Duality of Bures and Shape Distances with Implications for Comparing Neural Representations

Sarah E. Harvey (sharvey@flatironinstitute.org)
Center for Computational Neuroscience, Flatiron Institute
New York, NY, 10010
Brett W. Larsen (brettlarsen@flatironinstitute.org)
Center for Computational Neuroscience, Flatiron Institute
New York, NY, 10010
Alex H. Williams (awilliams@flatironinstitute.org)
Center for Neural Science, New York University
New York, NY, 10003
Center for Computational Neuroscience, Flatiron Institute
New York, NY, 10010

Abstract

How should neuroscientists mathematically evaluate whether two individuals or networks have similar neural representations? A multitude of (dis)similarity measures between neural network representations have been proposed, resulting in a fragmented research landscape. Most of these measures fall into one of two categories. First, measures such as linear regression, canonical correlations analysis (CCA), and shape distances, all learn explicit mappings between neural units to quantify similarity while accounting for expected invariances. Second, measures such as representational similarity analysis (RSA), centered kernel alignment (CKA), and normalized Bures similarity (NBS) all quantify similarity in summary statistics, such as stimulus-by-stimulus kernel matrices, which are already invariant to expected symmetries. Here, we take steps towards unifying these two broad categories of methods by observing that the cosine of the Riemannian shape distance (from category 1) is equal to NBS (from category 2). We explore how this connection leads to new interpretations of shape distances and NBS, and draw contrasts of these measures with CKA, a popular similarity measure in the deep learning literature.

Keywords: representational similarity analysis; shape metrics, Bures distance.

Quantifying similarity between neural network representations is a well-recognized problem in computational neuroscience and deep learning (Klabunde et al., 2023; Sucholutsky et al., 2023). In neuroscience, measures of representational similarity have been used to benchmark models of biological systems (Kietzmann et al., 2019; Storrs et al., 2021), and to compare neural activity across different species (Kriegeskorte et al., 2008). In deep learning, they have been used to characterize learning dynamics (Morcos, Raghu, & Bengio, 2018), model robustness (Jones, Springer, Kenyon, & Moore, 2022), and the effects of changing model architecture (Maheswaranathan et al., 2019; Nguyen, Raghu, & Kornblith, 2021). Interest in this area has sparked a proliferation of measures to quantify representational (dis)similarity including: representational similarity analysis (RSA; (Kriegeskorte et al., 2008)), centered kernel alignment (CKA; (Kornblith et al., 2019)), generalized shape distances (Williams et al., 2021), canonical correlations analysis (CCA; (Raghu et al., 2017)), normalized Bures similarity (NBS; (Tang et al., 2020)), and the Riemannian covariance distance (Shahbazi et al., 2021). While all of these methods aim to quantify similar aspects of neural data, much more work is needed to characterize and understand the meaningful differences between these competing methods.

Here we develop a duality principle that links shape distances (Kendall et al., 2009; Williams et al., 2021) to well-known quantities in optimal transport (Malagò, Montrucchio, & Pistone, 2018) and quantum information theory (Nielsen & Chuang, 2000; Mendonça, et al., 2008; Watrous, 2018). We

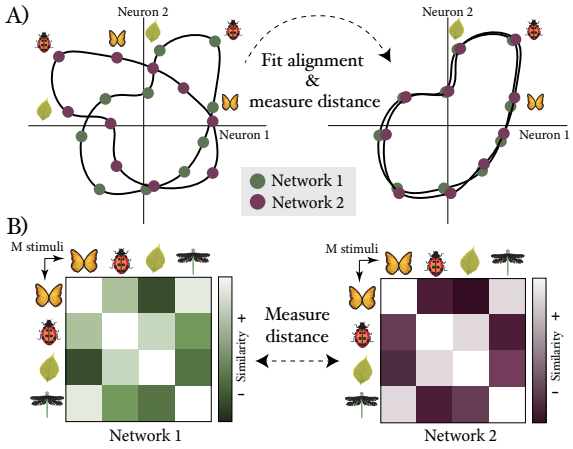

Figure 1: Two methods for measuring network similarity. (A) Methods that measure distance after aligning neural dimensions. (B) Methods that compare stimulus-by-stimulus representational similarity matrices.

