Revisiting Cosine Similarity via Normalized ICA-transformed Embeddings

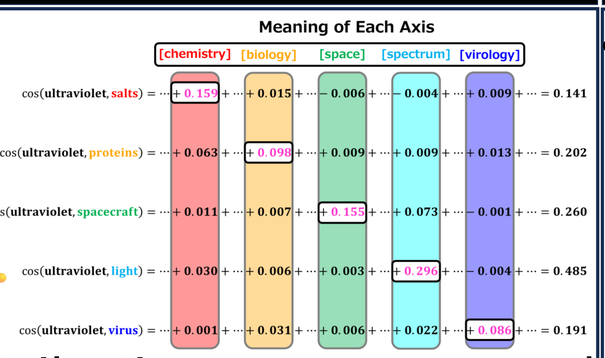
Hiroaki Yamagiwa[1], Momose Oyama[1,2], and Hidetoshi Shimodaira[1,2]
[1]Kyoto University [2]RIKEN AIP

The 31st International Conference on Computational Linguistics
COLING 2025 • Abu Dhabi

arXiv

Main Contributions

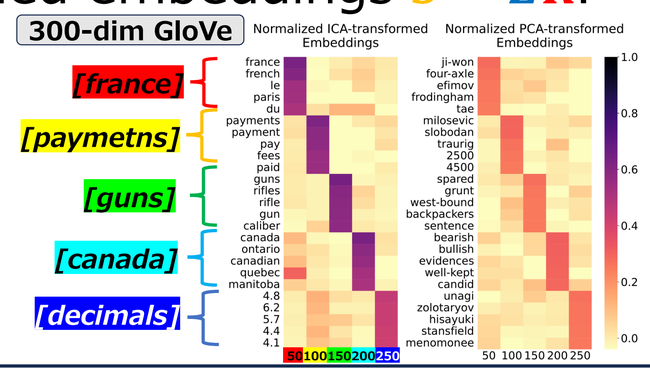
- We proposed interpreting cosine similarity as the sum of semantic similarities across axes. 👉



- By deriving probability distributions, we proposed a method to select statistically significant axes.
- We demonstrated that ICA provides better interpretability than PCA, and that normalization enhances this interpretability.

Background

- For embeddings $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ (n words, d-dim), the **PCA**-transformed embeddings are $\boldsymbol{Z} = \boldsymbol{XA}$. FastICA [Hyvärinen, 1999] computes the **orthogonal** matrix $\boldsymbol{R} \in \mathbb{R}^{d \times d}$ for the **ICA**-transformed embeddings $\boldsymbol{S} = \boldsymbol{ZR}$.
- The axes of the ICA-transformed embeddings are interpretable [Musil+2024]. 👉



Cosine Similarity Interpretation

Cosine Similarity The ICA-transformed embedding of a word w_i, denoted by $\boldsymbol{s}_i \in \mathbb{R}^d$, is normalized to $\hat{\boldsymbol{s}}_i$:

$$\hat{\boldsymbol{s}}_i = \boldsymbol{s}_i / \|\boldsymbol{s}_i\| = \left(\ldots, \hat{s}_i^{(\ell)}, \ldots\right).$$

For words w_i and w_j,

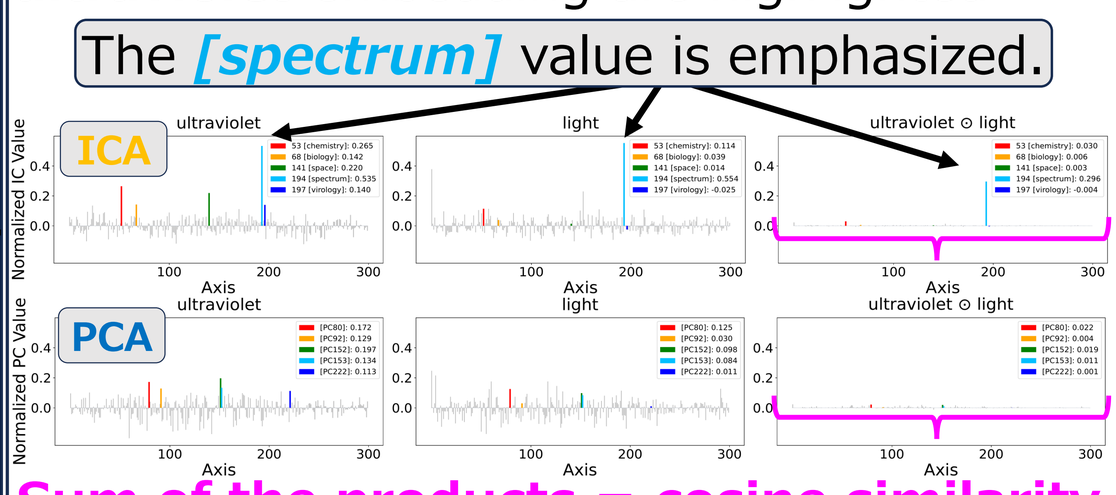
$$\cos(w_i, w_j) = \hat{\boldsymbol{s}}_i^{\top} \hat{\boldsymbol{s}}_j = \sum_{\ell=1}^{d} \hat{s}_i^{(\ell)} \hat{s}_j^{(\ell)}.$$

