Multilingual Representation Learning

601.764

4/6/2023

Isotropy

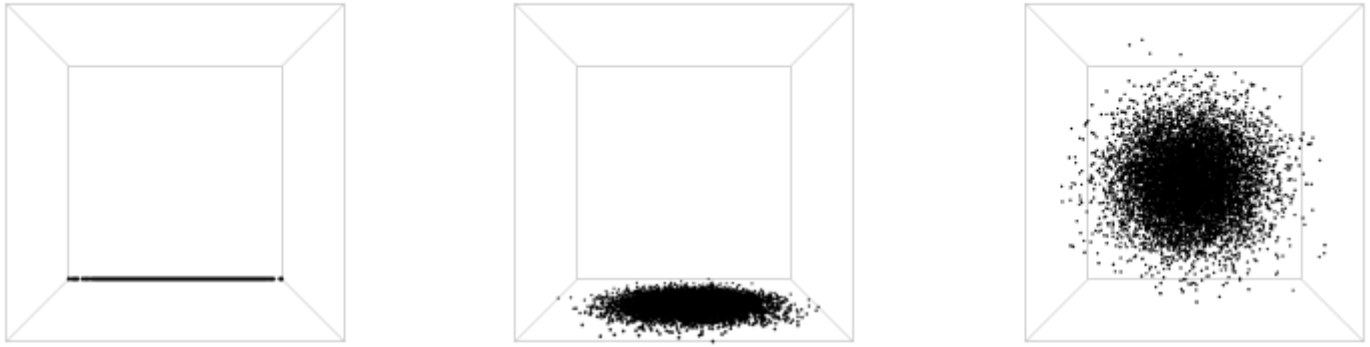


Figure 1: From left to right, a line, disk, and ball embedded in 3D space.

Rudman et al., 2022

Isotropic Distribution

- Variance is uniformly distributed across all dimensions
- Fully isotropic when the covariance matrix is proportional to the identity matrix

2016

A Latent Variable Model Approach to PMI-based Word Embeddings

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, Andrej Risteski
Computer Science Department, Princeton University
35 Olden St, Princeton, NJ 08540
{arora,yuanzhil,yingyul,tengyu,risteski}@cs.princeton.edu

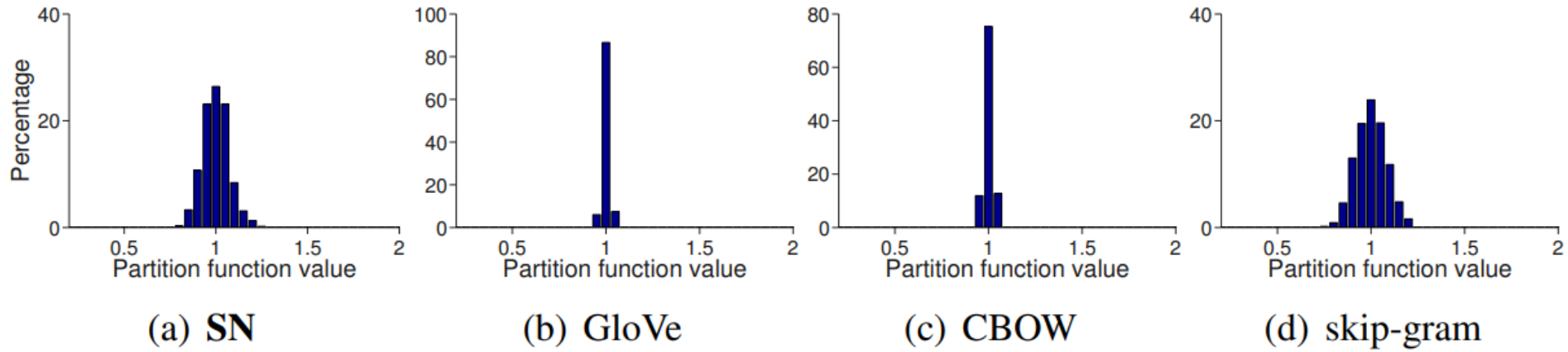
Arora et al., 2017

- “This provides a theoretical justification for nonlinear models like PMI, word2vec, and GloVe, as well as some hyperparameter choices. It also helps explain why low-dimensional semantic embeddings contain linear algebraic structure that allows solution of word analogies, as shown by Mikolov et al. (2013a) and many subsequent papers.”
- “Experimental support is provided for the generative model assumptions, the most important of which is that latent word vectors are fairly uniformly dispersed in space.”

Arora et al., 2017

- “The following lemma (whose proof appears in the appendix) is central to the analysis. It says that under the Bayesian prior, the partition … which is the implied normalization in equation (2.1), is close to some constant Z for most of the discourses c. This can be seen as a plausible theoretical explanation of a phenomenon called self-normalization in log-linear models: ignoring the partition function or treating it as a constant (which greatly simplifies training) is known to often give good results. This has also been studied in (Andreas and Klein, 2014)”

Arora et al., 2017



(a) **SN** (b) GloVe (c) CBOW (d) skip-gram

Figure 1: The partition function Z_c. The figure shows the histogram of Z_c for 1000 random vectors c of appropriate norm, as defined in the text. The x-axis is normalized by the mean of the values. The values Z_c for different c concentrate around the mean, mostly in $[0.9, 1.1]$. This concentration phenomenon is predicted by our analysis.

Arora et al., 2017

- The spatial isotropy of word vectors is both an assumption in our model, and also a new empirical finding of our paper.

2017

The strange geometry of skip-gram with negative sampling

David Mimno and **Laure Thompson**
Cornell University
mimno@cornell.edu, laurejt@cs.cornell.edu

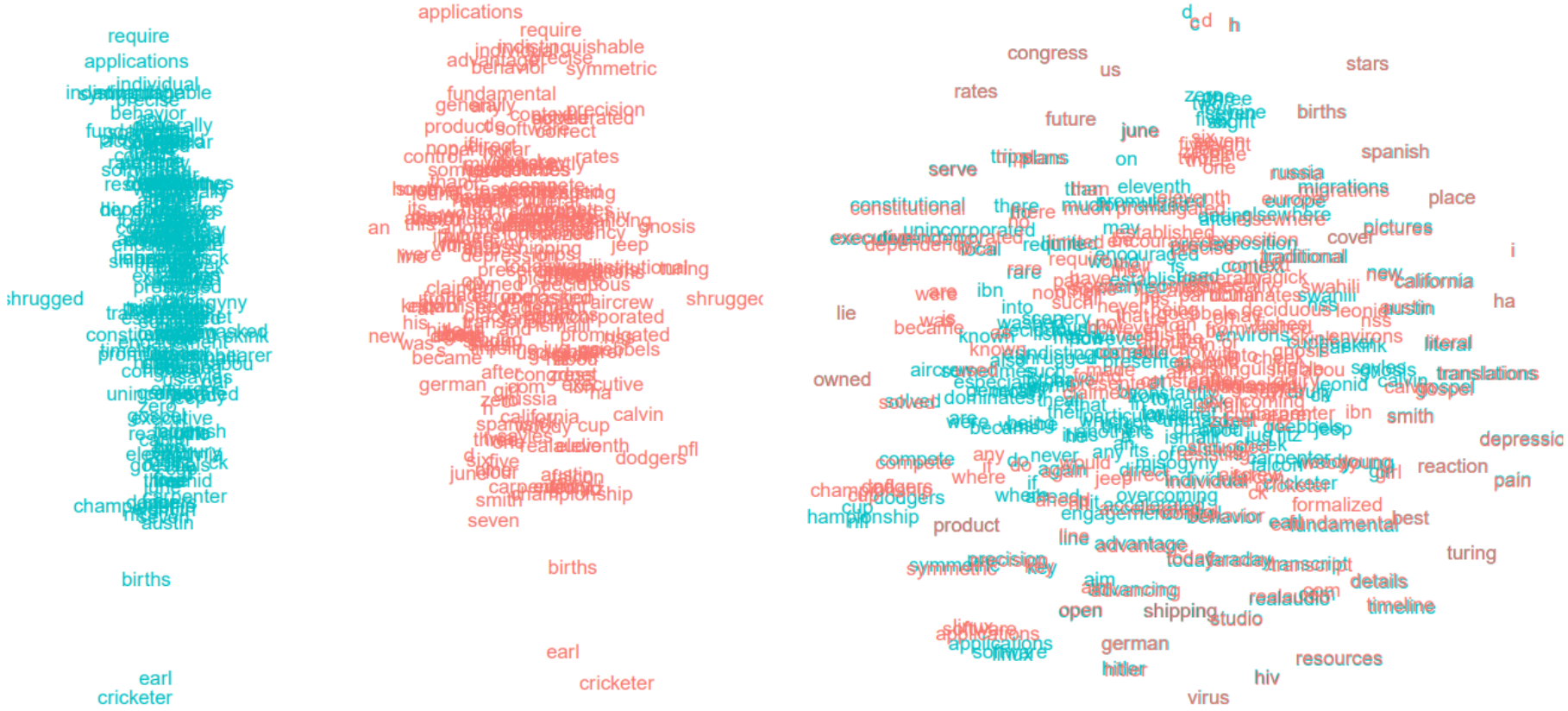


Figure 1: SGNS word vectors and their context vectors projected using PCA (left) and t-SNE (right). t-SNE provides a more readable layout, but masks the divergence between word and context vectors.

