Discovering Universal Geometry in Embeddings with ICA

arXiv: 2305.13175

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

Purpose

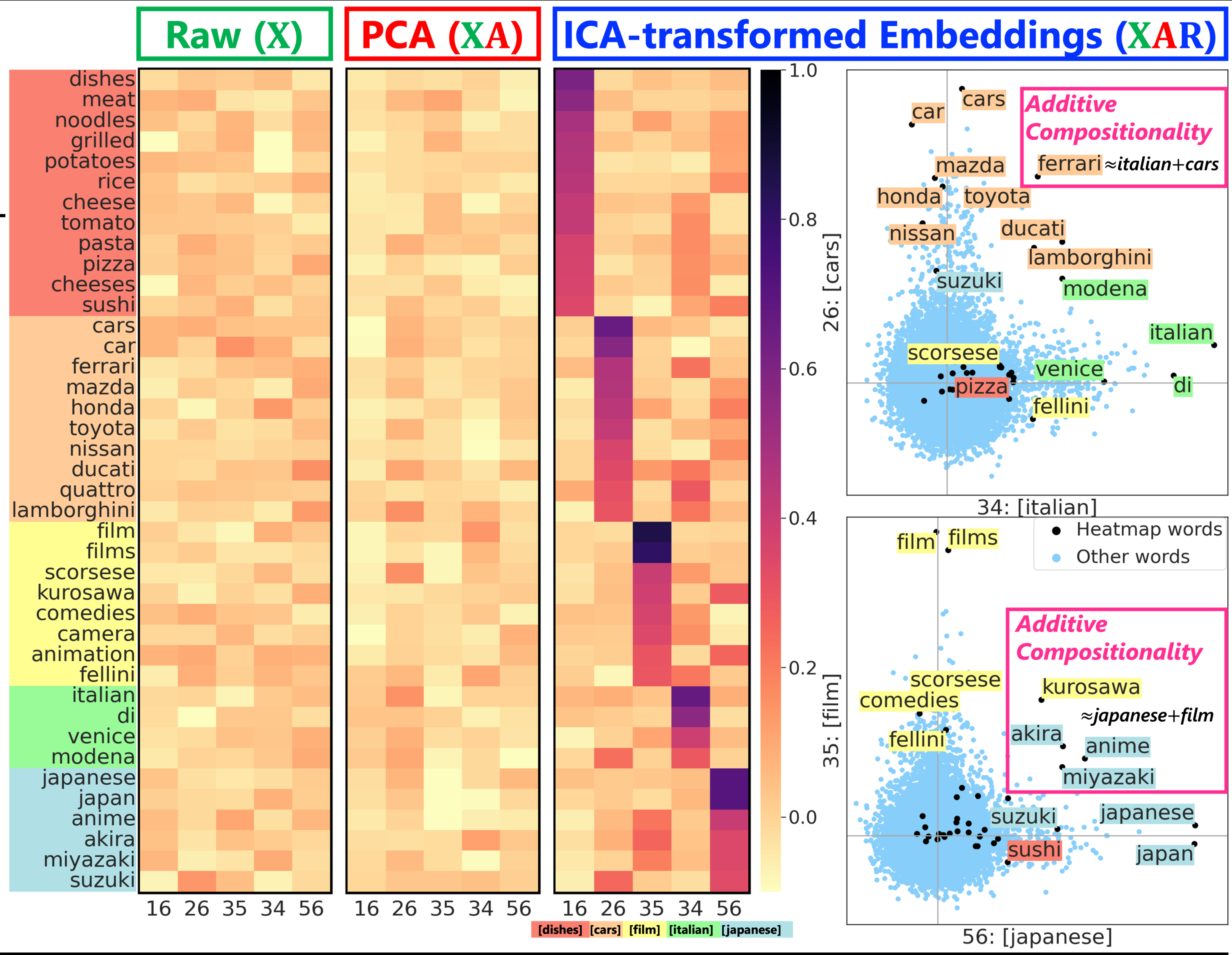
1. Understanding How Embedding Geometry Encodes Meaning
2. Exploring the Universality of Geometric-Meaning Relationships in Embeddings