The normalized PCA-transformed embedding $\hat{\boldsymbol{z}}_i$ satisfies $\hat{\boldsymbol{z}}_i = \boldsymbol{R}\hat{\boldsymbol{s}}_i$. Then

$$\hat{\boldsymbol{z}}_i^{\top} \hat{\boldsymbol{z}}_j = \hat{\boldsymbol{s}}_i^{\top} \hat{\boldsymbol{s}}_j = \cos(w_i, w_j).$$

Example Bar graphs of normalized GloVe embeddings (***ultraviolet***, ***light***, and **their component-wise products**) are shown for **ICA** and **PCA**. The top five axes of the ***ultraviolet*** embedding are highlighted.

The *[spectrum]* value is emphasized.



Sum of the products = cosine similarity.
ICA shows greater sparsity than PCA.

Interpretation The semantic similarity on the ℓ-th axis is defined as:

$$\mathrm{sem}_\ell(w_i, w_j) \coloneqq \hat{s}_i^{(\ell)} \hat{s}_j^{(\ell)}.$$

e.g. for ***ultraviolet*** and ***light***, *the semantic* similarity on the *[spectrum]* axis is $\mathrm{sem}_\ell(\textit{ultraviolet, light}) = 0.296$.

We interpret $\cos(w_i, w_j)$ as the sum of $\mathrm{sem}_\ell(w_i, w_j)$:

$$\cos(w_i, w_j) = \sum_{\ell=1}^{d} \mathrm{sem}_\ell(w_i, w_j).$$

additive compositionality of semantic similarities

Selection of Statistically Significant Axes

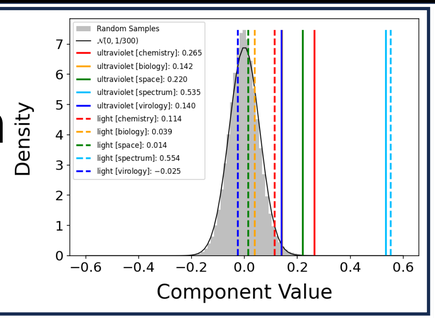
Distribution of Components

$\hat{s}_i^{(\ell)}$ follows a normal distribution

$$\hat{s}_i^{(\ell)} \sim \mathcal{N}(0, 1/d).$$

mean: 0
variance: $1/d$

Histogram fits with the curve.



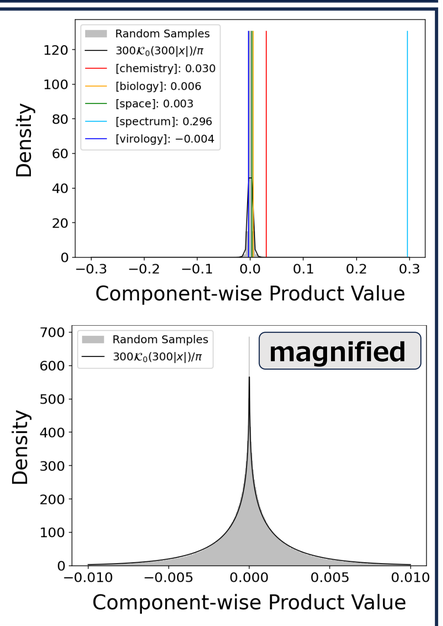
Distribution of Products

The probability density function of $\hat{s}_i^{(\ell)} \hat{s}_j^{(\ell)}$ ($i \neq j$, $\hat{s}_i^{(\ell)} \perp \hat{s}_j^{(\ell)}$) is

$$\hat{s}_i^{(\ell)} \hat{s}_j^{(\ell)} \sim (d/\pi) K_0\left(d \left| \hat{s}_i^{(\ell)} \hat{s}_j^{(\ell)} \right|\right),$$

where $K_0(\cdot)$ is the modified Bessel function of the second kind of order zero.

Histogram fits with the curve.



