

Step 1: Start with a point cloud $X \subseteq \mathbb{R}^n$. IsoScore takes as input a finite subset of $\mathbb{R}^n$ and outputs a number in the interval $[0, 1]$ that represents the extent to which X is isotropic.

Step 2: PCA-reorientation of data set. Execute PCA on X, where the target dimension remains the original n. Performing PCA reorients the axes of X so that the i'th coordinate accounts for the i'th greatest variance. Further, it eliminates all correlation between dimensions making the covariance matrix diagonal. We denote the transformed space as X^{PCA}.

Step 3: Compute variance vector of reoriented data. Compute the $n \times n$ covariance matrix of X^{PCA}; denote this matrix by Σ. Let Σ_D denote the diagonal of the covariance matrix. We refer to Σ_D as the *variance vector,* and we identify Σ_D as a vector in $\mathbb{R}^n$. Performing Step 2 causes all off-diagonal entries of the covariance matrix of X_T to vanish, which allows us to ignore off-diagonal elements for the rest of the computation.

Step 4: Length normalization of variance vector. We define the *normalized variance vector* to be

$$\hat{\Sigma}_D := \sqrt{n} \cdot \frac{\Sigma_D}{\|\Sigma_D\|},$$

where $\|(x_1, ..., x_n)\| := \sqrt{x_1^2 + \cdots + x_n^2}$ denotes the standard Euclidean norm on $\mathbb{R}^n$. Note that as a result of this normalization, we have $\|\hat{\Sigma}_D\| = \sqrt{n}$.

Step 5: Compute the distance between the covariance matrix and identity matrix. Denote the diagonal of the $n \times n$ identity matrix by $\mathbf{1} \in \mathbb{R}^n$. Then we define the *isotropy defect* of X to be

$$\delta(X) := \frac{\|\hat{\Sigma}_D - \mathbf{1}\|}{\sqrt{2(n - \sqrt{n})}}.$$

By definition of the Euclidean norm, we have $\|\hat{\Sigma}_D\| = \|\mathbf{1}\| = \sqrt{n}$. It follows from the triangle inequality that $\|\hat{\Sigma}_D - \mathbf{1}\| \in [0, 2\sqrt{n}]$. Crucially, we prove in Appendix C that achieving a value of $2\sqrt{n}$ using a valid covariance matrix is impossible. The largest value that can be attained is with the matrix $(a_{ij})_{i,j=1,...,n}$ defined by

Algorithm 1 IsoScore

begin Let $X \subset \mathbb{R}^n$ be a finite collection of points.
  Let X^{PCA} denote the points in X transformed by the first n principal components.
  Define $\Sigma_D \in \mathbb{R}^n$ as the diagonal of the covariance matrix of X^{PCA}.
  Normalize diagonal to $\hat{\Sigma}_D := \sqrt{n} \cdot \Sigma_D / \|\Sigma_D\|$, where $\|\cdot\|$ is the standard Euclidean norm.
  The isotropy defect is $\delta(X) := \|\hat{\Sigma}_D - \mathbf{1}\| / \sqrt{2(n - \sqrt{n})}$, where $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^n$.
  X uniformly occupies $\phi(X) := (n - \delta(X)^2(n - \sqrt{n}))^2 / n^2$ percent of ambient dimensions.
  Transform $\phi(X)$ so it can take values in $[0, 1]$, via $\iota(X) := (n \cdot \phi(X) - 1)/(n - 1)$.
  return: $\iota(X)$
end

$a_{11} = \sqrt{n}$ and $a_{ii} = 0$ whenever $i > 1$. One can compute that the Euclidean norm in this case is $\|\hat{\Sigma}_D - \mathbf{1}\| = \sqrt{2(n - \sqrt{n})}$. Choosing this normalization factor guarantees that $\delta(X) \in [0, 1]$, where 0 represents a perfectly isotropic space and 1 represents a perfectly anisotropic space.

Step 6: Use the isotropy defect to compute percentage of dimensions isotropically utilized. We argue in Heuristic D.1 that if X has isotropy defect $\delta(X)$, then X isotropically occupies approximately $k(X) = (n - \delta(X)^2(n - \sqrt{n}))^2/n$ dimensions in $\mathbb{R}^n$. Because $\delta(X) \in [0, 1]$, one can estimate that $k(X) \in [1, n]$ so the fraction of dimensions utilized is $\phi(X) := k(X)/n \in [1/n, 1]$.

Step 7: Linearly scale percentage of dimensions utilized to obtain IsoScore. The fraction of dimensions utilized, $\phi(X)$, is close to the final IsoScore, but it falls within the interval $[1/n, 1]$. As we want the possible range of scores to fill the interval $[0, 1]$, we apply the affine function that maps $1/n \mapsto 0$ and $1 \mapsto 1$. Thus, $S : [1/n, 1] \to [0, 1] : x \mapsto (nx - 1)/(n - 1)$. Once we compose these transformations, we obtain IsoScore:

$$\iota(X) := \frac{(n - \delta(X)^2(n - \sqrt{n}))^2 - n}{n(n - 1)}. \quad (4.1)$$

4.2 Geometric Interpretation for IsoScore

In Subsection 4.1 we described how to compute an IsoScore $\iota(X)$ for any point cloud $X \subseteq \mathbb{R}^n$. We will now present a heuristic interpretation for a given IsoScore. Intuitively, our heuristic says that $\iota(X)$ is roughly the fraction of dimensions of $\mathbb{R}^n$ utilized by X. More precisely, the quantity of dimensions of $\mathbb{R}^n$ utilized by X is some number inside the interval $[\iota(X)n, \iota(X)n + 1] \cap [1, n]$. We formalize this below.