can quantify similarity in neural representations using existing conventions (Klabunde et al., 2023). Let $f_x : \mathcal{Z} \mapsto \mathbb{R}^{N_x}$ and $f_y : \mathcal{Z} \mapsto \mathbb{R}^{N_y}$ be two neural networks that map inputs over a domain $\mathcal{Z}$ to high dimensional neural activation vectors (e.g. hidden layer activations or neuronal firing rates). Here, N_x and N_y denote the number of neurons in each network. How similar are the networks $f_x(\cdot)$ and $f_y(\cdot)$? That is, how similar are these functions over a collection of inputs $z_1, \ldots, z_M \in \mathcal{Z}$? We proceed by stacking neural responses $f_x(z_1) \ldots f_x(z_M)$ row-wise into a matrix $X \in \mathbb{R}^{M \times N_x}$. Likewise, we form a matrix $Y \in \mathbb{R}^{M \times N_y}$ from the second network's responses, $f_y(z_1) \ldots f_y(z_M)$. Lastly, we mean-center the columns of these matrices to remove overall translations in firing rate. One may view these matrices as approximations to each network's input-output mapping over a discrete set of M inputs. In general, $N_x \neq N_y$, but even if $N_x = N_y$, we should not expect the raw Euclidean distance, $\|X - Y\|_F$ to be meaningful since neurons are often indexed arbitrarily. Instead, we are interested in distances *that are invariant to a specified set of nuisance transformations* in the representations. For example, if we would like to ignore orthogonal transformations (including permutations of the neuron indices), we ought to develop distance functions for which $d(X,Y) = d(X,YQ)$ and also $d(X,XQ) = 0$ for any orthogonal matrix Q. This can be formalized by defining an equivalence relation between neural responses and defining a metric over the corresponding equivalence classes (Williams et al., 2021). Many measures of representational similarity either fit a nuisance transformation that aligns neural dimensions as well as possible (Fig. 1 A) or directly compare stimulus-by-stimulus (dis)similarities (Fig. 1 B), which are already invariant to certain transformations. The former encourages us to reason about geometric features in the space of neural activations, such as curvature or tangling of manifold structure which feature in theories of neural computation (Hénaff et al., 2021). The latter avoids

aligning neural axes, and connects to a rich literature in psychology that uses pairwise similarity judgements to interrogate the structure of cognition (Edelman, 1998).

From the first category, we consider *shape distances* (Kendall et al., 2009), which have been established in the computational neuroscience literature as a method to compare neural recordings across animals or brain regions (Williams et al., 2021). Assuming X and Y are both $M \times N$ matrices, we can define the angular distance: $\theta(X,Y) = \cos^{-1}\left(\mathrm{Tr}[X^\top Y]/\sqrt{\mathrm{Tr}[X^\top X]\mathrm{Tr}[Y^\top Y]}\right)$ which generalizes the elementary formula for the angle between two vectors. The *Riemannian shape distance* is the length of the shortest geodesic path between the two 'shapes' defined by X and Y, given by (Kendall et al., 2009): $\theta^*(X,Y) = \min \theta(X,YQ)$, minimized over nuisance transformations Q in the set of orthogonal matrices. Closely related to the Riemannian shape distance is a quantity called the *Procrustes size-and-shape distance* (Kendall et al., 2009) (or simply Procrustes distance): $\mathcal{P}(X,Y) = \min_{Q^\top Q=I} \|X - YQ\|_F$. An alternative from the second category begins by computing $M \times M$ (i.e. stimulus $\times$ stimulus) covariance matrices: $K_X = XX^\top$ and $K_Y = YY^\top$. We now measure similarity between the linear kernel matrices without any alignment (since the neural dimension has been removed, and the optimization problem above avoided). This is the basic idea behind RSA (Kriegeskorte et al., 2008), which is well-established in neuroimaging.