We use ICA

Independent Component Analysis (ICA) Finds Statistically Independent Axes

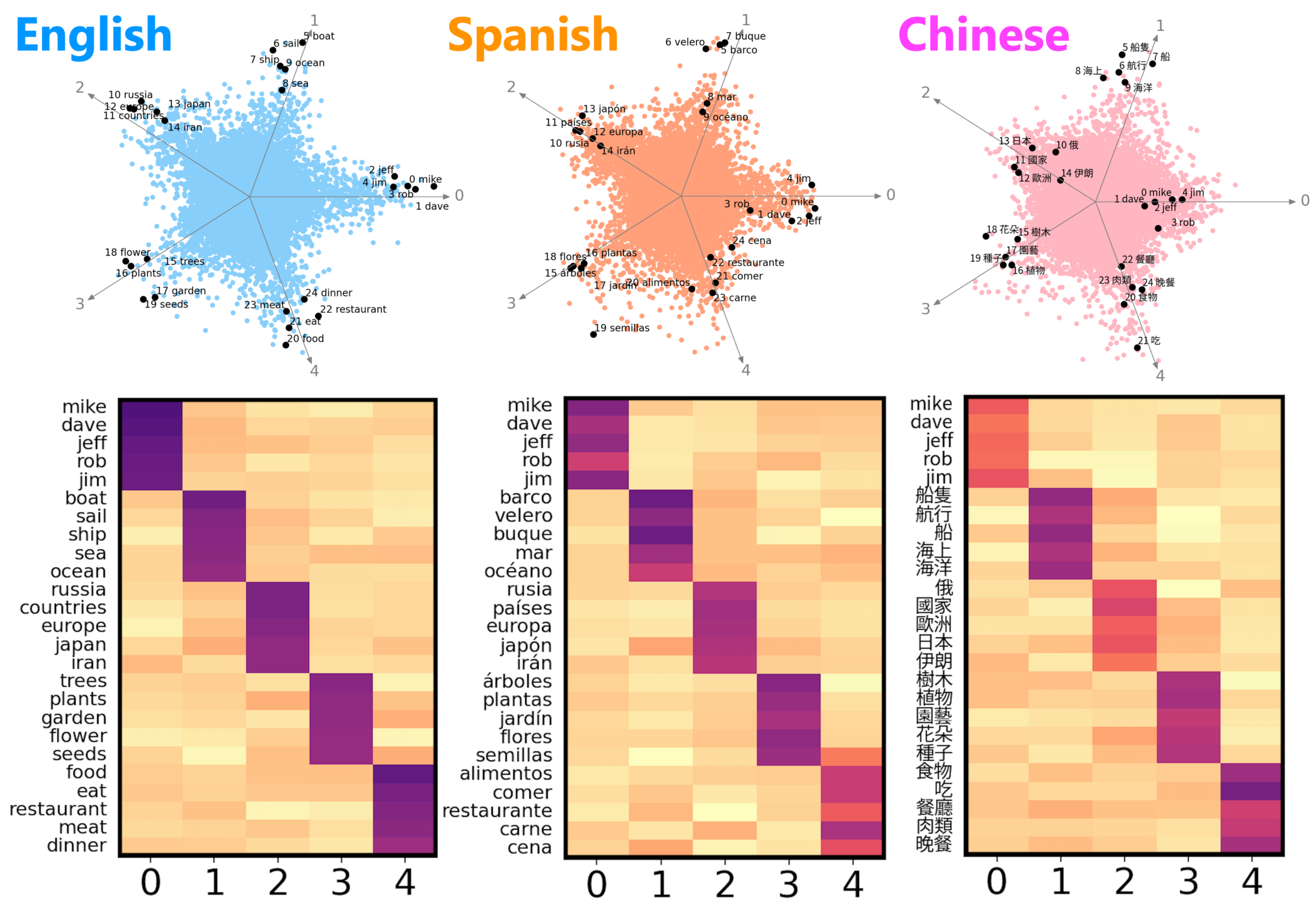
Results

1. Independent Axes in Embeddings are "**Spiky**" and **Interpretable**
2. "**Spiky**" and **Interpretable** Axes are Universal in Various Embeddings