Heuristic 4.1. *When the ambient space $\mathbb{R}^n$ has large dimension, the IsoScore $\iota(X)$ is approximately the fraction of dimensions uniformly utilized by X.*

We prove this heuristic in Appendix D. Note in particular that $\iota(X) = 0$ implies that D.1 simplifies to a single dimension utilized and $\iota(X) = 1$ implies that D.1 simplifies to all n dimensions utilized.

Because IsoScore covers a continuous spectrum, one should carefully interpret what we mean when we say that X occupies approximately k dimensions of $\mathbb{R}^n$. For example, consider the 2D Gaussian distributions depicted in Figure 3. Heuristic D.1 predicts $k = 1.9996, 1.6105, 1.0281$ dimensions are used when $x = 1, 3, 75$, respectively. These should be interpreted as follows: "when $x = 75$, the points sampled are mostly using one direction of space" and "when $x = 3$, the points sampled are using somewhere between one and two dimensions of space."

5 Experiments

In Subsection 5.1, we present results from numerical experiments designed to test each of the isotropy scores presented in this paper against the six essential properties outlined in Section 3.3. Exact descriptions of the numerical experiments are provided in Appendix E. We reiterate that each of the essential conditions have been derived directly from the mathematical definition of isotropy and violating any of the essential properties disqualifies a method from being a correct measure of isotropy.

In Subsection 5.2, we demonstrate the merit of IsoScore by recreating the experimental setup presented in (Cai et al., 2021). We create word embeddings for tokens from the WikiText-2 corpus using GPT (Radford and Narasimhan, 2018), GPT-2 (Radford et al., 2019), BERT (Devlin et al., 2018) and DistilBERT (Sanh et al., 2019) and calculate isotropy scores for each layer of the model.

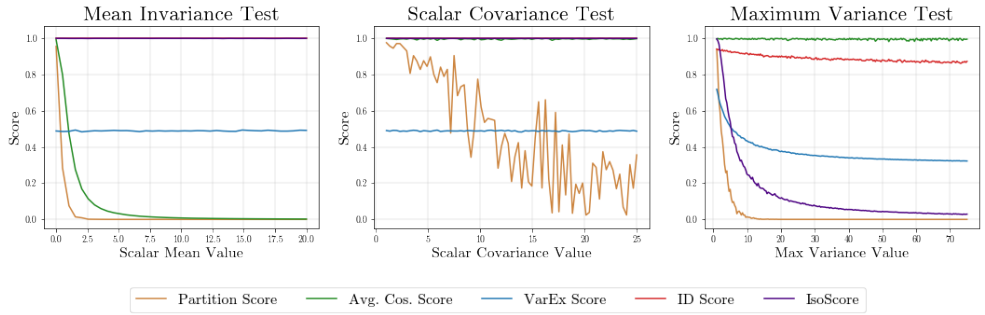

Figure 6: Left: Scores of points sampled from a 10-dimensional Gaussian with identity covariance and common mean vector ranging from 0 to 20. Center: Scores for the scalar covariance test for a 5-dimensional, zero-mean Gaussian. Right: Scores for the Maximum Variance test for 10-dimensional, zero-mean Gaussians.

5.1 Testing methods against the essential properties

Test 1: Mean Agnostic. When the covariance matrix of a distribution is proportional to the identity matrix, measures of isotropy should return a score of 1 regardless of the value of the mean. Figure 6 demonstrates that neither average random cosine similarity nor the partition score are mean-agnostic. IsoScore is mean-agnostic since it is a function of the covariance matrix. Importantly average random cosine similarity and the partition score are skewed by non-zero mean data. Our results show that, for an isotropic Gaussian with covariance matrix $\lambda \cdot I_n$ and mean vector $M = [\mu, \mu, ..., \mu]$, the average random cosine similarity of points sampled from this distribution will approach 0 as we increase the ratio between μ/λ. *Consequently, zero-centering data will cause average random cosine similarity to return maximally isotropic scores without impacting the distribution of the variance.*

Test 2: Scalar Changes to the Covariance Matrix. For a 5-dimensional Gaussian distribution with a zero mean vector and covariance matrix $\lambda \cdot I_n$, scores should reflect uniform utilization of space for any $\lambda > 0$. Figure 6 shows that IsoScore and the intrinsic dimensionality score are the only metrics that are agnostic to scalar multiplication to the covariance matrix and return a score 1 for each value of λ. In Step 4 of IsoScore, we normalize the diagonal of the covariance matrix to have the same norm as the diagonal of the identity matrix, which ensures IsoScore is invariant to scalar changes in covariance.

Test 3: Maximum Variance. An effective score should monotonically decrease to 0 as we increase the difference between the maximum variance value and average variance. Steps 4 and 5 of IsoScore ensure that the less equitably the mass in the covariance vector is distributed, the greater the isotropy defect will be. Figure 3 visualizes this phenomenon for a 2 Dimensional Gaussian. The ID Score fails this test since the intrinsic dimensionality estimate is 2.0 for all point clouds depicted in Figure 3.

Table 3: Performance of current methods on Test 4: Rotation Invariance

	IsoScore	**AvgCosSim**	**Partition**	**ID Score**	**VarEx**
X	0.216	0.990	0.445	1.000	0.500
$X^{120°}$	0.216	0.968	0.673	1.000	0.500
$X^{240°}$	0.216	0.981	0.669	1.000	0.500
X^{PCA}	0.216	0.993	0.446	1.000	0.500

Test 4: Rotation Invariance. We rotate our baseline point cloud X by $120°$ and $240°$. Lastly, we project X using PCA reorientation while retaining dimensionality to obtain a point cloud X^{PCA}. We record results in Table **??**. Only IsoScore, ID Score, and VarEx Score return constant values. The partition score would return a constant value if it were feasible to compute the true optimization problem. The approximate version of the partition score, however, depends too strongly on the basis. IsoScore is rotation invariant by design. In Step 2, IsoScore projects the point cloud of data in the directions of maximum variance before computing the covariance matrix of the data.