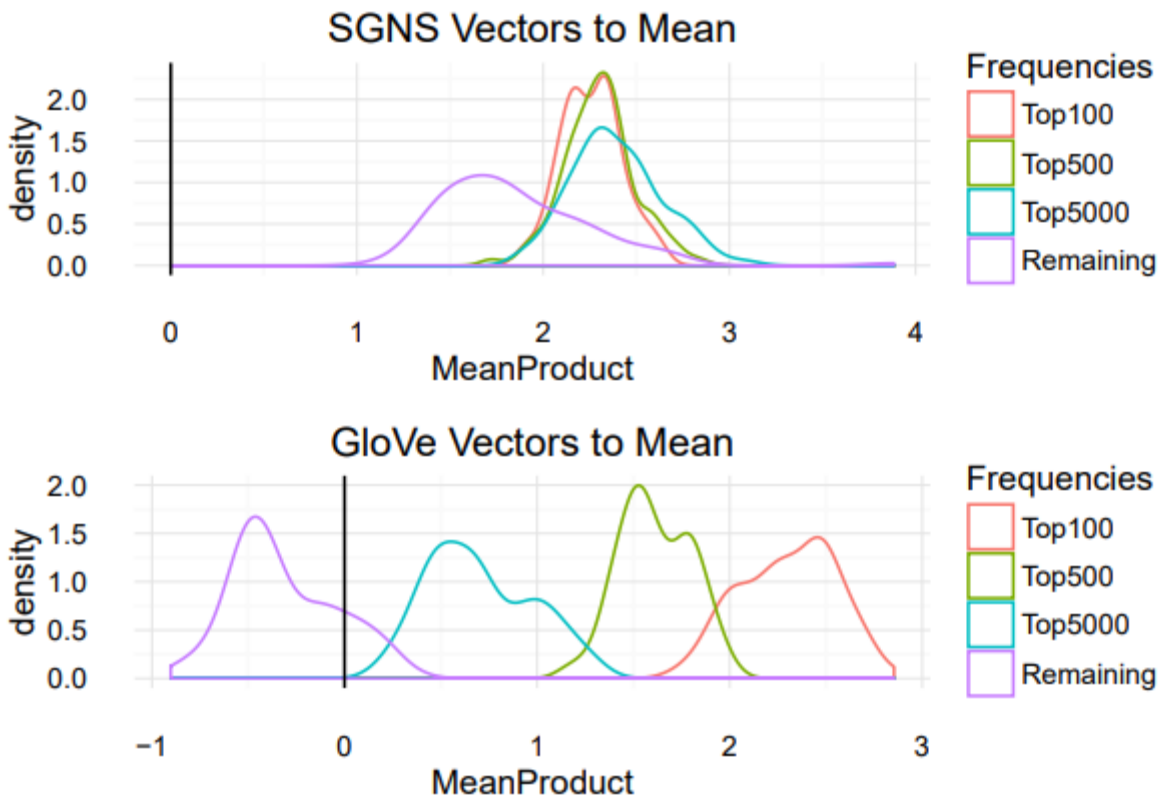


Figure 2: SGNS-trained vectors mostly point in the same direction, towards a mean vector $\hat{w}$.

Mimno and Thompson 2017

- “SGNS vectors are arranged along a primary axis. Our first observation is that SGNS-trained vectors all point in roughly the same direction. We can define a mean vector $\bar{w}$ by averaging the vectors of the complete vocabulary w. We sample a balanced set of 400 total words with 100 each from the four frequency categories. Figure 2 shows the distribution of inner products between these 400 sampled words and their mean vector $\hat{w}$. All vectors have a large, positive inner product with the mean, indicating that they are not evenly dispersed through the space.”

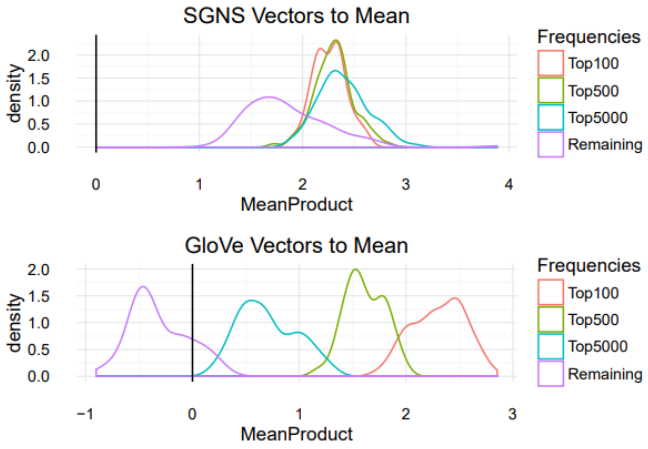


Figure 2: SGNS-trained vectors mostly point in the same direction, towards a mean vector $\hat{w}$.

2019

How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings

Kawin Ethayarajh*
Stanford University
kawin@stanford.edu

Ethayarajh 2019

- “In all layers of all three models, the contextualized word representations of all words are not isotropic: they are not uniformly distributed with respect to direction. Instead, they are anisotropic, occupying a narrow cone in the vector space. The anisotropy in GPT-2’s last layer is so extreme that two random words will on average have almost perfect cosine similarity! Given that isotropy has both theoretical and empirical benefits for static embeddings (Mu et al., 2018), the extent of anisotropy in contextualized representations is surprising”

Ethayarajh 2019

- “Occurrences of the same word in different contexts have non-identical vector representations. Where vector similarity is defined as cosine similarity, these representations are more dissimilar to each other in upper layers. This suggests that, much like how upper layers of LSTMs produce more task-specific representations (Liu et al., 2019a), upper layers of contextualizing models produce more context-specific representations”

Ethayarajh 2019

- “Context-specificity manifests very differently in ELMo, BERT, and GPT-2. In ELMo, representations of words in the same sentence grow more similar to each other as context-specificity increases in upper layers; in BERT, they become more dissimilar to each other in upper layers but are still more similar than randomly sampled words are on average; in GPT-2, however, words in the same sentence are no more similar to each other than two randomly chosen words.”

Ethayarajh 2019

- “After adjusting for the effect of anisotropy, on average, less than 5% of the variance in a word’s contextualized representations can be explained by their first principal component. This holds across all layers of all models. This suggests that contextualized representations do not correspond to a finite number of word-sense representations, and even in the best possible scenario, static embeddings would be a poor replacement for contextualized ones. Still, static embeddings created by taking the first principal component of a word’s contextualized representations outperform GloVe and FastText embeddings on many word vector benchmarks.”

Ethayarajh 2019

- “These insights help justify why the use of contextualized representations has led to such significant improvements on many NLP tasks.”

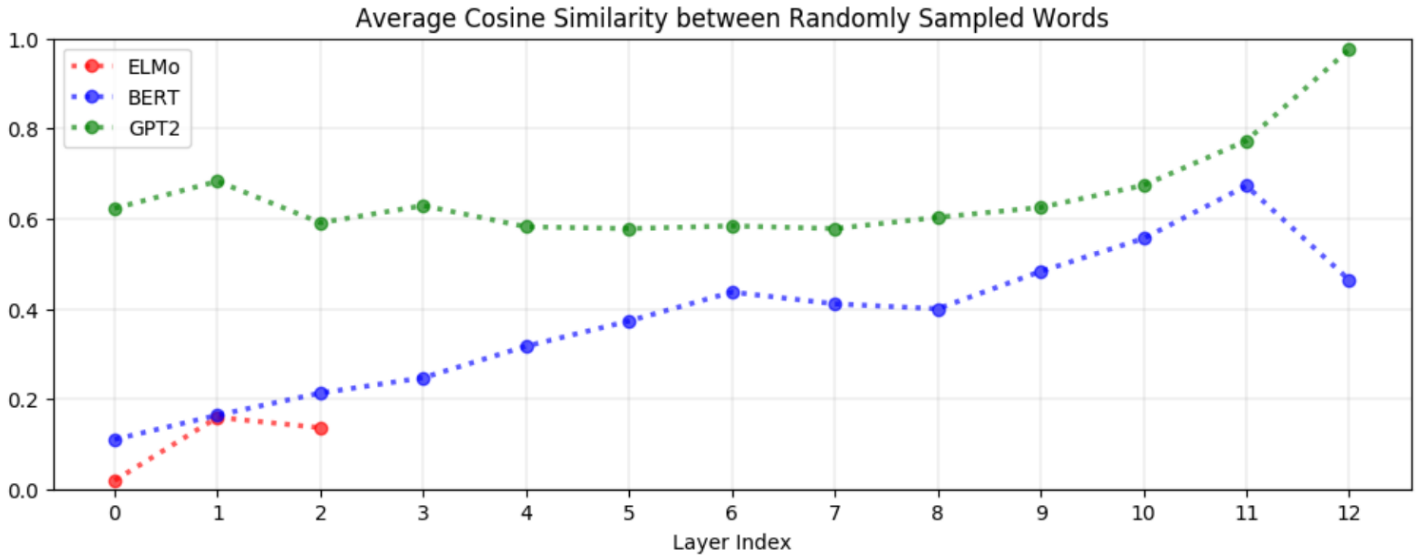


Figure 1: In almost all layers of BERT, ELMo, and GPT-2, the word representations are anisotropic (i.e., not directionally uniform): the average cosine similarity between uniformly randomly sampled words is non-zero. The one exception is ELMo's input layer; this is not surprising given that it generates character-level embeddings without using context. Representations in higher layers are generally more anisotropic than those in lower ones.

Contextualized representations are generally more anisotropic in higher layers. As seen in Figure 1, for GPT-2, the average cosine similarity between uniformly randomly words is roughly 0.6 in layers 2 through 8 but increases exponentially from layers 8 through 12. In fact, word representations in GPT-2's last layer are so anisotropic that any two words have on average an almost perfect cosine similarity! This pattern holds for BERT and

4.1 (An)Isotropy

Contextualized representations are anisotropic in all non-input layers. If word representations from a particular layer were isotropic (i.e., directionally uniform), then the average cosine similarity between uniformly randomly sampled words would be 0 (Arora et al., 2017). The closer this average is to 1, the more anisotropic the representations. The geometric interpretation of anisotropy is that the word representations all occupy a narrow cone in the vector space rather than being uniform in all directions; the greater the anisotropy, the narrower this cone (Mimno and Thompson, 2017). As seen in Figure 1, this implies that in almost all layers of BERT, ELMo and GPT-2, the representations of all words occupy a narrow cone in the vector space. The only exception is ELMo's input layer, which produces static character-level embeddings without using contextual or even positional information (Peters et al., 2018). It should be noted that not all static embeddings are necessarily isotropic, however; Mimno and Thompson (2017) found that skipgram embeddings, which are also static, are not isotropic.

Published as a conference paper at ICLR 2018

ALL-BUT-THE-TOP: SIMPLE AND EFFECTIVE POST-PROCESSING FOR WORD REPRESENTATIONS

Jiaqi Mu, Pramod Viswanath
University of Illinois at Urbana Champaign
{jiaqimu2, pramodv}@illinois.edu

The idea of isotropy comes from the partition function defined in (Arora et al., 2016),

$$Z(c) = \sum_{w \in \mathcal{V}} \exp\left(c^\top v(w)\right),$$

where $Z(c)$ should approximately be a constant with any unit vector c (c.f. Lemma 2.1 in (Arora et al., 2016)). Hence, we mathematically define a measure of isotropy as follows,

$$I(\{v(w)\}) = \frac{\min_{\|c\|=1} Z(c)}{\max_{\|c\|=1} Z(c)}, \tag{1}$$

where $I(\{v(w)\})$ ranges from 0 to 1, and $I(\{v(w)\})$ closer to 1 indicates that $\{v(w)\}$ is more isotropic. The intuition behind our postprocessing algorithm can also be motivated by letting $I(\{v(w)\}) \to 1$.

All-But-The-Top

Observation *Every* representation we tested, in many languages, has the following properties:

- The word representations have *non-zero mean* – indeed, word vectors share a large common vector (with norm up to a half of the average norm of word vector).
- After removing the common mean vector, the representations are *far from* isotropic – indeed, much of the energy of most word vectors is contained in a very low dimensional subspace (say, 8 dimensions out of 300).