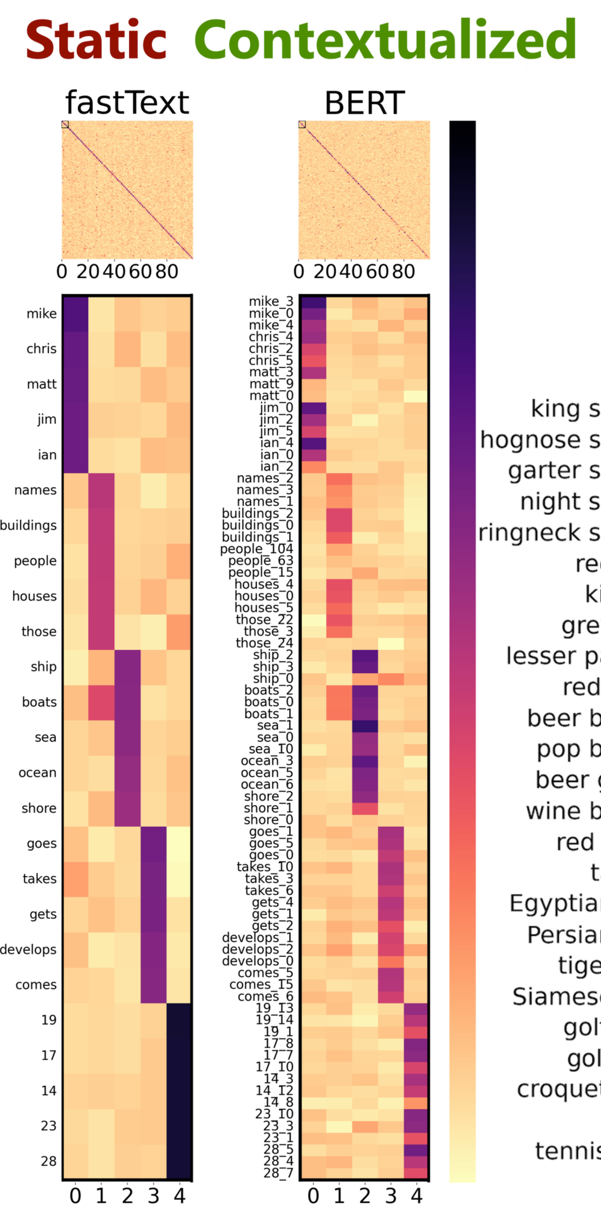
Universal Geometry in Embeddings

① Cross-Lingual

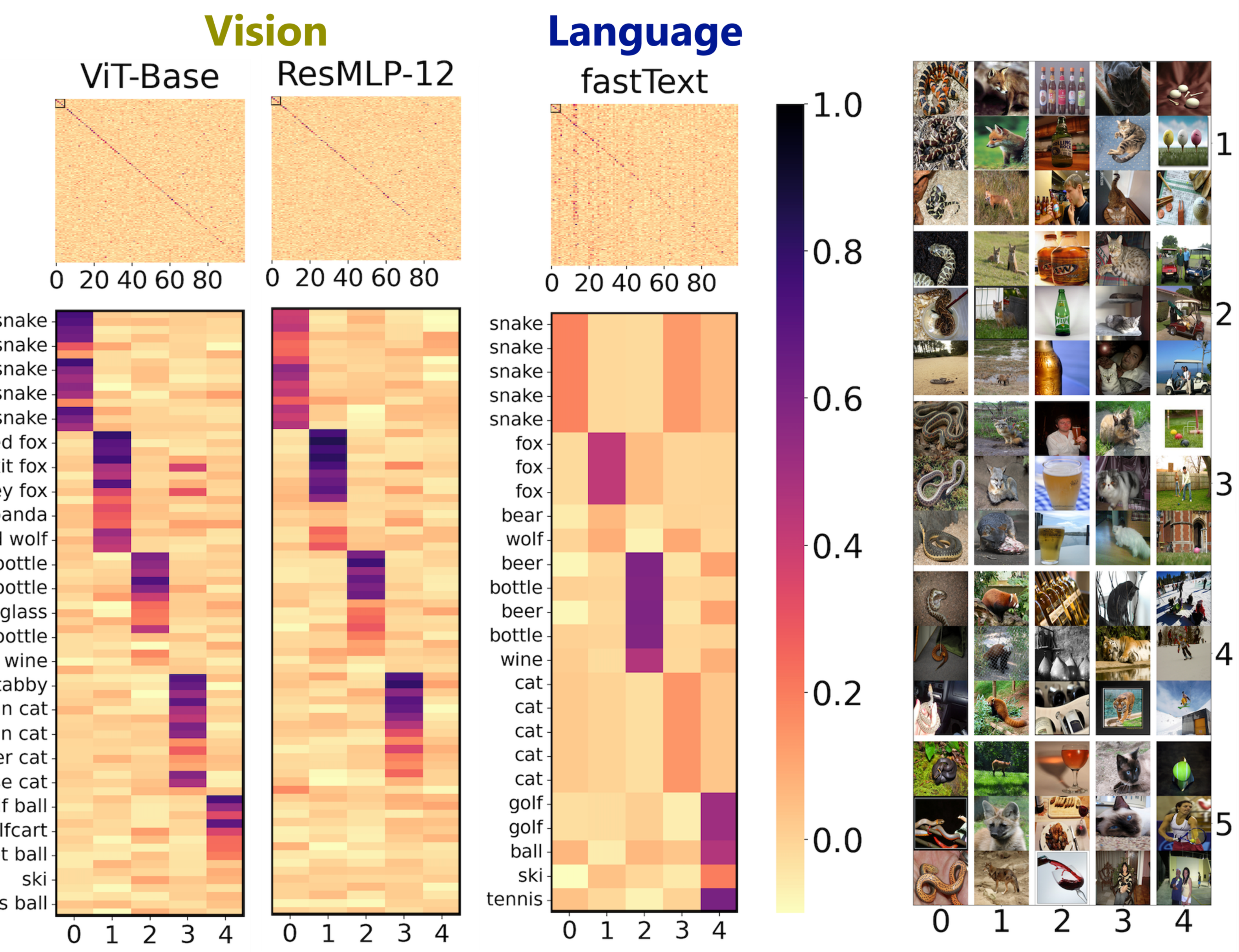


1. Apply ICA to each embedding
2. Permute axes by correlation for translation pairs

② Model



③ Modality