Test 5: Dimensions Used (Fraction of Dimensions Used Test). The number of dimensions used in a point cloud $X \subset \mathbb{R}^n$ provides a sense of how uniformly X utilizes the ambient space. A reliable metric should return scores near $0.0, 0.5$, and 1.0 when number of dimensions used is $1, \lfloor n/2 \rfloor$, and n, respectively. Figure 7 shows that only IsoScore models ideal behavior for the dimensions used test.

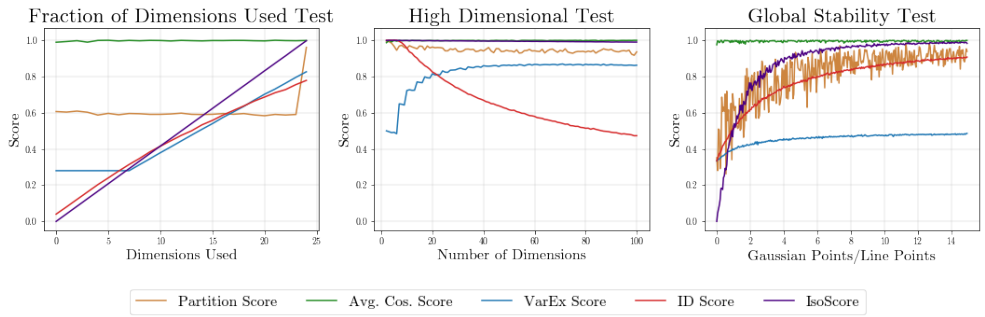

Figure 7: Left and center: Scores for the two Dimensions Used tests. Right: Scores for the "skewered meatball" test in 3 dimensions.

A rigorous explanation of why IsoScore reflects the percentage of 1s present in the diagonal of the covariance matrix is provided in Heuristic 4.1. Although the intrinsic dimensionality score monotonically increases as we increase k, it fails to reach 1 when all dimensions are uniformly utilized. Average cosine similarity fails this test, as it stays constant near 1 regardless of the fraction of dimensions uniformly utilized.

Test 5: Dimensions Used (High Dimensional Test). Metrics of spatial utilization should allow for easy comparison between different vector spaces even when the dimensionality of the two spaces is different. Figure 7 illustrates that IsoScore, the average cosine similarity score, and the partition score pass this test, as they stay constant near 1. Note that the line for IsoScore decreases slightly. By the law of large numbers, the more data points we sample from the Gaussian distribution, the closer the covariance matrix will be to the covariance matrix from which it was sampled. The VarEx Score is not stable under an increase in dimension primarily because it requires the user to specify the percentage of principal components used in calculating the score. Note that the ID Score begins to decrease simply by increasing the dimensionality of the space since the MLE method is not very well suited for estimating the intrinsic dimension of isotropic Gaussian balls.

Test 6: Global Stability. To evaluate which scores are not skewed by highly concentrated subspaces, we design the "skewered meatball test" (see Figure 5 for a geometric rendering). As we increase the ratio between the number of points sampled from a 3D isotropic Gaussian and a 1D anisotropic line, we should see isotropy scores increase from 0 to 1, and hit 0.5 precisely when the number of points sampled from the Gaussian distribution and the line are equal. Results from the skewered meatball test in Figure 7 indicate that the partition score, IsoScore and intrinsic dimensionality estimation are the only metrics that are global estimators of the data.

5.2 Isotropy in Contextualized Embeddings

Recent literature suggests that contextualized word embeddings are anisotropic. However, as demonstrated in Subsection 5.1, no existing methods truly measure isotropy. We replicate experiments by (Cai et al., 2021), and present isotropy scores for the vector space of token embeddings generated from the WikiText-2 corpus for GPT (110M parameters) and GPT2 (117M parameters) in Figure 8, as well as the scores for BERT (base, uncased) and DistilBERT (base, uncased) in Figure 9.

Figure 8: The 5 scores for each of the 12 layers of GPT-2 and GPT

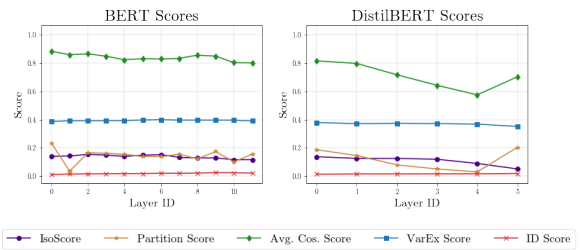

Figure 9: The 5 scores for the 12 layers of BERT, and the 6 layers of DistilBERT

Our findings using IsoScore challenge and extend upon the literature in the following ways. Contextualized embedding models (i) utilize even fewer dimensions than previously thought; (ii) do not utilize fewer dimensions in deeper layers; and (iii) in agreement with Biś et al. (2021), contextualized embedding models do not necessarily occupy a "narrow cone" in space.