While these two approaches seem quite different, our main result shows that there exist cases where they coincide exactly, providing a theoretical bridge between the two classes of techniques. Specifically, there is a measure of distance between K_X and K_Y that equals the shape distance between X and Y. We provide a theorem stating that the Procrustes distance $\mathcal{P}(X,Y)$ is equivalent to the Bures distance between linear kernel matrices $\mathcal{B}(K_X,K_Y)$. Furthermore, the normalized Bures similarity (NBS) is equal to the cosine of the Riemannian shape distance θ^*. These results can also be generalized for nonlinear kernels.

Theorem 1. *Let K_X and K_Y be centered linear kernel matrices. Then, $\mathcal{B}(K_X,K_Y) = \mathcal{P}(X,Y)$ and $\mathrm{NBS}(K_X,K_Y) = \cos\theta^*(X,Y)$.*

The *normalized Bures similarity* (NBS) is a similarity measure that takes into account the geometry of the manifold of positive semidefinite (PSD) covariance matrices, and is defined as (Tang et al., 2020): $\mathrm{NBS}(K_X,K_Y) = \mathcal{F}(K_X,K_Y)/\sqrt{\mathrm{Tr}[K_X]\mathrm{Tr}[K_Y]}$, with $\mathcal{F}(K_X,K_Y) = \mathrm{Tr}[(K_X^{1/2}K_Y K_X^{1/2})^{1/2}]$. The quantity $\mathcal{F}(K_X,K_Y)$ is known as the *fidelity* and is used in quantum theory as a measure of quantum state distinguishability (Watrous, 2018). Analogous to the shape metrics case, the related distance is the *Bures distance* (Bhatia, Jain, & Lim, 2019): $\mathcal{B}(K_X,K_Y)^2 = \mathrm{Tr}[K_X] + \mathrm{Tr}[K_Y] - 2\mathcal{F}(K_X,K_Y)$.

One consequence of our work is a generalization of shape distances to the case where $N_x \neq N_y$. Furthermore, Thm. 1 allows us to draw on extensive literature to theoretically characterize shape/Bures distances. For example, it is well known that the Bures distance and $\arccos(\mathrm{NBS})$ on PSD matrices both satisfy the criteria of a metric space, including the triangle inequality. We conclude that the generalized definitions of Procrustes and Riemannian shape distance are also metrics, even though most classical work on shape theory does not consider datasets with unequal dimensions ($N_x \neq N_y$). Thm. 1 also makes clear that normalized shape and Bures distances converge to reasonable values when either $M \to \infty$ or $N \to \infty$.

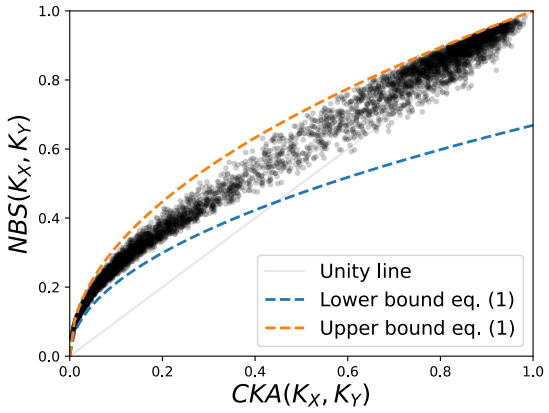

Figure 2: Comparing CKA and NBS. Points show similarity between pairs of 10×10 PSD matrices generated by sampling two Wishart distributions, of rank 1 and rank 5.

Our second set of results investigates the relationship between NBS and CKA (Kornblith et al., 2019), a popular approach in the deep learning literature which also compares stimulus-by-stimulus covariance matrices. From their definitions, one may guess that CKA is related to NBS (and to Riemannian shape distance by theorem 1). We show that CKA scores between networks can differ substantially (e.g. two- to three-fold) from NBS scores. We also derive upper and lower bounds that relate CKA and NBS in terms of the rank of the covariance matrices; and confirm their rather loose relationship. Setting $r(\cdot) = \mathrm{rank}(\cdot)$, we find:

$$\begin{aligned} \mathrm{CKA}(K_X,K_Y)/\sqrt{r(K_X)r(K_Y)} \;&\leq\; \mathrm{NBS}(K_X,K_Y)^2 \\ \mathrm{NBS}(K_X,K_Y)^2 \;&\leq \min(r(K_X),r(K_Y))\,\mathrm{CKA}(K_X,K_Y). \end{aligned} \tag{1}$$