Significance of Nulled Vectors Consider the representation of the words as viewed in terms of the top D PCA coefficients $\alpha_\ell(w)$, for $1 \leq \ell \leq D$. We find that these few coefficients encode the *frequency* of the word to a significant degree; Figure 2 illustrates the relation between the $(\alpha_1(w), \alpha_2(w))$ and the unigram probabilty $p(w)$, where the correlation is geometrically visible.

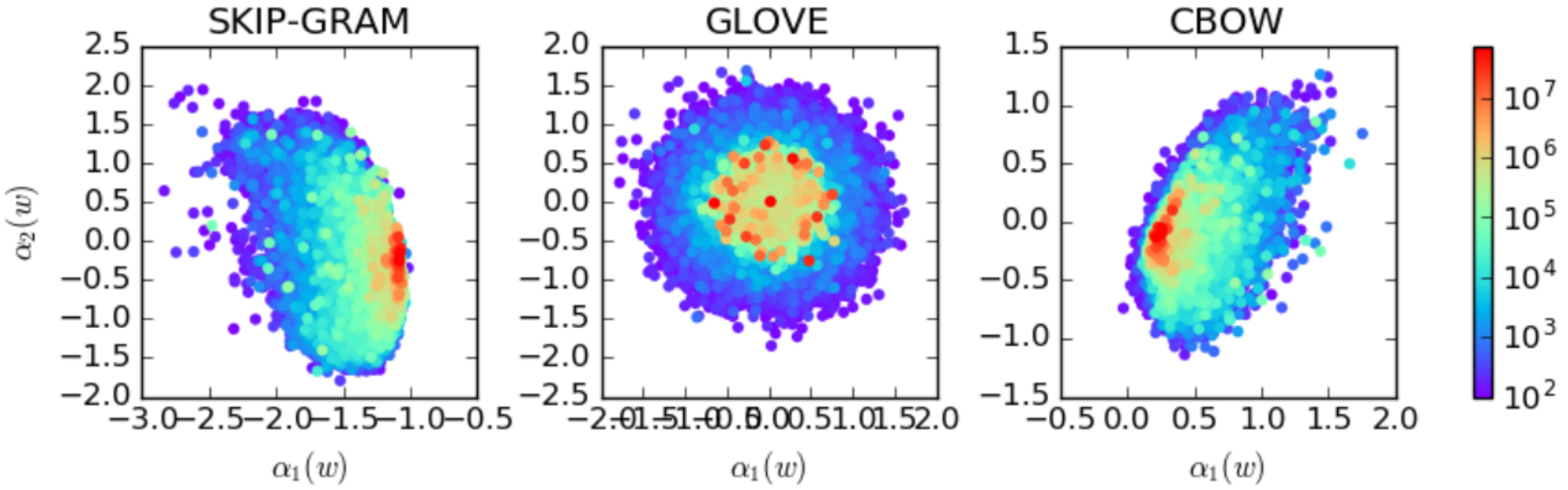


Figure 2: The top two PCA directions (i.e, $\alpha_1(w)$ and $\alpha_2(w)$) encode frequency.

5 CONCLUSION

We present a simple postprocessing operation that renders word representations even stronger, by eliminating the top principal components of all words. Such an simple operation could be used for word embeddings in downstream tasks or as intializations for training task-specific embeddings. Due to their popularity, we have used the published representations of WORD2VEC and GLOVE in English in the main text of this paper; postprocessing continues to be successful for other representations and in multilingual settings – the detailed empirical results are tabulated in Appendix C.

2021

An Isotropy Analysis in the Multilingual BERT Embedding Space

Sara Rajaee★ and **Mohammad Taher Pilehvar**☆
★Iran University of Science and Technology, Tehran, Iran
☆Tehran Institute for Advanced Studies, Khatam University, Iran
sara_rajaee@comp.iust.ac.ir
mp792@cam.ac.uk

Cosine Similarity. Ethayarajh (2019) used cosine similarity between random embeddings as an approximation of isotropy in the space. As mentioned before, random embeddings with an isotropic distribution have near-zero cosine similarities. The metric can be formulated as follows:

$$I_{Cos}(\mathcal{W}) = \frac{1}{N} \sum_{i=1, x_i \neq y_i}^{N} Cos(x_i, y_i) \quad (1)$$

where $x_i \in X, y_i \in Y$, X and Y are the sets of randomly sampled embeddings, and $\mathcal{W}$ is the embedding matrix. N is the number of sampled pairs that is set to 1000 in our experiments. Lower $I_{Cos}(\mathcal{W})$ values indicate higher isotropy.

Principal Components. Mu and Viswanath (2018) proposed a metric based on principal components (PCs), approximated as follows:

$$I_{PC}(\mathcal{W}) \approx \frac{\min_{u \in U} F(u)}{\max_{u \in U} F(u)}, F(u) = \sum_{i=1}^{M} \exp(u^T w_i) \quad (2)$$

where w_i is the i^{th} word embedding, M is the number of all representations in the space, U is the set of eigenvectors of the embedding matrix, and $F(u)$ is the partition function described in Equation 2. Arora et al. (2016) proved that $F(u)$ could be approximated using a constant for isotropic embedding spaces. Therefore, $I_{PC}(\mathcal{W})$ would be close to one in an isotropic embedding space.

	$I_{Cos}(\mathcal{W})$	**First**	**Second**	**Third**
BERT	0.34	0.385	0.005	0.005
English	0.24	0.041	0.029	0.020
Spanish	0.27	0.033	0.029	0.018
Arabic	0.27	0.033	0.025	0.022
Turkish	0.25	0.036	0.024	0.024
Sundanese	0.25	0.036	0.016	0.016
Swahili	0.27	0.025	0.018	0.014

Table 2: The contribution of top-three dimensions to the expected cosine similarity ($I_{Cos}(\mathcal{W})$) in BERT and mBERT models.

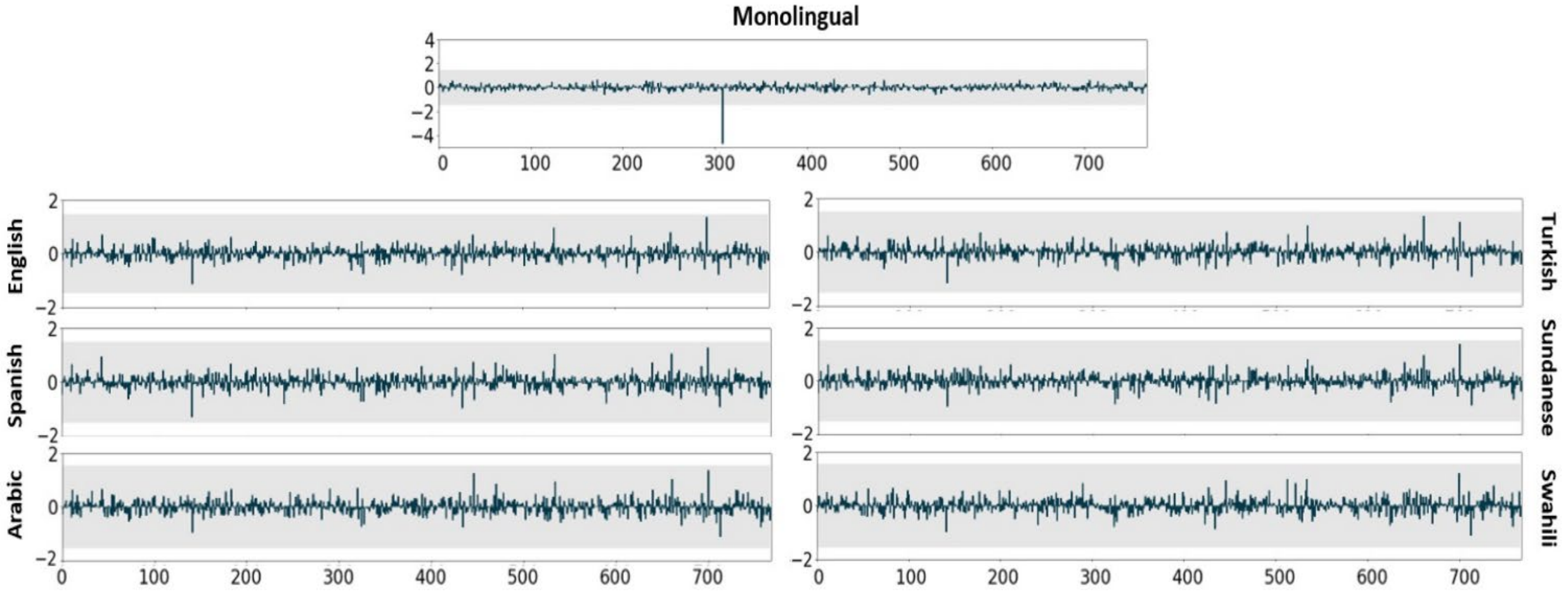


Figure 2: The average representation in English BERT (top) and mBERT (bottom). The shaded area denotes 3σ. While an outlier has emerged in the former, we do not see any major outliers in the multilingual space.

2021

Too Much in Common: Shifting of Embeddings in Transformer Language Models and its Implications

Daniel Biś
Florida State University
Tallahassee, USA
bis@cs.fsu.edu

Maksim Podkorytov
Florida State University
Tallahassee, USA
maksim@cs.fsu.edu

Xiuwen Liu
Florida State University
Tallahassee, USA
liux@cs.fsu.edu

Abstract

The success of language models based on the Transformer architecture appears to be inconsistent with observed anisotropic properties of representations learned by such models. We resolve this by showing, contrary to previous studies, that the representations do not occupy a narrow cone, but rather drift in common directions. At any training step, all of the embeddings except for the ground-truth target embedding are updated with gradient in the same direction. Compounded over the training set, the embeddings drift and share common components, manifested in their shape in all the models we have empirically tested. Our experiments show that isotropy can be restored using a simple transformation.[1]