IsoScore returns values of less than 0.18 for every considered contextualized embedding model. GPT and GPT-2 embeddings do not even isotropically utilize a single dimension in space, in the sense of Heuristic D.1. Using average random cosine similarity, Cai et al. concluded that earlier layers in contextualized embedding models are more isotropic than layers deeper in the network. While this may appear to be true using inaccurate measures of isotropy, there is no significant decrease in IsoScore between the earlier and later layers of contextualized embedding models. Biś et al. (2021) argue that isotropy improves performance for contextualized embedding models and that enforcing zero mean embeddings recovers "isotropy". The author's claim to improve isotropy by subtracting the mean vector from the point clouds of embeddings produced from BERT, GPT-2 and RoBERTa, however, the authors use the partition score in attempts to measure isotropy which will return values close to 1 when the data is zero-mean. As demonstrated throughout the paper, isotropy is strictly a property of the covariance matrix of a distribution and is by definition mean-agnostic.

Note that our average random cosine similarity score finds contextualized embedding models to be much more isotropic then previously reported. When computing the average random cosine similarity score for contextualized word embeddings we sample 250,000 pairs of points. Prior studies such as Ethayarajh (2019) and Cai et al. (2021) sample as few as 1000 pairs of points when calculating average random cosine similarity. In both cases, the point clouds contain millions of tokens embedded into 768 dimensional vector space and differences in reported scores are likely due to sampling noise. We found empirically that the quantity of points sampled should be orders of magnitude larger than the dimension.

The notion of isotropy is often conflated with geometry. The geometry of isotropic vector spaces, however, will differ depending on the distribution that generates the points in space. For example, multivariate isotropic Gaussians form n-dimensional balls and uniform distributions form n-dimensional cubes, yet both distributions receive an IsoScore of 1. For an illustrated example of points generated from different isotropic distributions, consult Appendix F. It is therefore not necessarily the case that even highly anisotropic embedding spaces form narrow, anisotropic cones.

6 Conclusion & Future Works

Several studies have attempted to study isotropy in contextualized embedding models. Using mathematically rigorous tests, we demonstrate that current methods do not accurately measure isotropy. This paper presents IsoScore: a novel method for measuring isotropy that corrects the current misunderstandings in the literature. IsoScore is the only tool that is mean agnostic, robust to scalar changes to the covariance matrix and rotation invariant. Furthermore, IsoScore scales linearly with the number dimensions used and is stable when distributions contain highly isotropic subspaces. Future studies should avoid using existing methods to measure isotropy as resulting conclusions will be misleading or altogether inaccurate.

There are several promising directions for future work. Current studies have used inaccurate methods to claim that increasing isotropy improves the performance of contextualized embedding models. However, we believe that further decreasing isotropy could improve performance, especially in language modeling applications. IsoScore could be used as a regularizer when fine tuning word embeddings to penalize distributions that exhibit isotropy.

As point clouds of data arise in nearly all deep learning applications, IsoScore presents itself as a useful tool to study and refine a variety of models beyond the domain of NLP.

Acknowledgements

This research is supported in part by the NSF (IIS-1956221). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF, or the U.S. Government.

References

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2015. Random walks on context spaces: Towards an explanation of the mysteries of semantic word embeddings. *CoRR*, abs/1502.03520.

Daniel Biś, Maksim Podkorytov, and Xiuwen Liu. 2021. Too much in common: Shifting of embeddings in transformer language models and its implications. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5117–5130, Online. Association for Computational Linguistics.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*.

Paola Campadelli, Elena Casiraghi, Claudio Ceruti, and Alessandro Rozza. 2015. Intrinsic dimension estimation: Relevant techniques and a benchmark framework. *Mathematical Problems in Engineering*, 2015:1–21.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, F. Viégas, and M. Wattenberg. 2019a. Visualizing and measuring the geometry of bert. In *NeurIPS*.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 2019b. Visualizing and measuring the geometry of bert.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and GPT-2 embeddings. *CoRR*, abs/1909.00512.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models.

Chengyue Gong, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2018. Frage: Frequency-agnostic word representation. *ArXiv*, abs/1809.06858.

S. Hasan and E. Curry. 2017. Word re-embedding via manifold dimensionality retention. In *EMNLP*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *NAACL-HLT*.

Elizaveta Levina and Peter J. Bickel. 2004. Maximum likelihood estimation of intrinsic dimension. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'04, page 777–784, Cambridge, MA, USA. MIT Press.

Yuxin Liang, Rui Cao, Jie Zheng, Jie Ren, and Ling Gao. 2021. Learning to remove: Towards isotropic pre-trained BERT embedding. *CoRR*, abs/2104.05274.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models.

Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees van Deemter. 2019. What do you mean, bert? assessing BERT as a distributional semantics model. *CoRR*, abs/1911.05758.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2017. All-but-the-top: Simple and effective postprocessing for word representations. *CoRR*, abs/1702.01417.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Dilin Wang, Chengyue Gong, and Qiang Liu. 2019. Improving neural language modeling via adversarial training. *ArXiv*, abs/1906.03805.

Chu Yonghe, Hongfei Lin, Liang Yang, Yufeng Diao, Zhang Shaowu, and Fan Xiaochao. 2019. Refining word representations by manifold learning. pages 5394–5400.

Tianyuan Zhou, João Sedoc, and J. Rodu. 2019. Getting in shape: Word embedding subspaces. In *IJCAI*.

Wenxuan Zhou, Bill Yuchen Lin, and Xiang Ren. 2021. Isobn: Fine-tuning bert with isotropic batch normalization. In *AAAI*.

A Interpretation: IsoScore as a Summary Statistic

We will now provide an intuitive interpretation for the IsoScore of a point cloud $X \subseteq \mathbb{R}^n$. The interested reader should consult Appendix D for an in-depth explanation of this heuristic.

Heuristic A.1. *The IsoScore of X is roughly the fraction of dimensions uniformly utilized by X.*

For example, an IsoScore near 0.5 indicates that around half of the dimensions are utilized; and more generally, an IsoScore near $\alpha \in [0, 1]$ indicates that approximately $n \cdot \alpha$ of the dimensions of $\mathbb{R}^n$ are uniformly utilized by X. Table 2 illustrates this trend where IsoScore increases linearly as more dimensions are uniformly utilized in $\mathbb{R}^9$.