ICA Discovers Spiky Axes

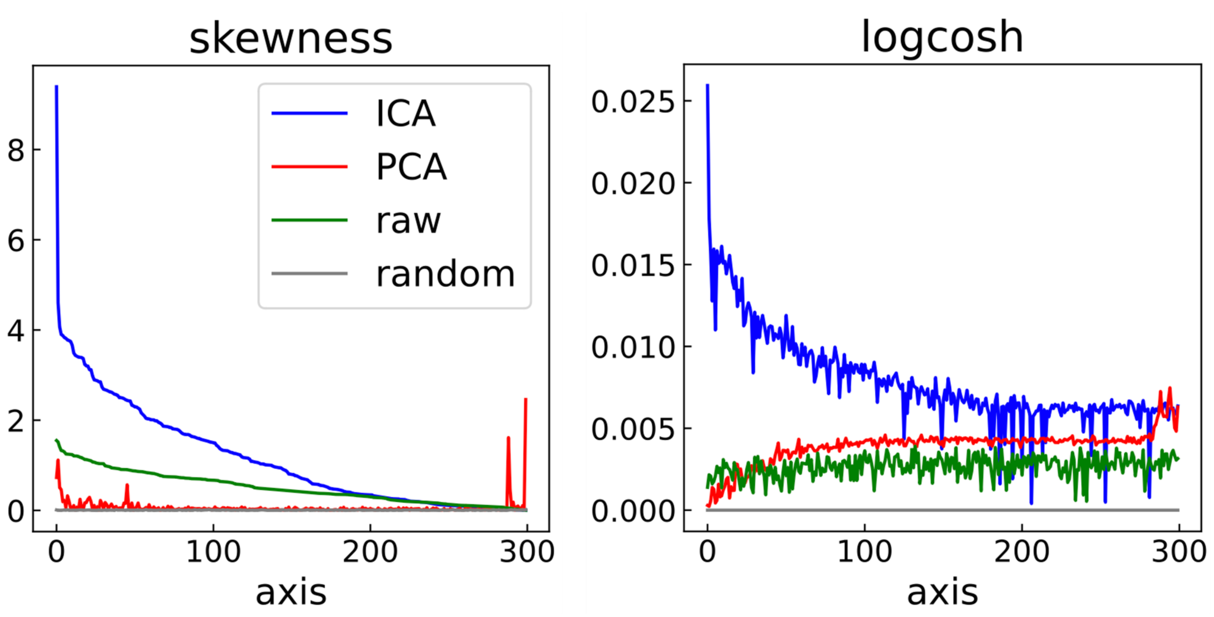
- Two steps for **ICA** to find independent axes
  1. **Whitening (PCA)** : Make each axis uncorrelated
  2. **Orthogonal Transformation**: Maximize the **non-Gaussianity** of each axis