Axis Selection Using Distributions The p-value p_1 can be computed for each axis. To select from d axes, we account for the multiplicity issue (Bonferroni correction) and use $p_d = \min\{dp_1, 1\}$. We compare the top components for ***ultraviolet*** and ***light*** with $\alpha = 0.05$.

ICA

Axis	Meaning	Value	p-value	Bonferroni
194	*[spectrum]*	0.535	9.97×10^{-21}	2.99×10^{-18}
53	*[chemistry]*	0.265	2.17×10^{-6}	6.51×10^{-4}
141	*[space]*	0.220	6.97×10^{-5}	2.09×10^{-2}
68	*[biology]*	0.142	6.88×10^{-3}	1.00
197	*[virology]*	0.140	7.53×10^{-3}	1.00

(a) *ultraviolet*

Axis	Meaning	Value	p-value	Bonferroni
194	*[spectrum]*	0.554	4.44×10^{-22}	1.33×10^{-19}
1	*[function words]*	0.379	2.67×10^{-11}	8.02×10^{-9}
153	*[boxing]*	0.230	3.35×10^{-5}	1.00×10^{-2}
53	*[chemistry]*	0.114	2.37×10^{-2}	1.00
58	*[ordinal]*	0.114	2.46×10^{-2}	1.00

(b) *light*

Axis	Meaning	Value	p-value	Bonferroni
194	*[spectrum]*	0.296	1.12×10^{-40}	3.36×10^{-38}
53	*[chemistry]*	0.030	1.38×10^{-5}	4.15×10^{-3}
1	*[function words]*	0.022	1.69×10^{-4}	5.07×10^{-2}
153	*[boxing]*	0.014	2.56×10^{-3}	7.68×10^{-1}
158	*[police]*	0.007	2.76×10^{-2}	1.00

(c) *ultraviolet* ⊙ *light*

PCA

Axis	Value	p-value	Bonferroni
152	0.197	3.21×10^{-4}	9.62×10^{-2}
80	0.172	1.41×10^{-3}	4.22×10^{-1}
153	0.134	1.01×10^{-2}	1.00
92	0.129	1.30×10^{-2}	1.00
222	0.113	2.52×10^{-2}	1.00

(a) *ultraviolet*

Axis	Value	p-value	Bonferroni
1	0.247	9.12×10^{-6}	2.74×10^{-3}
5	0.216	9.38×10^{-5}	2.81×10^{-2}
61	0.176	1.15×10^{-3}	3.44×10^{-1}
118	0.176	1.16×10^{-3}	3.47×10^{-1}
127	0.146	5.76×10^{-3}	1.00

(b) *light*

Axis	Value	p-value	Bonferroni
1	0.024	1.09×10^{-4}	3.26×10^{-2}
80	0.022	2.21×10^{-4}	6.62×10^{-2}
131	0.020	3.80×10^{-4}	1.14×10^{-1}
152	0.019	4.58×10^{-4}	1.38×10^{-1}
60	0.017	9.65×10^{-4}	2.90×10^{-1}

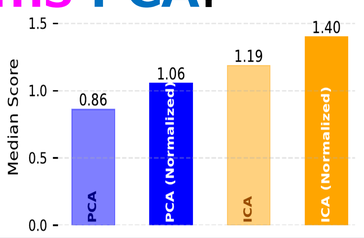
(c) *ultraviolet* ⊙ *light*

ICA results in more statistically significant axes.

Comparative Experiments on ICA and PCA

Normalization We compute the median of interpretability scores for each axis by the word intrusion task [Sun et al., 2016].

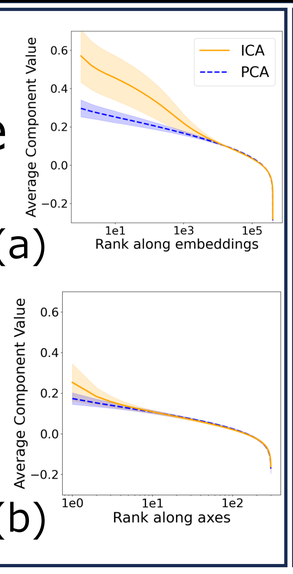
- **ICA outperforms PCA.**
- **Normalization improves further.**



Component Values

Component values are averaged after sorting in descending order (a) along embeddings and (b) along axes.

Both shows larger top values for ICA.



Component-wise Products

Cosine similarity is computed using the **top** k in $\left\{\hat{s}_i^{(\ell)} \hat{s}_j^{(\ell)}\right\}_{\ell=1}^{d}$ for word similarity & analogy tasks.

ICA shows greater sparsity than PCA.