B Pre-existing metrics, in detail

Average Cosine Similarity: We define the *Average Cosine Similarity Score* as 1 minus the average cosine similarity of N randomly sampled pairs of points from X. That is,

$$\text{AvgCosSim}(X) := 1 - \left| \sum_{i=1}^{N} \frac{\cos(x_i, y_i)}{N} \right|, \quad \text{(B.1)}$$

where $\{(x_1, y_1), \ldots, (x_N, y_N)\} \subseteq X \times X$ are randomly chosen with $x_i \neq y_i$ for all i, and $\cos(x_i, y_i)$ denotes the cosine similarity of x_i and y_i. Some authors define the average cosine similarity score to be exactly the average, rather than one minus the average. However, for ease of comparison to other metrics, our score ensures that $\text{AvgCosSim}(X)$ is between 0 and 1. Under our convention, it is commonly believed that a score of 0 indicates that the point cloud X is anisotropic and a score of 1 indicates that X is isotropic. In Section 5, we demonstrate that this is not the case.

Partition Isotropy Score: For any unit vector $c \in \mathbb{R}^n$, let the partition function be denoted as $Z(c) := \sum_{x \in X} \exp(c^{\mathrm{T}} x)$. Mu et al. (2017) measure isotropy as $I(X) := (\min_{||c||=1} Z(c))/(\max_{||c||=1} Z(c))$. It is believed that a score closer to zero indicates an anisotropic space while a score closer to one indicates an isotropic space. Mu et al. (2017) demonstrate that a score of 1 implies that the eigenspectrum of X is flat. Computing $I(X)$ explicitly is intractable since the set of unit vectors is infinite. Accordingly, Mu et al. (2017) approximate $I(X)$ by

$$I(X) \approx \frac{\min_{c \in C} Z(c)}{\max_{c \in C} Z(c)} \quad \text{(B.2)}$$

where C is the set of eigenvectors of $X^{\mathrm{T}} X$. For the remainder of the paper we refer to (B.2) as the *Partition Score.*

Intrinsic Dimensionality: Given a point cloud $X \subseteq \mathbb{R}^n$, it is sometimes useful to assume that X is sampled from a manifold of dimension less than n. For example, points in the left panel in Figure 1 are sampled from a 1-dimensional space and points in the middle panel are sampled from a 2-dimensional space. Algorithms for intrinsic dimensionality aim to estimate the true dimension of a given manifold from which we assume a point cloud has been sampled. Intrinsic dimensionality has been used to argue that word embedding models are anisotropic (Cai et al., 2021). For a point cloud $X \subset \mathbb{R}^n$, it is commonly thought that the more isotropic X is, the closer the intrinsic dimensionality of X is to n. Dividing the intrinsic dimensionality of X by n provides us with a normalized score of isotropy, which we refer to as the *ID Score*. We use the maximum likelihood estimation (MLE) method to calculate intrinsic dimensionality. For a detailed description of the MLE method for intrinsic dimensionality estimation please consult (Levina and Bickel, 2004; Campadelli et al., 2015).

Variance Explained Ratio: The variance explained ratio measures how much total variance is explained by the first k principal components of the data. Note that when all principal components are considered, the variance explained ratio is equal to 1. Examining the eigenspectrum of principal components is undoubtedly a useful tool in quantifying the spatial distribution of high dimensional data. However, the variance explained ratio requires us to specify *a priori* the number of principal components we wish to examine. We divide the variance explained by the first k principal components by k/n to convert the variance explained ratio into a normalized score.

C Bounds on IsoScore

Proposition C.1. *Let $X \subseteq \mathbb{R}^n$ be a finite set. Then $\iota(X) \in [0, 1]$.*

Proof. Define Σ to be the $n \times n$ sample covariance matrix of X^{PCA}. Let $c > 0$ be so that if we define $\hat{\Sigma} := c \cdot \Sigma$, then $\|\hat{\Sigma}_D\| = \sqrt{n}$. Let us enumerate the entries of this vector as $\hat{\Sigma}_D = (\text{Var}(x_1), \ldots, \text{Var}(x_n))$. In order to show that $\iota(X) \in [0, 1]$, it is equivalent to show that $\|\hat{\Sigma}_D - \mathbf{1}\| \in [0, \sqrt{2(n - \sqrt{n})}]$, and by definition of the Euclidian norm, the latter estimate is equivalent to

$$2(n - \sqrt{n}) \geq \sum_{i=1}^{n} (\text{Var}(x_i) - 1)^2. \quad \text{(C.1)}$$

But the identity $\|\hat{\Sigma}_D\| = \sqrt{n}$ implies that $\sum_{i=1}^{n} \text{Var}(x_i)^2 = n$, so in fact (C.1) is equivalent to

$$\sum_{i=1}^{n} \text{Var}(x_i) \geq \sqrt{n}.$$

If this inequality were flipped, then we could esti-

mate that

$$\begin{aligned} n &= \mathrm{Var}(x_1)^2 + \cdots + \mathrm{Var}(x_n)^2 \\ &\leq (\mathrm{Var}(x_1) + \cdots + \mathrm{Var}(x_n))^2 \\ &< n, \end{aligned}$$

which is a contradiction. □

D Interpretation of IsoScore, in Detail

This appendix provides rigorous mathematical justification for the claims that we made in Appendix A about the interpretation of IsoScore. It is split into two parts. In Appendix D.1 we formalize, and prove, the claim that the IsoScore for a point cloud X is approximately the fraction of dimensions uniformly utilized by X. And in Appendix D.2 we argue that IsoScore is an honest indicator of uniform spatial utilization.