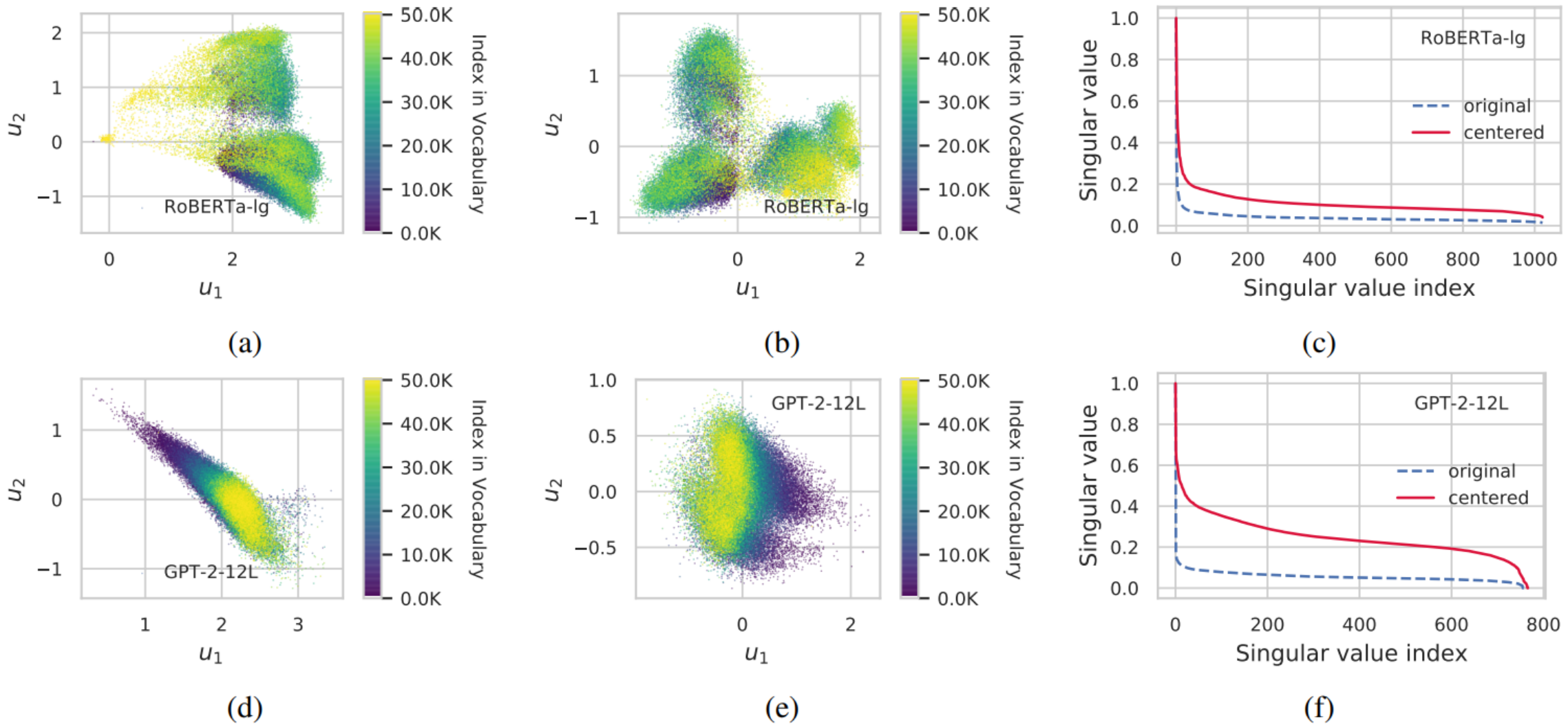


Figure 1: Top: RoBERTa-large. Bottom: GPT-2 (12 layers). (1a, 1d): Word embeddings projected onto first two singular vectors. (1b, 1e) Centered word embeddings projected onto first two singular vectors. (1c, 1f) Singular values of embedding matrix before and after centering. Centering the embedding matrix increases isotropy of embeddings.

2021

Isotropy in the Contextual Embedding Space: Clusters and Manifolds

Xingyu Cai, Jiaji Huang, Yuchen Bian, Kenneth Church
Baidu Research, 1195 Bordeaux Dr, Sunnyvale, CA 94089, USA
{xingyucai,huangjiaji,yuchenbian,kennethchurch}@baidu.com

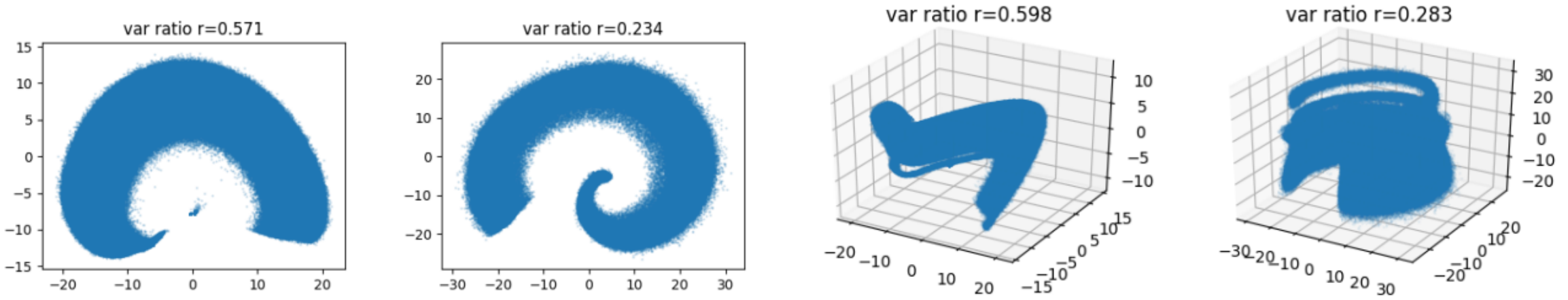
Cai et al., 2021

ABSTRACT

The geometric properties of contextual embedding spaces for deep language models such as BERT and ERNIE, have attracted considerable attention in recent years. Investigations on the contextual embeddings demonstrate a strong anisotropic space such that most of the vectors fall within a narrow cone, leading to high cosine similarities. It is surprising that these LMs are as successful as they are, given that most of their embedding vectors are as similar to one another as they are. In this paper, we argue that the isotropy indeed exists in the space, from a different but more constructive perspective. We identify isolated clusters and low dimensional manifolds in the contextual embedding space, and introduce tools to both qualitatively and quantitatively analyze them. We hope the study in this paper could provide insights towards a better understanding of the deep language models.

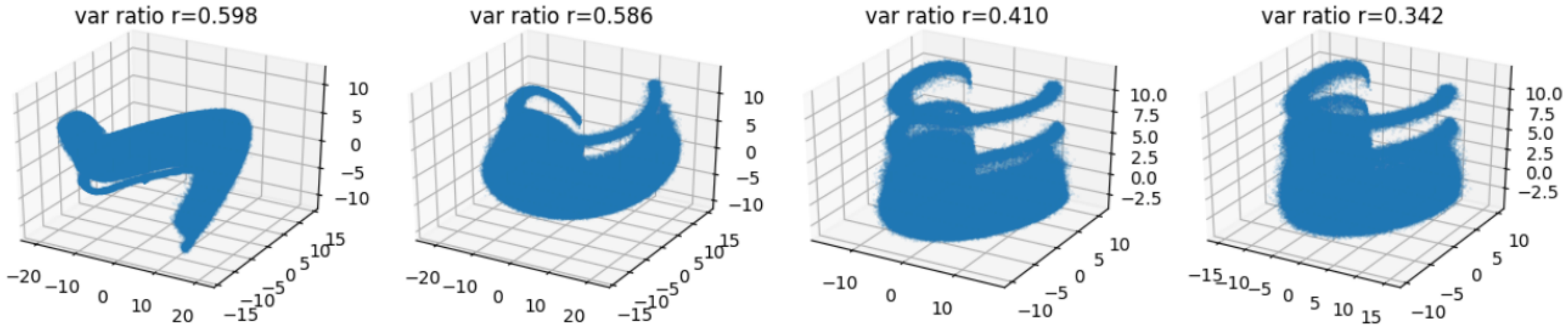
5 Conclusions and Future Work

Previous works have reported the strong anisotropy in deep LMs, which is hard to explain the superior performance achieved by these models. We suggest that the anisotropy is a global view, being largely misled by distinct clusters resided in the space. Our analysis show that it is more constructive to isolate and transform the space to measure the isotropy. From this view, within the clusters, the spaces of different models all have nearly perfect isotropy that could explain the large model capacity. In addition, we investigate the space geometry for different models. Our visualization demonstrates a low-dimensional Swiss Roll manifold for GPT and GPT2 embeddings, that has not been reported before. The tokens and word frequencies are presented to qualitatively show the manifold structure. We propose to use the approximate LID to quantitatively measure the local subspace, and compared with static embedding spaces. The results show smaller LID values for the contextual embedding models, which can be seen as a local anisotropy in the space. We hope this line of research could bring a comprehensive geometric view of contextual embedding space, and gain insights on how the embeddings are affected by attention, compression, multilingualism, etc. Therefore the model performance could be further improved based on the findings.



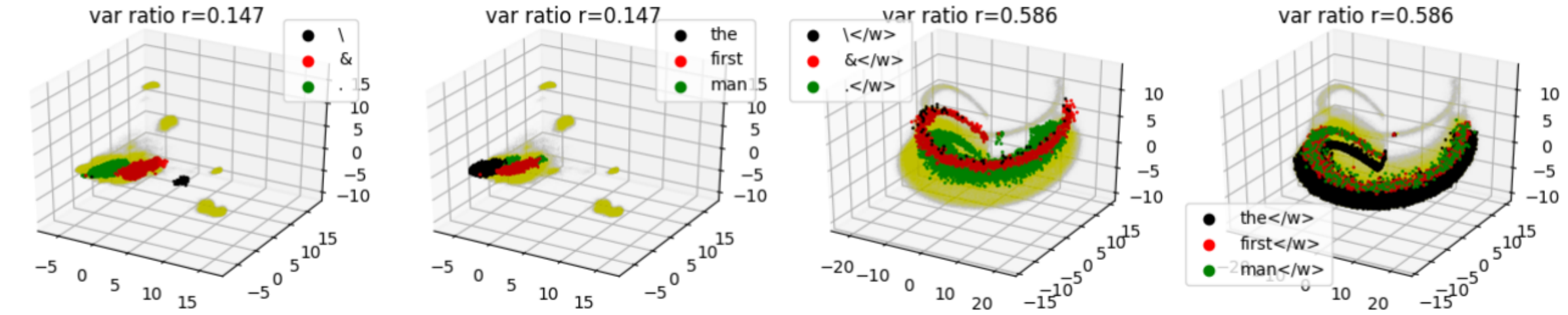
(a) GPT layer 2 (b) GPT2 layer 2 (c) GPT layer 2 3-D view (d) GPT2 layer 2 3-D view

Figure 5: The 2-D and 3-D view of low-dimensional manifold in GPT/GPT2's embedding spaces



(a) GPT layer 2 (b) GPT layer 3 (c) GPT layer 4 (d) GPT layer 5

Figure 6: The evolution from a narrow band into a taller and taller Swiss Roll with deeper layers.



(a) BERT, symbol tokens (b) BERT, word tokens (c) GPT, symbol tokens (d) GPT, word tokens

Figure 7: Embeddings for symbol tokens and word tokens, in layer 3 of BERT and GPT. This shows that GPT has manifold structure, such that vectors are along the spiral band. BERT's space is closer to a Euclidean space as similar vectors are in concentrated clusters.

