Similarity-based Representation Factorization for Understanding Representations in Minds, Brains and Machines

Florian P. Mahner (mahner@cbs.mpg.de)
Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig 04103, Germany
Department of Medicine, Justus Liebig University, Giessen 35390, Germany
Center for Mind, Brain and Behavior, Universities of Marburg, Giessen and Darmstadt

Martin N. Hebart (hebart@cbs.mpg.de)
Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig 04103, Germany
Department of Medicine, Justus Liebig University, Giessen 35390, Germany
Center for Mind, Brain and Behavior, Universities of Marburg, Giessen and Darmstadt

Abstract

Understanding representations is a major aim in cognitive computational neuroscience, yet existing data-driven methods are limited in providing interpretable dimensions that also capture the underlying data structure. Here we propose Similarity-based Representation Factorization (SRF), a method that reliably decomposes data structures into interpretable, non-negative components based on similarity matrices. Through simulations and empirical data, we demonstrate that SRF is robust to noise and capable of revealing interpretable dimensions in both synthetic and behavioral similarity data. SRF opens new possibilities for uncovering the dimensions that underlie similarity even in smaller and noisier datasets, thus offering a principled approach for interpreting representational structure.

Keywords: RSA; NMF; Representation Learning; Interpretability

Introduction

Understanding the nature of representations in minds, brains, and machines is a central question in cognitive computational neuroscience. To address this, both theory-driven and data-driven approaches have been developed to reveal, evaluate and compare representations. Prominent theory-driven approaches include encoding and decoding methods (Naselaris, Kay, Nishimoto, & Gallant, 2011) and representational similarity analysis (RSA; Kriegeskorte, Mur, and Bandettini, (2008)), while data-driven approaches include principal component analysis (PCA), multi-dimensional scaling (MDS) or t-SNE (Van der Maaten & Hinton, 2008).

Among these, RSA has become widely used due to its ability to operate directly on the representational level, abstracting away from specific implementations and enabling comparisons across different systems. RSA quantifies the degree of similarity between representational geometries, providing insights into how similar different models or brain regions are in their encoding of information. However, RSA typically lacks interpretability and researchers have often resorted to visualization techniques like MDS or t-SNE to identify interpretable axes of variation on top of representational geometries, but these techniques can be hard to interpret beyond three dimensions (t-SNE) or are not optimized for interpretability (MDS).

Providing a multi-dimensional interpretable explanation about the underlying factors driving representational similarity is desirable, since it would al to move beyond merely comparing models and hypothesis testing (as classically done with RSA) to a more fine-grained description that not only reveals *how much* two representations are similar to each other, but also *why* they are similar.

A canonical approach for having interpretable explanations derived from data is non-negative matrix factorization (NMF). NMF creates part-based and additive representations by decomposing data into non-negative components, making it particularly well-suited for interpretability. In this paper we introduce a new toolkit for the cognitive sciences that marries the benefits of representational geometries with the interpretability gained NMF. To this end, we introduce a framework we term *Similarity-based Representation Factorization (SRF)* that applies NMF to similarity matrices to yield interpretable latent representational geometries. Unlike RSA, *SRF* does not require a separate statistical framework and is embedded within classical inferential statistics, making it more accessible and flexible for a wide range of applications. This framework allows for the identification of interpretable, part-based representations from any symmetric similarity matrix, including those derived from behavioral responses or kernel-based methods. By providing a more interpretable way to analyze representations, *SRF* has the potential to advance our understanding of cognitive neuroscience, psychology, and artificial intelligence.

Methods

SRF applies Symmetric Non-negative Matrix Factorization (SymNMF; (Kuang, Ding, & Park, 2012)) to decompose similarity matrices into interpretable latent components. Given a symmetric similarity matrix $S \in \mathbb{R}^{n \times n}$, SRF finds a low-rank factorization $S \approx WH^T$, where $W, H \in \mathbb{R}^{n \times k}$ are non-negative matrices and $k \ll n$. Following (Zhu, Li, Liu, & Li, 2018), we relax the strict symmetry constraint by allowing W and H to differ slightly, which improves optimization stability and flexibility. A regularization term encourages, but does not enforce, $W \approx H$, accounting for minor asymmetries that arise during optimization. The objective function becomes:

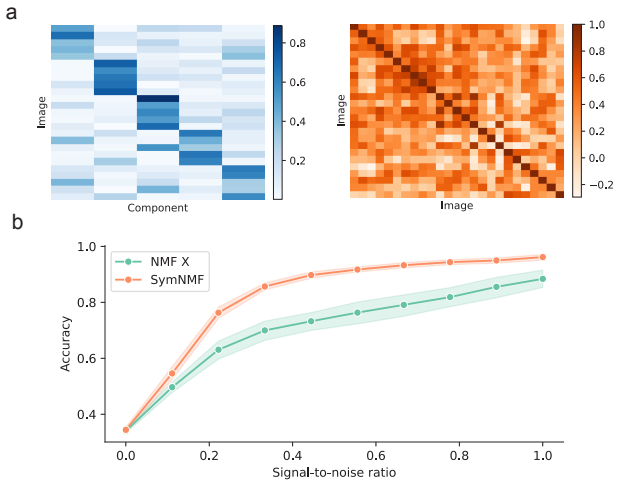

Figure 1: Simulation experiments. **a.** Ground-truth latent structure: each object is assigned to a unique cluster with minimal overlap. The resulting cosine similarity matrix reveals clear block structure reflecting the underlying clusters. **b.** SymNMF recovers the ground-truth similarity better compared to NMF.