D.1 IsoScore Reflects the Fraction of Dimensions Uniformly Utilized

In Section A we provided an interpretation for the value of the IsoScore $\iota(X)$ in Heuristic A.1. Intuitively, our heuristic says that $\iota(X)$ is roughly the fraction of dimensions of $\mathbb{R}^n$ utilized by X. We will now explain and justify this heuristic in detail. We formalize our heuristic below.

Heuristic D.1. *Suppose that a point cloud $X \subseteq \mathbb{R}^n$ gives an IsoScore $\iota(X)$. Then X occupies approximately*

$$k(X) := \iota(X) \cdot n + 1 - \iota(X) \tag{D.1}$$

dimensions of $\mathbb{R}^n$.

Note in particular that $\iota(X) = 0$ implies that (D.1) simplifies to a single dimension utilized and $\iota(X) = 1$ implies that (D.1) simplifies to all n dimensions utilized.

In the remainder of this subsection, we will justify the above heuristic. We will make reference to the notations and equations in Section 4. Fix $n \geq 1$ and $k \in \{1, \ldots, n\}$, and consider the matrix $I_n^{(k)}$. Recall that $I_n^{(k)}$ is the covariance matrix for a k-dimensional uncorrelated Gaussian distribution in $\mathbb{R}^n$. For example, spaces sampled using the matrices $I_3^{(k)}$, for $k = 1, 2, 3$ are rendered in Figure 1 as a line, a circle, and a ball, respectively. One can compute directly that the IsoScores for these three spaces are

$$\iota(I_3^{(1)}) \approx 0.0, \qquad \iota(I_3^{(2)}) \approx 0.5, \qquad \iota(I_3^{(3)}) \approx 1.0.$$

Our main insight in this section is that it is worthwhile to apply these statistics for reverse reasoning in the following sense: suppose you have some point cloud $X \subseteq \mathbb{R}^3$ which satisfies $\iota(X) \approx 1/2$. Then this IsoScore should allow you to infer that X uniformly occupies approximately 2 dimensions of $\mathbb{R}^3$.

In Heuristic D.1, we provide the closed formula (D.1) for generalizing the above reasoning to all dimensions n. We will now prove this formula.

Proof of Heuristic D.1. Once we normalize $I_n^{(k)}$ so that its Euclidean norm is $\sqrt{n}$, we get that the first k diagonal entries are $\sqrt{n/k}$. Therefore, the isotropy defect is

$$\begin{aligned} \delta(I_n^{(k)}) &= \frac{\|\hat{I}_n^{(k)} - \mathbf{1}\|}{\sqrt{2(n - \sqrt{n})}} && \text{(D.2)} \\ &= \frac{\sqrt{k(1 - \sqrt{n/k})^2 + n - k}}{\sqrt{2(n - \sqrt{n})}} && \text{(D.3)} \\ &= \frac{\sqrt{n - \sqrt{nk}}}{\sqrt{n - \sqrt{n}}}. \end{aligned}$$

It is natural to consider the map $k \mapsto \delta(I_n^{(k)})$. A priori, this is a discrete function defined on $\{1, \ldots, n\}$; a fortiori, this is in fact a continuous, monotonically decreasing bijection on the connected interval $[1, n]$. Therefore, the function defined by

$$\tilde{\delta}_n : [1, n] \to [0, 1] : k \mapsto \delta(I_n^{(k)})$$

is invertible, and one can compute that its inverse is

$$\tilde{\delta}_n^{-1} : [0, 1] \to [1, n] : d \mapsto \frac{(n - d^2(n - \sqrt{n}))^2}{n}.$$

The truth of this heuristic rests upon the validity of the following assumption, which is reasonable to use in many contexts.

Assumption Underpinning The Heuristic. *The isotropy defect corresponding to a point cloud sampled using the covariance matrix $I_n^{(k)}$ is the prototypical isotropy defect for any point cloud in $\mathbb{R}^n$ which uniformly utilizes k dimensions.*

We will now invoke this assumption. Let $\delta(X)$ be the isotropy defect for an arbitrary point cloud X. If we assume that we are in the nontrivial case where $\delta(X) > 0$, then $\tilde{\delta}_n^{-1}(\delta(X))$ is in the interval $[1, n)$. Because $\tilde{\delta}_n^{-1}$ is bijective, there exists

a unique $k \in \{1, \ldots, n-1\}$ with the property that $\tilde{\delta}_n^{-1}(\delta(X)) \in [k, k+1)$. But by construction, $[k, k+1) = [\tilde{\delta}_n^{-1}(\delta(I_n^{(k)})), \tilde{\delta}_n^{-1}(\delta(I_n^{(k+1)})))$. By monotonicity of $\tilde{\delta}_n^{-1}$, this implies that

$$\delta(X) \in [\delta(I_n^{(k)}), \delta(I_n^{(k+1)})).$$

Therefore, by the assumption underpinning the heuristic, we can deduce that X is uniformly utilizing between k and $k+1$ dimensions of $\mathbb{R}^n$. To be specific, we say that X is uniformly utilizing $\tilde{\delta}_n^{-1}(\delta(X)) \in [k, k+1)$ dimensions. Recalling Section 4, we can recognize that in Step 6, the formula for $k(X)$, the quantity of dimensions uniformly utilized by X, is precisely $k(X) := \tilde{\delta}_n^{-1}(\delta(X))$; likewise, the formula for $\phi(X)$, the fraction of dimensions uniformly utilized by X, is $\phi(X) := \tilde{\delta}_n^{-1}(\delta(X))/n$.