"As shown in Figure 7a 7b, the BERT model indeed group similar embeddings into small regions in the space (the red, black and green clusters). However, the GPT models are assigning similar embeddings along the manifold we observed before. In Figure 7c 7d, the embeddings for the tokens occupy a spiral band that almost cross the entire space. It does not comply with the Euclidean space geometry as points in such a spiral band would not have high cosine similarity. A Riemannian metric must exist, such that the manifold has larger distance between two spiral bands, but smaller distance on the band. Note that the 3-D plots are obtained using PCA, so there is no density-based nor non-linear reduction involved. Therefore, the manifold structures in GPT embedding spaces are verified."

2022

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

Isotropy & Embedding Models

- Numerous claims that isotropy correlates with improved performance
- Laundry list of citations from past 5 years
- However do not truly measure isotropy
- Fundamental shortcomings

IsoScore

- Claim to be the first score that incorporates the *mathematical definition* of isotropy
- Global measure of how uniformly distributed points are in a vector space
- Robust to changes in distribution mean and scalar changes in covariance
- Rotation Invariant
- Increases linearly with more dimensions
- Not skewed by highly isotropic subspaces within the data

Narrow Cones

- Several studies conclude anisotropic
- Most calculate the average cosine similarity of a small number of randomly sampled pairs of points
- Ethayarajh 2019 claims in some cases, contextualized embedding models have avg. random cosine similarity that approaches one
- → all points oriented in the same direction

Zero-M

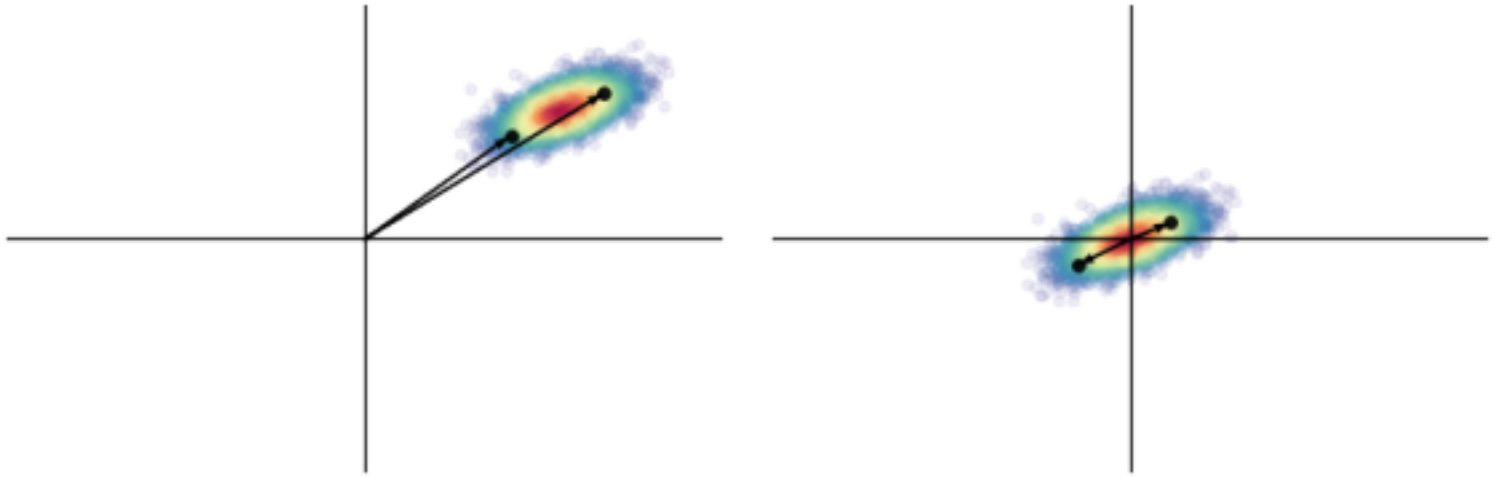


Figure 2: *Left:* Point cloud $X \subset \mathbb{R}^2$. *Right:* Result of applying a zero-mean transform to X.

- Average random cosine similarity & partition score are significantly influenced by the mean of the data irrespective of how data points are distributed in vector space.
- Normalize data to have zero-mean → average random cosine similarity & partition score will artificially produce a score that reflects maximal isotropy

word embedding models have non-zero mean vectors

“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.”

Cosine Similarity

- Long-Used to capture “semantic” difference in *static* word embeddings
- Adapting for *contextualized* obscures true distribution

Existing Methods

Average Random Cosine Similarity

- 100,000 randomly sampled pairs
- 1 - score (compare to other methods)

Partition Isotropy

The idea of isotropy comes from the partition function defined in (Arora et al., 2016),

$$Z(c) = \sum_{w \in \mathcal{V}} \exp\left(c^\top v(w)\right),$$

Intrinsic Dimensionality

- Try and calculate the true dimension of a given manifold
- Been used to argue word embeddings are anisotropic
- Here they use MLE (Levina and Bickel, 2004)
- X ⊆ R^n … divide by n

Variance Explained Ratio

- How much of the total variance is explained by the first k principal components
- k/n
- Specify a priori the number of principal components we wish to examine
- Makes comparisons between vector spaces with different dimensions difficult and results in undesirable behavior
- Particularly when the dimension of the vector space is large

Definitions

Definition of Isotropy

- Variance is uniformly distributed across all dimensions
- Anisotropic is when variance is dominated by single dimension
- Robust scores should be high for balls and low for lines
- Medium isotropy: use ~ n/2 dimensions
- Dimension needs to be > 1
- Exclude single points

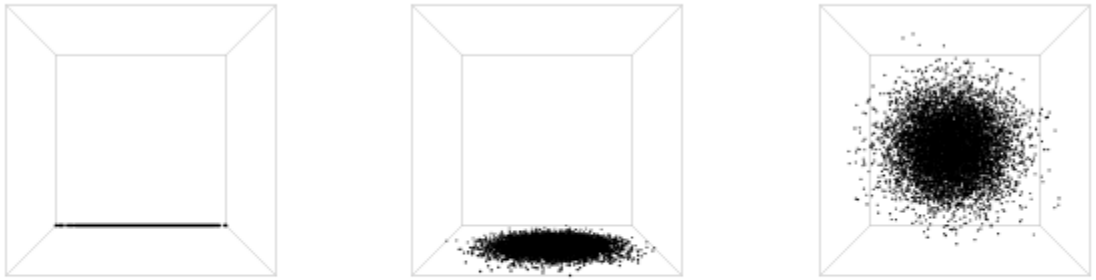


Figure 1: From left to right, a line, disk, and ball embedded in 3D space.

Given a point cloud $X \subseteq \mathbb{R}^n$

$I_n^{(k)}$

$n \times n$

$a_{i,i} = 1 \text{ for } i \in \{1, 2, ..., k\}$

When n = k → Identity Matrix

1	0	0	.3
0	1	0	.3
0	0	1	.3
.3	.3	.3	.1

$I_3^{(1)}$ $I_3^{(2)}$ $I_3^{(3)}$

$\mathbb{R}^3$

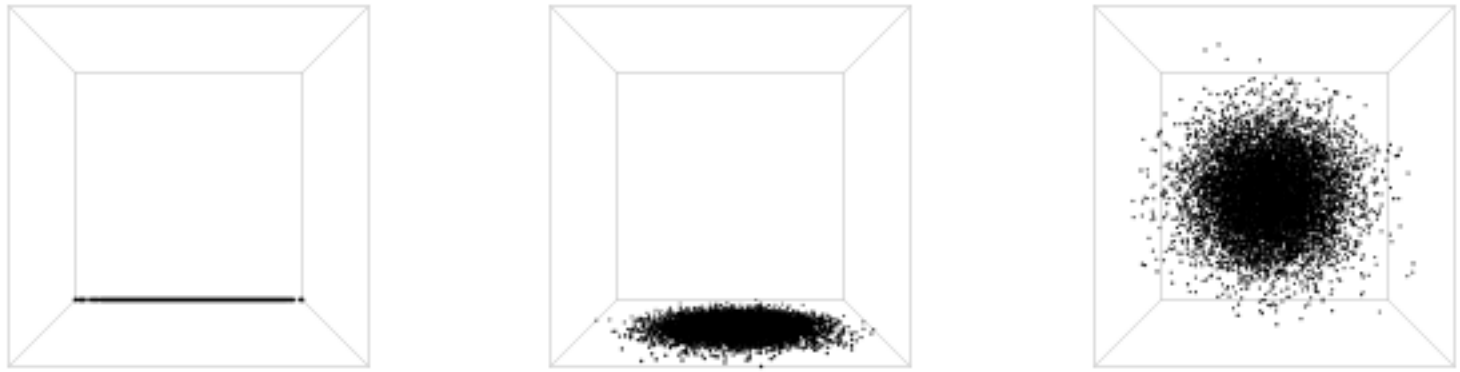


Figure 1: From left to right, a line, disk, and ball embedded in 3D space.

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.*

- The leftmost panel uniformly utilizes all dimensions of 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.