Both of these bounds are equality when $r(K_X) = r(K_Y) = 1$. Fig. 2 shows that while NBS is bound to an envelope by CKA set by the matrix ranks, there is in general not a one-to-one relationship between them and the discrepancy between the two can be large compared with the range of $[0,1]$. We conclude that one should not expect CKA and NBS to behave similarly in practical scenarios.

Overall, our results demonstrate a theoretical equivalence between seemingly disparate methods, and an empirical divergence between superficially similar methods—motivating the need for more careful study into the theoretical landscape of representational similarity.

References

Bhatia, R., Jain, T., & Lim, Y. (2019). On the bures–wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, *37*(2), 165-191. doi: https://doi.org/10.1016/j.exmath.2018.01.002

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and brain sciences*, *21*(4), 449–467.

Hénaff, O. J., Bai, Y., Charlton, J. A., Nauhaus, I., Simoncelli, E. P., & Goris, R. L. (2021). Primary visual cortex straightens natural video trajectories. *Nature communications*, *12*(1), 5982.

Jones, H. T., Springer, J. M., Kenyon, G. T., & Moore, J. S. (2022, 01–05 Aug). If you've trained one you've trained them all: inter-architecture similarity increases with robustness. In J. Cussens & K. Zhang (Eds.), *Proceedings of the thirty-eighth conference on uncertainty in artificial intelligence* (Vol. 180, pp. 928–937). PMLR.

Kendall et al. (2009). *Shape and shape theory*. John Wiley & Sons.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116*(43), 21854-21863. doi: 10.1073/pnas.1905544116

Klabunde et al. (2023). *Similarity of neural network models: A survey of functional and representational measures.*

Kornblith et al. (2019, 09–15 Jun). Similarity of neural network representations revisited. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning* (Vol. 97, pp. 3519–3529). PMLR.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., ... Bandettini, P. A. (2008, December). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141.

Kriegeskorte et al. (2008, November). Representational similarity analysis - connecting the branches of systems neuroscience. *Front. Syst. Neurosci.*, *2*, 4.

Maheswaranathan et al. (2019). Universality and individuality in neural dynamics across large populations of recurrent networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc.

Malagò, L., Montrucchio, L., & Pistone, G. (2018). Wasserstein riemannian geometry of gaussian densities. *Information Geometry*, *1*(2), 137–179.

Mendonça, et al. (2008, Nov). Alternative fidelity measure between quantum states. *Phys. Rev. A*, *78*, 052330. doi: 10.1103/PhysRevA.78.052330

Morcos, A., Raghu, M., & Bengio, S. (2018). Insights on representational similarity in neural networks with canonical correlation. *Advances in neural information processing systems*, *31*.

Nguyen, T., Raghu, M., & Kornblith, S. (2021). Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. In *International conference on learning representations.*

Nielsen, M. A., & Chuang, I. L. (2000). *Quantum computation and quantum information*. Cambridge University Press.

Raghu et al. (2017). Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, *30*.

Shahbazi et al. (2021). Using distance on the riemannian manifold to compare representations in brain and in models. *NeuroImage*, *239*, 118271.

Storrs et al. (2021, 09). Diverse Deep Neural Networks All Predict Human Inferior Temporal Cortex Well, After Training and Fitting. *Journal of Cognitive Neuroscience*, *33*(10), 2044-2064. doi: 10.1162/jocn$_{a0}$1755

Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., ... others (2023). Getting aligned on representational alignment. *arXiv preprint:2310.13018*.

Tang et al. (2020). *Similarity of neural networks with gradients.*

Watrous, J. (2018). *The theory of quantum information*. Cambridge university press.

Williams et al. (2021). Generalized shape metrics on neural representations. In *Advances in neural information processing systems* (Vol. 34).