Now we are in a position to verify Equation D.1, the main claim of Heuristic D.1. By the assumption underpinning the heuristic, it is sufficient to verify Equation D.1 in the case of $I_n^{(k)}$, for $k = 1, \ldots, n$. This is because all functions that we will utilize are monotonic bijections. Using the notation in Steps 6 and 7 in Section 4, we can compute that

$$\begin{aligned} \iota(I_n^{(k)})(n-1)+1 &= S(\phi_n(I_n^{(k)}))(n-1)+1 \\ &= n \cdot \phi_n(I_n^{(k)}) \\ &= k(I_n^{(k)}). \end{aligned}$$

Using the formula $k(X) = (n - \delta(X)^2(n - \sqrt{n}))^2/n$, we can continue:

$$\begin{aligned} k(I_n^{(k)}) &= \frac{(n - \delta(I_n^{(k)})^2(n-\sqrt{n}))^2}{n} \\ &= \frac{(n - \frac{n-\sqrt{nk}}{n-\sqrt{n}}(n-\sqrt{n}))^2}{n} \\ &= k, \end{aligned}$$

where in the penultimate equality we used Equation D.2. This completes the proof. □

Because IsoScore covers a continuous spectrum, one should carefully interpret what we mean when we say that X occupies approximately k dimensions of $\mathbb{R}^n$. For example, consider the 2D Gaussian distributions depicted in Figure 3. Heuristic D.1 predicts $k = 1.9996, 1.6105, 1.0281$ dimensions are used when $x = 1, 3, 75$, respectively. These should be interpreted as follows: "when $x = 75$, the points sampled are mostly using one direction of space" and "when $x = 3$, the points sampled are using somewhere between one and two dimensions of space."

Heuristic D.1 suggests that an IsoScore near $1/2$ means that the corresponding point cloud X occupies approximately half of the dimensions of its ambient space. We can make this reasoning rigorous as follows: for any $n \geq 2$, one can compute that

$$\iota(I_n^{(k)}) = \frac{k-1}{n-1} \approx \frac{k}{n}, \qquad \text{for any } k = 1, \ldots, n. \tag{D.4}$$

Proof of (D.4). In Equation D.2 computed that the isotropy defect is $\delta(I_n^{(k)}) = \sqrt{n - \sqrt{nk}}/\sqrt{n - \sqrt{n}}$. If we substitute this expression into (4.1), then we obtain the formula $\iota(I_n^{(k)}) = \frac{k-1}{n-1}$. Furthermore, one can easily estimate that $|\frac{k-1}{n-1} - \frac{k}{n}| \leq \frac{1}{n}$. □

Table 2 illustrates this formula in the concrete case of $\mathbb{R}^9$. This formula implies the following key relationship:

$$\lim_{n\to\infty} \iota(I_n^{(\lfloor n/2 \rfloor)}) = 1/2.$$

Generalizing this line of reasoning yields our second heuristic explanation for the meaning of IsoScore, Heuristic 4.1. We copy it here:

Heuristic D.2. *When the ambient space $\mathbb{R}^n$ has large dimension, the IsoScore $\iota(X)$ is approximately the fraction of dimensions uniformly utilized by X.*

Proof of Heuristic 4.1. By the assumption underpinning Heuristic D.1, it suffices to show this in the case of matrices of the form $I_n^{(k)}$. Fix $\alpha \in [0, 1]$, and consider the covariance matrix $I_n^{(\lfloor \alpha n \rfloor)}$. For large n, the fraction of dimensions uniformly utilized by $I_n^{(\lfloor \alpha n \rfloor)}$ is approximately α, according to Definition 3.1. But by (D.4), we can compute that

$$\lim_{n\to\infty} \iota(I_n^{(\lfloor \alpha n \rfloor)}) = \lim_{n\to\infty} \frac{\lfloor \alpha n \rfloor - 1}{n-1} = \alpha.$$

This completes the proof. □

D.2 The IsoScore for $I_n^{(k)}$ Reflects Uniform Utilization of k Dimensions

We will now investigate what range of IsoScores are obtained by sample covariance matrices that utilize k out of n dimensions. It is easy to see

that these scores at least fill the interval $(0, \iota(I_n^{(k)})]$, since the map

$$[1,\infty) \to (0, \iota(I_n^{(k)})]$$
$$x \mapsto \iota(\text{diag}(x, 1, \ldots, 1, 0, \ldots, 0))$$

is surjective. Conversely, we can show that this interval is the only possible range of IsoScores corresponing to such covariance matrices. We make this claim rigorous in the following proposition.

Proposition D.3. *Fix $n \geq 2$. For any $k = 1, \ldots, n$, we have that*

$$I_n^{(k)} = \text{argmax}\{\iota(J) : J \text{ utilizes } k \text{ out of } n \text{ dimensions}\}. \quad \text{(D.5)}$$

This result justifies the use of IsoScore for measuring the extent to which a point cloud optimally utilizes all dimensions of the ambient space because it demonstrates that $\iota(I_n^{(k)})$ is the maximal IsoScore for any covariance matrix with k non-zero entries and $n - k$ zero entries.

Proof of Proposition D.3. In this section we let $\text{Diag}^+(n)$ denote the set of $n \times n$ real matrices which vanish away from the diagonal and whose diagonal entries are all non-negative. The set $\text{Diag}^+(n)$ parameterizes the set of all $n \times n$ sample covariance matrices after performing PCA-reorientation. We also let $\text{Diag}^+(n,k) \subseteq \text{Diag}^+(n)$ denote that subset whose first k diagonal entries are non zero and whose last $n-k$ diagonal entries are zero. The set $\text{Diag}^+(n,k)$ parameterizes the set of sample covariance matrices post-PCA reorientation which utilize k out of n dimensions of space. Covariance matrices in $\text{Diag}^+(n,k)$ represent point clouds with the property that $\text{Var}(x_i) > 0$ for $i = 1, \ldots, k$, and $\text{Var}(x_i) = 0$ for $i = k+1, \ldots, n$.