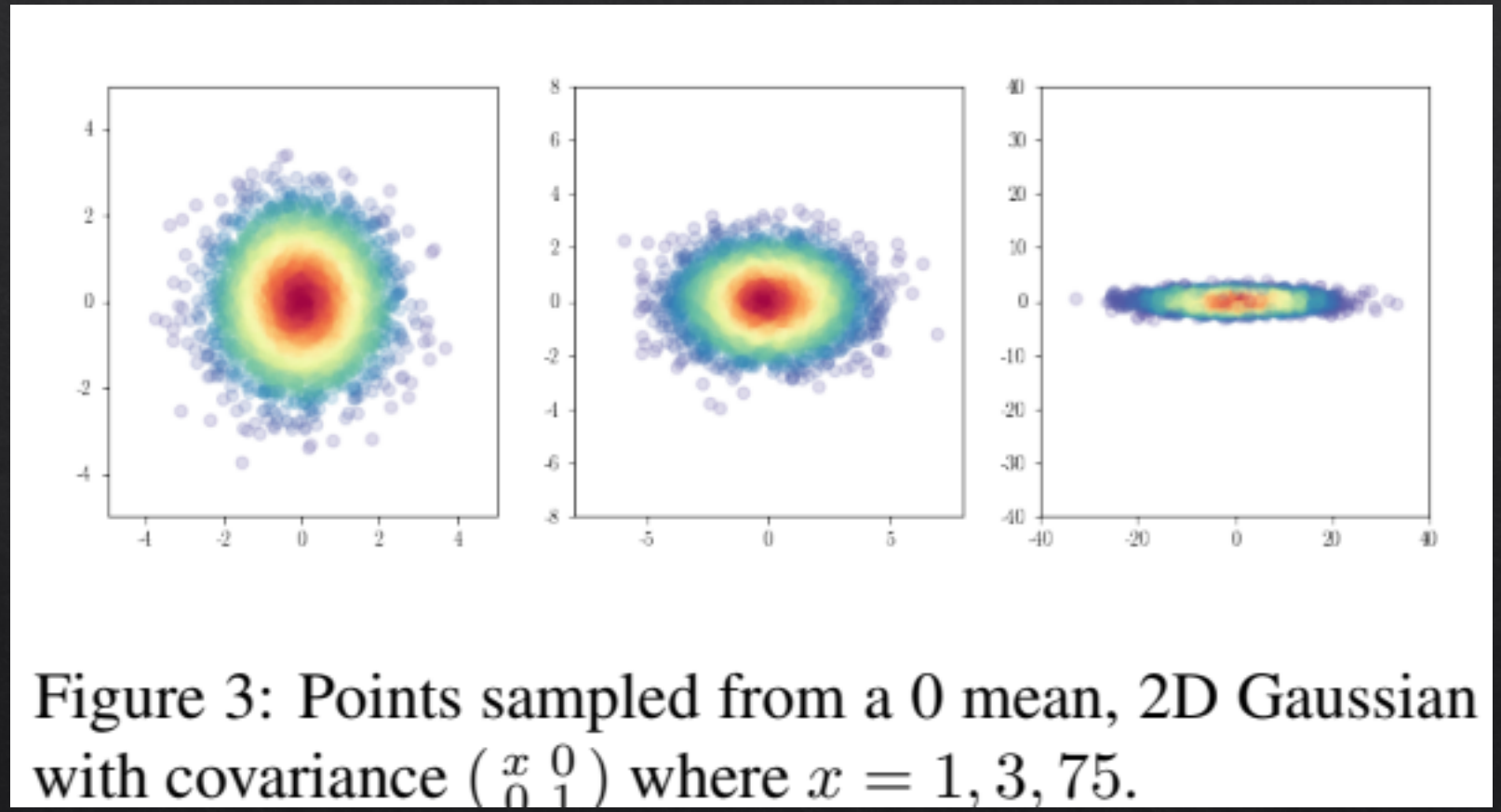


Figure 3: Points sampled from a 0 mean, 2D Gaussian with covariance $\left(\begin{smallmatrix} x & 0 \\ 0 & 1 \end{smallmatrix}\right)$ where $x = 1, 3, 75$.

Essential Properties of Isotropy

Mean Agnostic

- Isotropic if variance is uniform across dimensions
- 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.

Scalar Changes to the Covariance Matrix

- 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

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
- 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.
- Larger maximum variance values reduce the efficiency of spatial utilization.

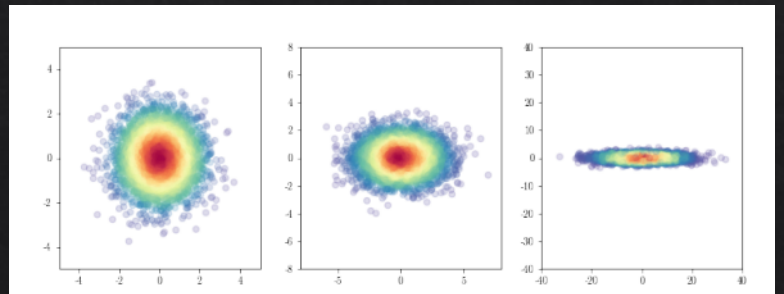


Figure 3: Points sampled from a 0 mean, 2D Gaussian with covariance $\begin{pmatrix} x & 0 \\ 0 & 1 \end{pmatrix}$ where $x = 1, 3, 75$.

Rotatio

- An i
  since
- Cons
  proje

s of X
n
er

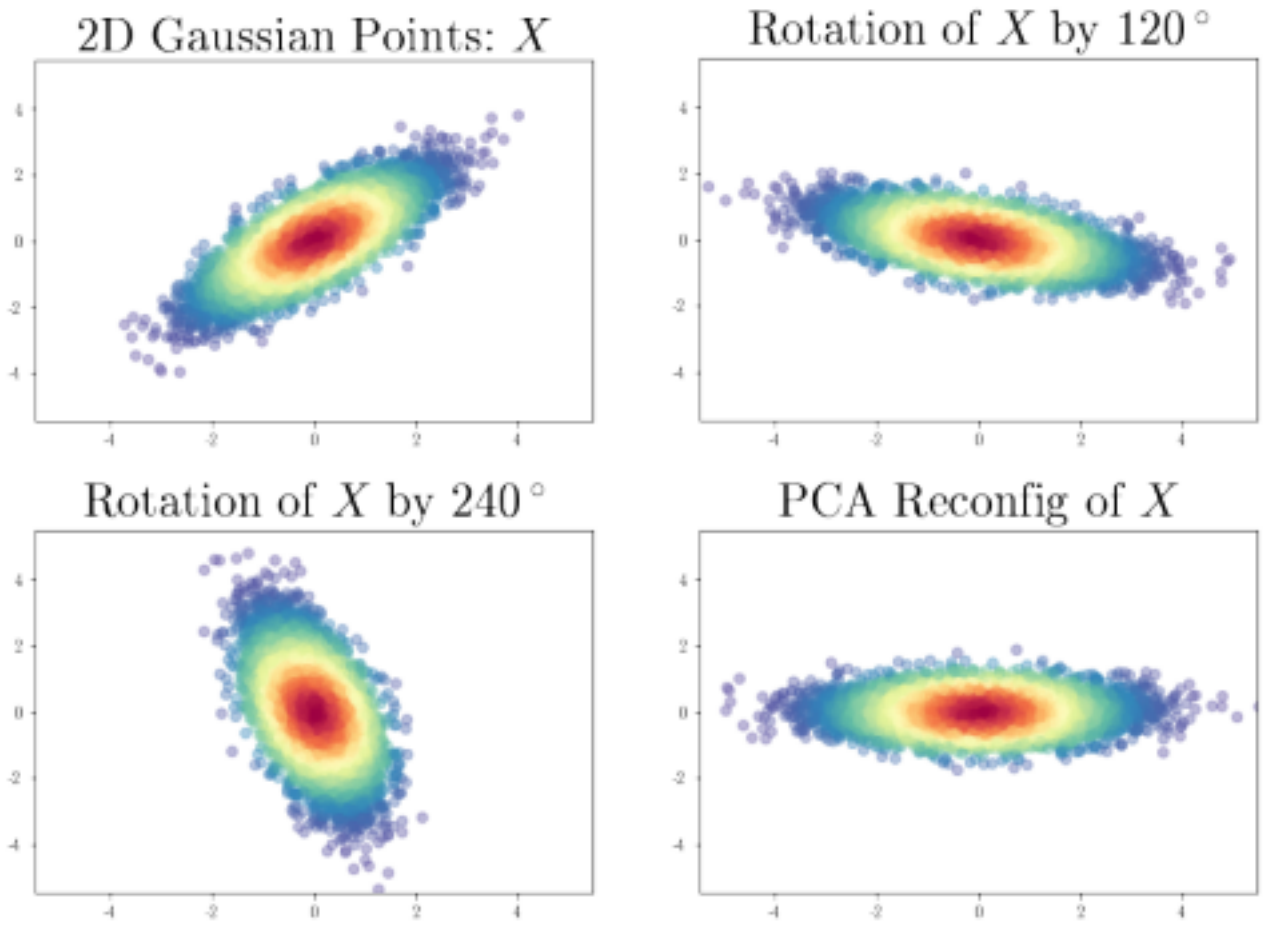


Figure 4: Left: 2D zero-mean Gaussian with covariance $\begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$. We rotate X by 120° and 240°, respectively. Right: Points after applying PCA reorientation.

Dimensions Used

- Good score of spatial utilization should increase linearly as we increase the number of dimensions uniformly utilized by the data

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 extr

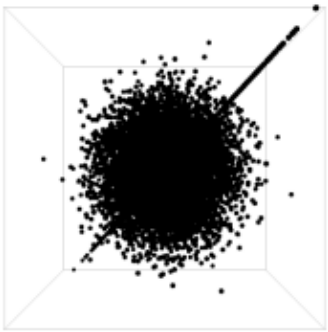


Figure 5: 2D rendering of a line in 3D space intersecting noisy sphere. AKA “skewered meatball.”

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.

IsoScore

Step 1: Start with a point cloud $X \subseteq \mathbb{R}^n$

IsoScore is a finite subset of R^n that outputs a score [0,1]

Step 2: PCA-reorientation of data set.

- Target dimension is original n
- Reorients the axes so ith coordinate accounts for ith greatest variance
- Eliminates correlation between dimensions and makes it diagonal

X^{PCA}

Step 3: Compute variance vector of reoriented data

Σ $n \times n$ covariance matrix X^{PCA}

Σ_D diagonal of the covariance matrix.

Σ_D as the *variance vector*,

Step 4: Length normalization of variance vector

normalized variance vector

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

$$\|(x_1, ..., x_n)\| := \sqrt{x_1^2 + \cdots + x_n^2}$$

$$\|\hat{\Sigma}_D\| = \sqrt{n}$$

Step 5: Compute the distance between the covariance matrix and identity matrix

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

$$\|\hat{\Sigma}_D\| = \|\mathbf{1}\| = \sqrt{n}$$

$$\|\hat{\Sigma}_D - \mathbf{1}\| \in [0, 2\sqrt{n}]$$

$$\|\hat{\Sigma}_D - \mathbf{1}\| = \sqrt{2(n - \sqrt{n})}$$

Step 6: Use the isotropy defect to compute percentage of dimensions isotropically utilized

$$k(X) = (n - \delta(X)^2(n - \sqrt{n}))^2/n$$

$$\delta(X) \in [0, 1] \qquad k(X) \in [1, n]$$

$$\phi(X) := k(X)/n \in [1/n, 1]$$

Step 7: Linearly scale percentage of dimensions utilized to obtain IsoScore

$\phi(X)$ $[1/n, 1]$.

$1/n \mapsto 0$ and $1 \mapsto 1$

$x \mapsto (nx - 1)/(n - 1)$

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

Geometric Interpretation

- Intuitively, our heuristic says that ι(X) is roughly the fraction of dimensions of R n utilized by X.
- k = 1.9996, 1.6105, 1.0281
- x = 1, 3, 75
- “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.”