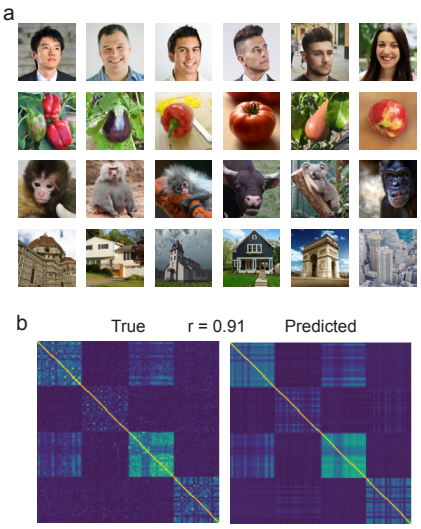

Figure 2: SRF on Behavioral Similarity **a.** SRF reveals interpretable components. Each row is a unique component and the top 6 images with highest numeric value are visualized. **b.** RSMs for predicted and true similarity.

$$\min_{W,H\geq 0} \|S - WH^T\|_F^2 + \alpha\|W - H\|_F^2, \tag{1}$$

where the first term ensures a low-rank approximation, and the second penalizes deviation from symmetry, weighted by α. Since Eq.1 is highly non-convex, we solve it iteratively using a two-block coordinate descent algorithm (Kim & Park, 2008).

Simulated Data Generation

We evaluated the clustering performance of SRF using synthetically generated data with known latent structure and controlled noise levels. Each dataset consisted of n samples and p features structured by k latent components. Sample-specific cluster memberships were drawn from a Dirichlet distribution, producing a non-negative matrix $\mathbf{M} \in \mathbb{R}^{n\times k}$. A latent feature matrix $\mathbf{X} \in \mathbb{R}^{k\times p}$ was sampled from a Gaussian distribution. These were combined to form the clean data matrix $\mathbf{D} = \mathbf{MX}$. We then simulated measured data by adding Gaussian noise scaled to the signal variance. Specifically, we computed the standard deviation of the clean signal $\mathbf{D}$, denoted as $\sigma = \text{std}(\mathbf{D})$, and generated noise $\boldsymbol{\eta} \sim \mathcal{N}(0, \sigma^2)$. The final data matrix $\mathbf{D}'$ was obtained via square-root mixing based on a predefined signal-to-noise ratio (SNR) parameter:

$$\mathbf{D}' = \sqrt{\text{SNR}} \cdot \mathbf{D} + \sqrt{1 - \text{SNR}} \cdot \boldsymbol{\eta},$$

where SNR is a scalar in the range $[0, 1]$. This procedure ensures that noise magnitude is matched to the variance of the underlying signal, while allowing us to systematically vary the contribution of noise across simulations. We then computed pairwise cosine similarities between the rows of $\mathbf{D}'$ to obtain a similarity matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$, with:

$$S_{ij} = \frac{\mathbf{d}'_i \cdot \mathbf{d}'_j}{\|\mathbf{d}'_i\| \, \|\mathbf{d}'_j\|},$$

and rescaled all entries of $\mathbf{S}$ to the range $[0, 1]$ via min-max normalization to ensure non-negativity.

Behavioral Similarity Matrix

We applied SRF to behavioral similarity matrices obtained from human participants performing a multi-arrangement task (Peterson, Abbott, & Griffiths, 2018). Participants arranged objects according to their perceived similarity, and pairwise object similarities were calculated from the aggregated participant arrangements, resulting in a symmetric behavioral similarity matrix.

Results and Conclusion

We first evaluated SRF on the synthetic data, testing its ability to recover known clustering structure. Clustering is closely linked to interpretability, as it groups data into discrete, often semantically meaningful categories, that can provide insight into the latent dimensions that organize the data. Each data point (row) had multiple cluster assignments with one dominant category (Figure 1a). The resulting cosine similarity matrix shows a distinct block pattern corresponding to the ground-truth cluster assignments. To quantify the effectiveness of SRF, we compared its cluster recovery performance to a conventional NMF approach applied directly to the noisy feature matrix. We systematically varied the noise level and measured how well each method recovered the original clusters. For this, we set the low-rank factorization to the same dimensionality as our original data rank ($k = 5$) and then eval-

uated the accuracy of predicting the dominant cluster category. As shown in Figure 1b, SRF was substantially more robust, maintaining better accuracy even at high noise levels.

Next, we applied SRF to empirical similarity data collected from human participants using a fixed rank of $k = 4$. As shown in Figure 2a, SRF uncovered interpretable latent dimensions that align with intuitive categorical and conceptual grouping of objects. Additionally, the low-dimensional embedding captured much of the ground-truth similarity (Figure 2b).

Together, our approach proposes a novel framework for the cognitive sciences to derive interpretable representations from similarity matrices, with broad applications for research in cognitive and computational neuroscience.

References

Kim, H., & Park, H. (2008). Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM journal on matrix analysis and applications*, *30*(2), 713–730.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 249.

Kuang, D., Ding, C., & Park, H. (2012). Symmetric nonnegative matrix factorization for graph clustering. In *Proceedings of the 2012 siam international conference on data mining* (pp. 106–117).

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fmri. *Neuroimage*, *56*(2), 400–410.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, *42*(8), 2648–2669.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11).

Zhu, Z., Li, X., Liu, K., & Li, Q. (2018). Dropping symmetry for fast symmetric nonnegative matrix factorization. *Advances in Neural Information Processing Systems*, *31*.