It suffices to show that, for every $J \in \text{Diag}^+(n,k)$, we have that $\iota(J) \leq \iota(I_n^{(k)})$, or equivalently, $\delta(J) \geq \delta(I_n^{(k)})$. Write $\hat{I}_{n,D}^{(k)} = (\sqrt{n/k}, \ldots, \sqrt{n/k}, 0, \ldots, 0)$ and $J_D = (a_1, \ldots, a_k, 0, \ldots, 0)$, where $a_1^2 + \cdots a_k^2 = n$. Then we must show that $\|J_D - \mathbf{1}\| \geq \|\hat{I}_{n,D}^{(k)} - \mathbf{1}\|$, or equivalently,

$$\sum_{i=1}^{k}(a_i - 1)^2 + n - k \geq \sum_{i=1}^{k}(\sqrt{n/k} - 1)^2 + n - k.$$

This latter estimate is equivalent to

$$\sum_{i=1}^{k} a_i \leq \sqrt{nk}.$$

By Jensen's inequality, applied with the convex function $f(x) = x^2$, we have that

$$f\left(\sum_{i=1}^{k} \frac{a_i}{k}\right) \leq \sum_{i=1}^{k} \frac{f(a_i)}{k}.$$

Simplifying, this implies that $(a_1 + \cdots + a_k)^2 \leq kn$. This completes the proof. □

E Numerical Experiments

In this section, we provide explicit details of how each test is designed. We provide code for all experiments at: *https://github.com/bcbi-edu/p_eickhoff_isoscore*.

1. **Test 1: Mean Invariance.** To assess whether the five scores are mean invariant, we start with $100,000$ points sampled from a 10-dimensional multivariate Gaussian distribution with covariance matrix equal to the identity and a common mean vector $M = [\mu, \mu, ..., \mu]$. We compute scores for $\mu = 0, 1, 2, ..., 20$.

2. **Test 2: Scalar Invariance.** We test for the property of scalar invariance by sampling $100,000$ points from a 5D Gaussian distribution with common mean vector $M = [3, 3, 3, 3, 3]$ and covariance matrix equal to $\lambda \cdot I_5$. We then compute scores for each point cloud as we increase λ from 1 to 25.

3. **Test 3: Maximum Variance.** We start by sampling $100,000$ points from a 10D multivariate Gaussian distribution with zero common mean vector and a diagonal covariance matrix with nine entries equal to 1 and one diagonal entry equal to x. In our experimental setup, we compute all five scores as we increase x from 1 to 75.

4. **Test 4: Rotation Invariance.** Our baseline point cloud $X \subset \mathbb{R}^n$ consists of $100,000$ points sampled from a 2D zero-mean Gaussian distribution with a covariance matrix equal to $\left(\begin{smallmatrix} 1 & 0.8 \\ 0.8 & 1 \end{smallmatrix}\right)$. We rotate X by $120°$ and $240°$. Lastly, we project X using PCA reorientation while retaining dimensionality to obtain a point cloud X^{PCA}.

5. **Test 5: Dimensions Used (Fraction of Dimensions Used Test).** For our first experiment, which we term the "fraction of dimensions used test," we sample $100,000$ points

from a 25D multivariate Gaussian distribution with a zero common mean vector and a diagonal covariance matrix where the first k entries are 1 and the remaining $n-k$ diagonal elements are 0. We refer to k as the number of dimensions uniformly used by our data (see Definition 3.1). For our experiment we let $k = 1, 2, 3, ..., 25$, and compute the corresponding scores.

6. **Test 5: Dimensions Used (High Dimensional Test).** A good score of spatial utilization should allow for easy comparison between different vector spaces even when the dimensionality of the two spaces is different. We sample $100,000$ points from a zero-mean Gaussian distribution with identity covariance matrix I_n and increase the dimension of the distribution from $n = 2, \ldots, 100$.

7. **Test 6: Global Stability.** We generate a "skewered meatball" by sampling $1,000$ points from a line in 3D space and increase the number of points sampled from a 3-Dimensional, zero-mean, isotropic Gaussian from 0 to $150,000$.

F Geometry of Isotropy

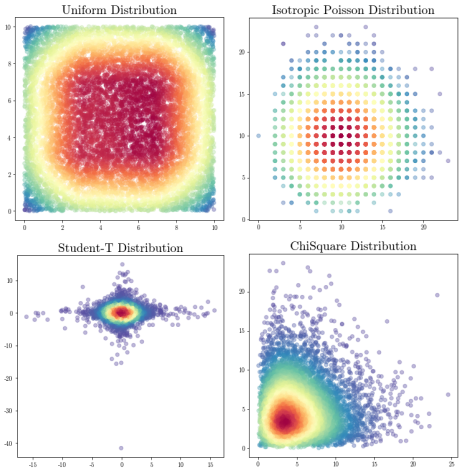

Figure 10: Points sampled from a Uniform distribution, Poisson distribution, Student-T distribution and ChiSquare distribution respectively

Each of the distributions illustrated in Figure 10 has a covariance matrix proportional to the identity and is therefore maximally isotropic. Namely, the variance is distributed equally in all directions. Despite receiving an IsoScore of 1, the geometry of the point clouds are vastly different. We can only comment on the geometry of the point cloud if the underlying distribution of the space is known.