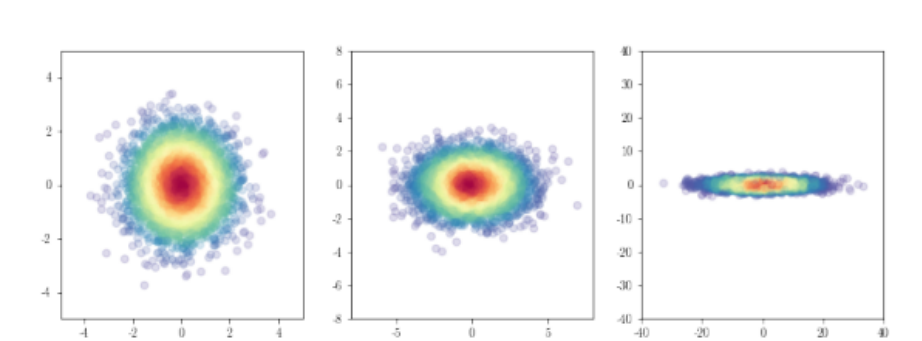


Figure 3: Points sampled from a 0 mean, 2D Gaussian with covariance $\left(\begin{smallmatrix} x & 0 \\ 0 & 1 \end{smallmatrix}\right)$ where $x = 1, 3, 75$.

$\iota(I_9^{(1)})$	$\iota(I_9^{(2)})$	$\iota(I_9^{(3)})$	$\iota(I_9^{(4)})$	$\iota(I_9^{(5)})$	$\iota(I_9^{(6)})$	$\iota(I_9^{(7)})$	$\iota(I_9^{(8)})$	$\iota(I_9^{(9)})$
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 $\iota(I_n^{(k)}) = \frac{k-1}{n-1}$.

Experiments

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

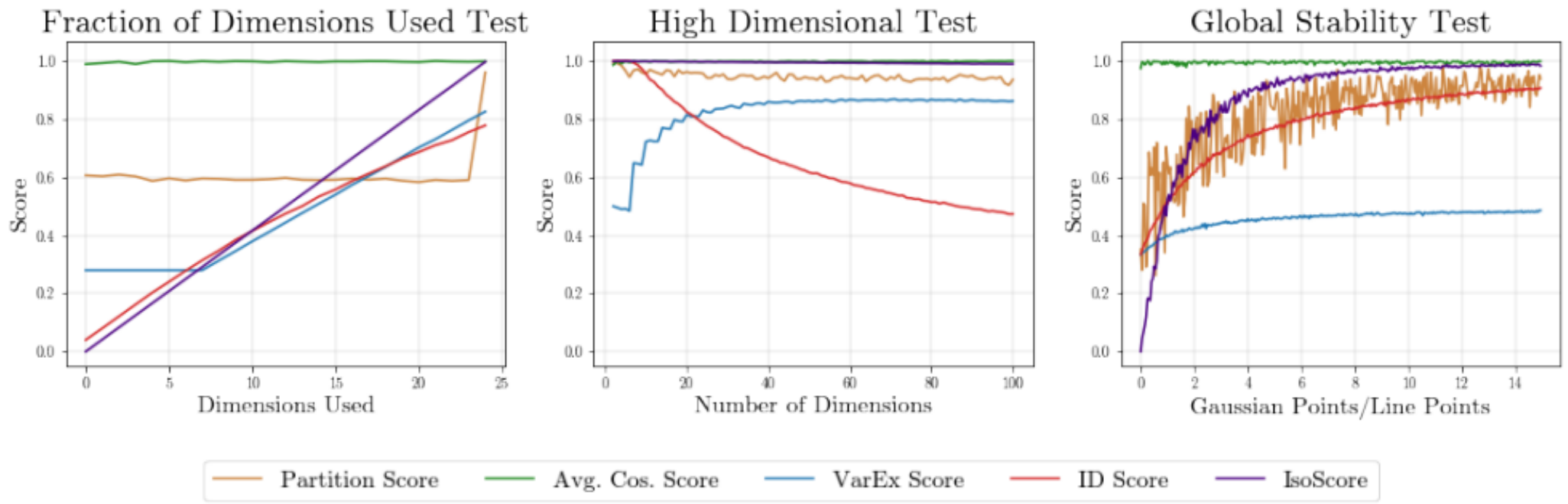


Figure 7: Left and center: Scores for the two Dimensions Used tests. Right: Scores for the "skewered meatball" test in 3 dimensions.

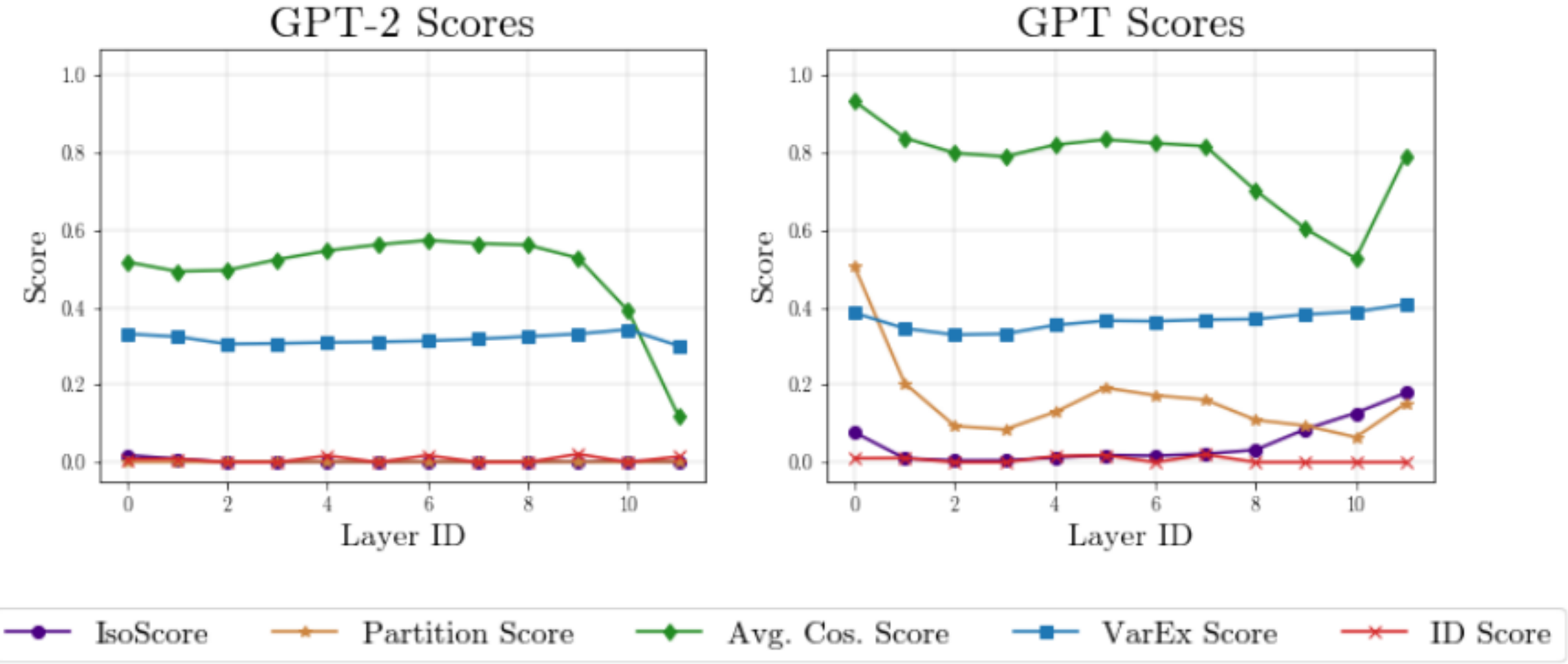


Figure 8: The 5 scores for each of the 12 layers of GPT-2 and GPT

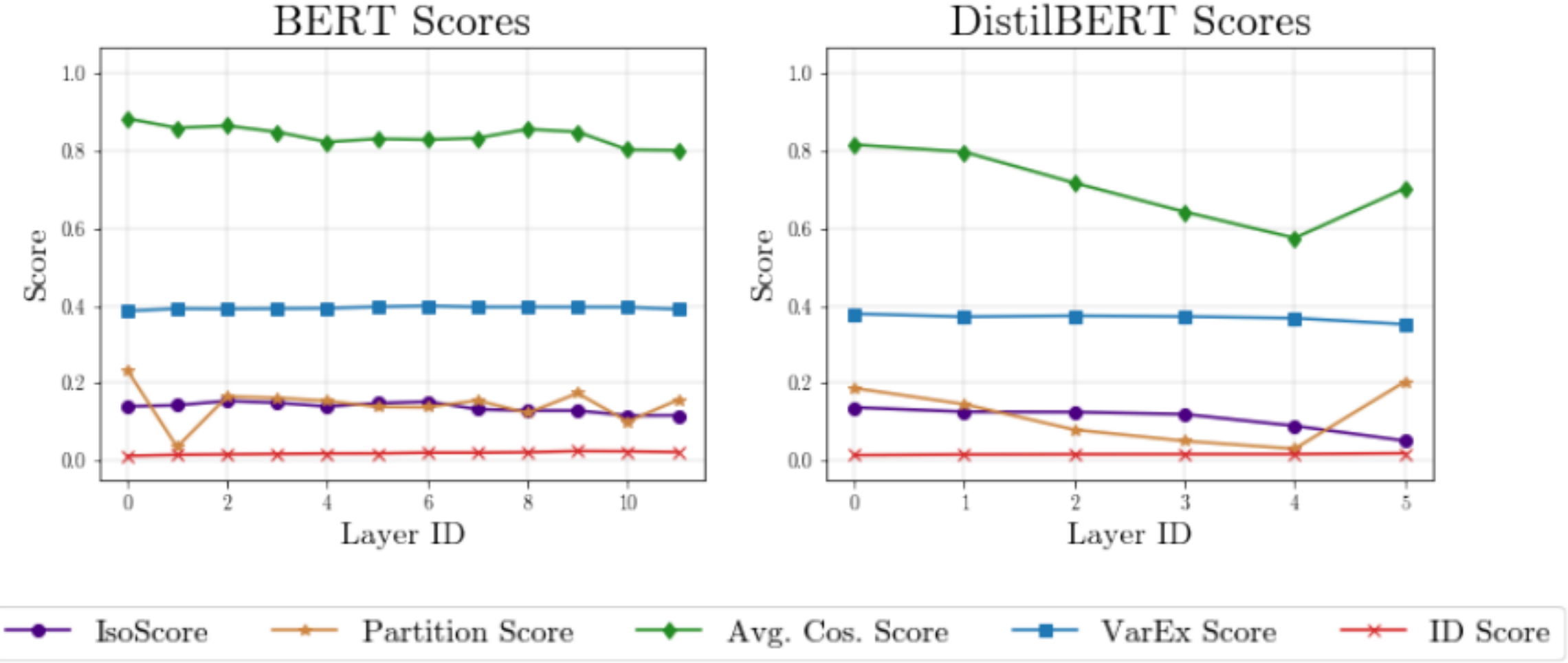


Figure 9: The 5 scores for the 12 layers of BERT, and the 6 layers of DistilBERT

Contextualized Embeddings

- (i) utilize even fewer dimensions than previously thought;
- (ii) do not utilize fewer dimensions in deeper layers;
- (iii) in agreement with Bis et al. ´ (2021), contextualized embedding models do not necessarily occupy a “narrow cone” in space