ICA-transformed

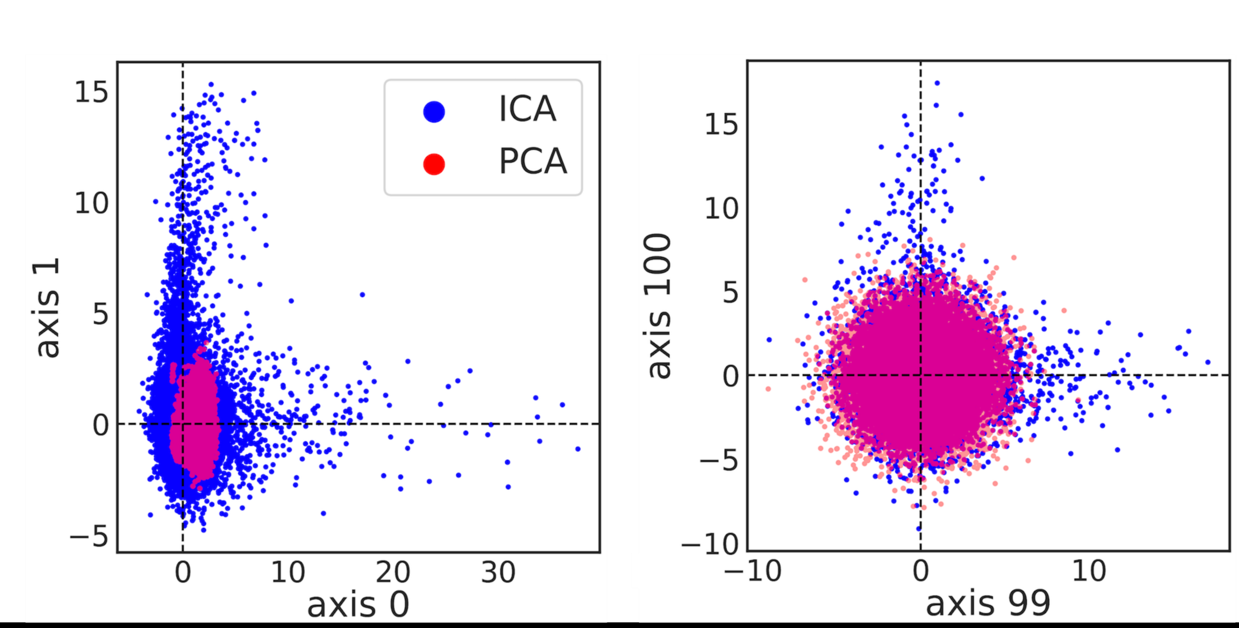
$$\mathbf{S} = \mathbf{XAR}$$

- Embedding distribution along independent axes is "**spiky**"
  - **PCA can't** find the "spiky" axes, providing ***isotropic*** embedding
  - **ICA can** find the "spiky" axes by focusing on ***anisotropic*** information

ICA Maximizes Non-Gaussianity of Each Axis

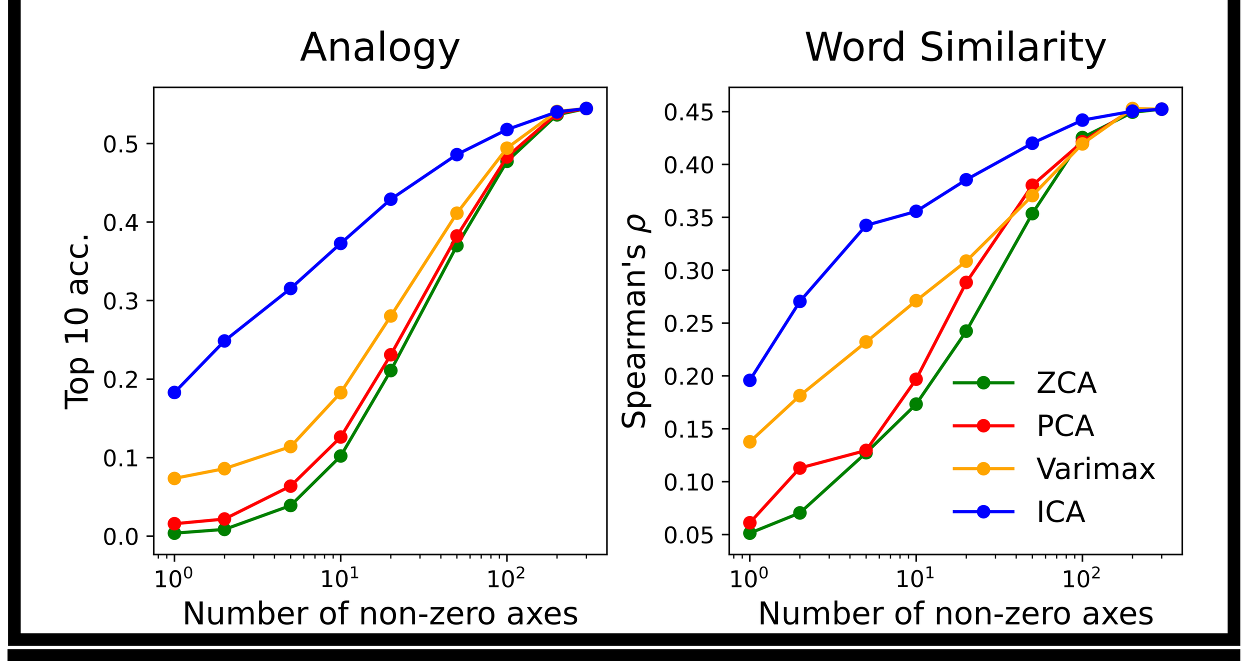


Embedding Distribution along ICA Axes is Spiky



Low-Dimensionality

- **ICA-transformed** embedding represents meaning with **few axes**
- Evaluation Tasks
  1. Analogy
  2. Word Similarity
- Evaluate performance by **reducing the number of non-zero axes**



Ref.
[1] Aapo Hyvärinen and Erkki Oja. Independent Component Analysis: Algorithms and Applications. (Neural Networks 2000)
[2] Tomáš Musil and David Mareček. Independent Components of Word Embeddings Represent Semantic Features (arXiv 2022)