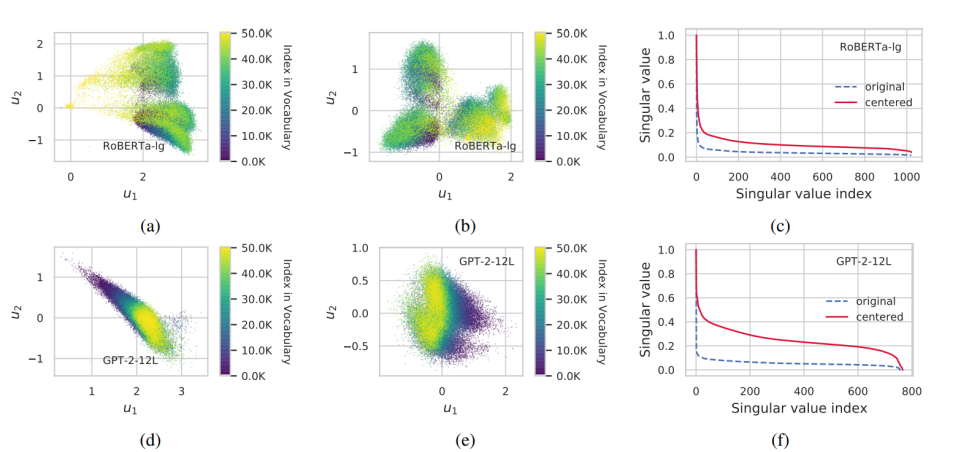


Figure 1: Top: RoBERTa-large. Bottom: GPT-2 (12 layers). (1a, 1d): Word embeddings projected onto first two singular vectors. (1b, 1e) Centered word embeddings projected onto first two singular vectors. (1c, 1f) Singular values of embedding matrix before and after centering. Centering the embedding matrix increases isotropy of embeddings.

Contextualized Embeddings

“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.”

Contextualized Embeddings

- Average random cosine similarity score finds contextualized embedding models to be much more isotropic then previously reported.
- 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.
- 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

Notion of isotropy is often conflated with geometry

- Geometry of isotropic vector spaces, will differ depending on the distribution that generates the points in space.
- Multivariate isotropic Gaussians form n-dimensional balls
- Uniform distributions form n-dimensional cubes
- Yet both distributions receive an IsoScore of 1

- Each has a covariance matrix proportional to the identity and is therefore maximally isotropic.
- The variance is distributed equally in all directions
- All receive 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.

F Geometry of Isotropy

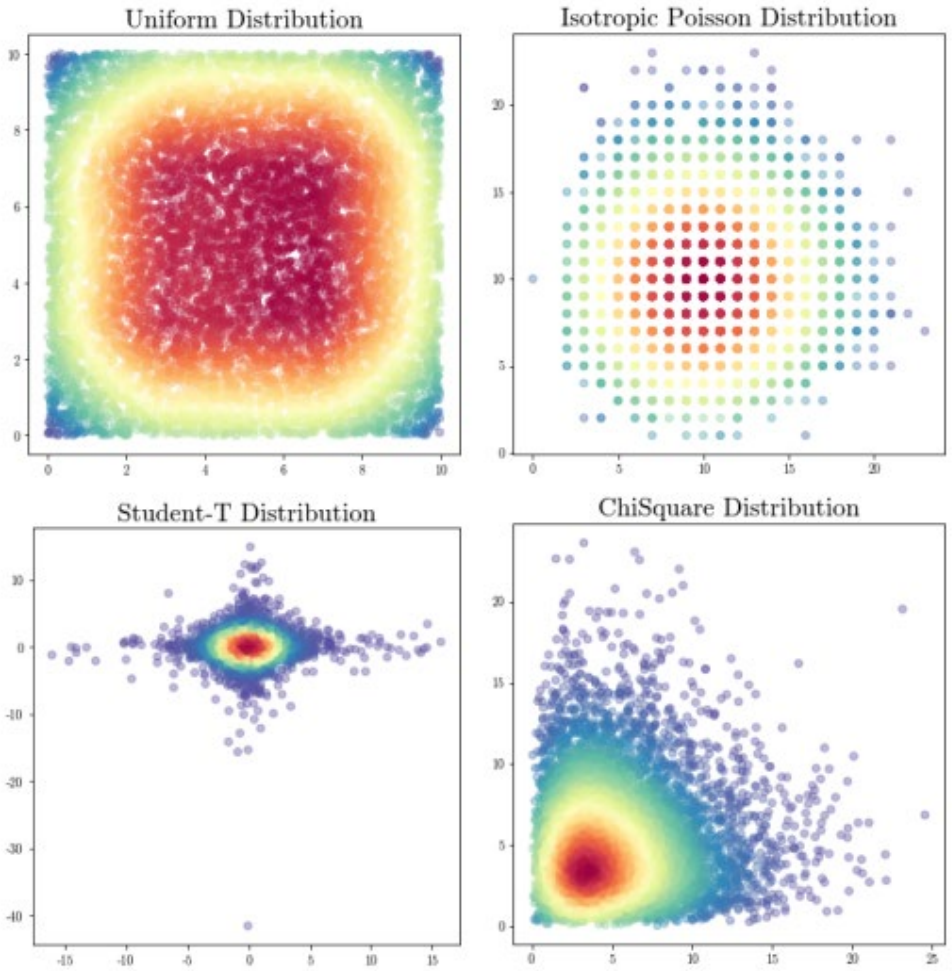


Figure 10: Points sampled from a Uniform distribution, Poisson distribution, Student-T distribution and ChiSquare distribution respectively

What's Missing?

- Same answer as always …. Multilingual

Rajee and Pilehvar 2022

- Homework 3!
- Additional Language
- Additional Models

Cosine Similarity. Ethayarajh (2019) used cosine similarity between random embeddings as an approximation of isotropy in the space. As mentioned before, random embeddings with an isotropic distribution have near-zero cosine similarities. The metric can be formulated as follows:

$$I_{Cos}(\mathcal{W}) = \frac{1}{N} \sum_{i=1, x_i \neq y_i}^{N} Cos(x_i, y_i) \quad (1)$$

where $x_i \in X, y_i \in Y$, X and Y are the sets of randomly sampled embeddings, and $\mathcal{W}$ is the embedding matrix. N is the number of sampled pairs that is set to 1000 in our experiments. Lower $I_{Cos}(\mathcal{W})$ values indicate higher isotropy.

	$I_{Cos}(\mathcal{W})$	**First**	**Second**	**Third**
BERT	0.34	0.385	0.005	0.005
English	0.24	0.041	0.029	0.020
Spanish	0.27	0.033	0.029	0.018
Arabic	0.27	0.033	0.025	0.022
Turkish	0.25	0.036	0.024	0.024
Sundanese	0.25	0.036	0.016	0.016
Swahili	0.27	0.025	0.018	0.014

Table 2: The contribution of top-three dimensions to the expected cosine similarity ($I_{Cos}(\mathcal{W})$) in BERT and mBERT models.

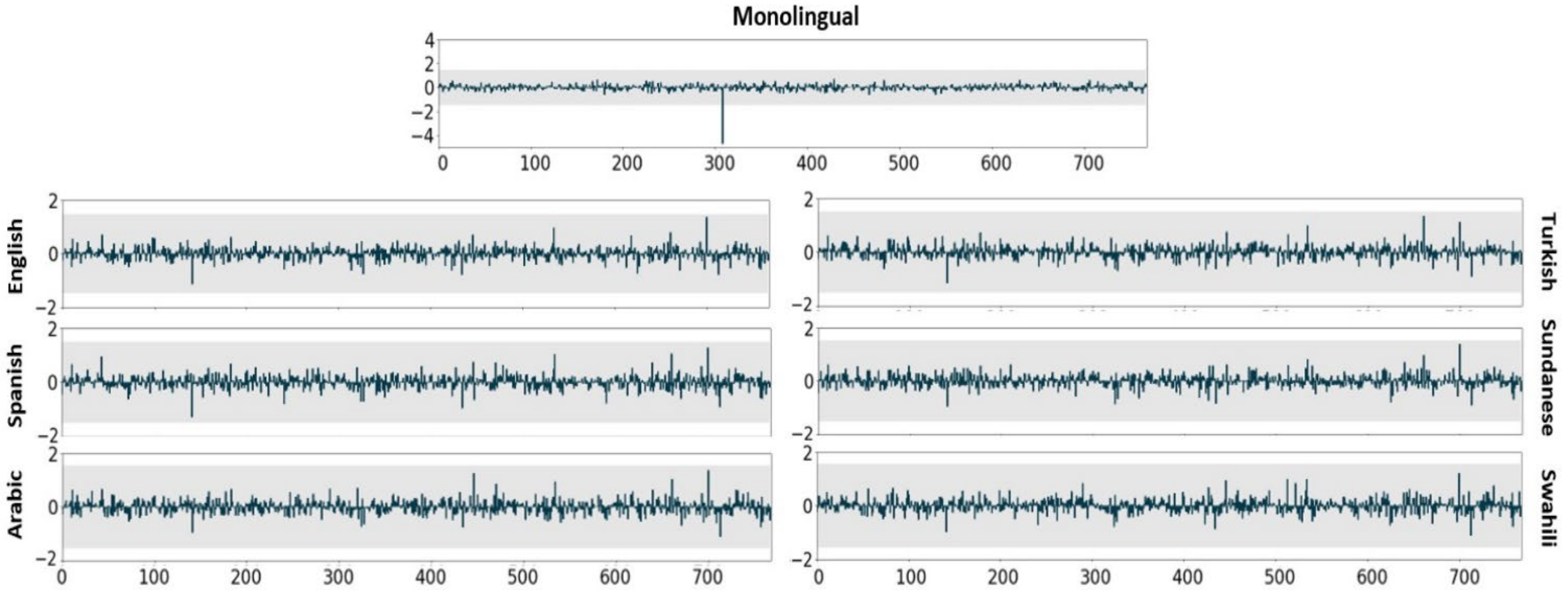


Figure 2: The average representation in English BERT (top) and mBERT (bottom). The shaded area denotes 3σ. While an outlier has emerged in the former, we do not see any major outliers in the multilingual space.

Multilingual Model IsoScore

- Rest of HW3