An Equivalence Between Representational Similarity Analysis and Centered Kernel Alignment

Alex H. Williams
Center for Neural Science, New York University
Center for Computational Neuroscience, Flatiron Institute

Abstract

Centered kernel alignment (CKA) and representational similarity analysis (RSA) of dissimilarity matrices are two popular methods for comparing neural systems in terms of representational geometry. Although they follow a conceptually similar approach, typical implementations of CKA and RSA tend to result in numerically different outcomes. Here, we show that these approaches become equivalent after incorporating a mean-centering step into RSA. This equivalence holds for both linear and nonlinear variants of these methods. By unifying these measures, this paper hopes to simplify a complex and fragmented literature on this subject.

Keywords: Neural Representations; Representational Similarity Analysis; Centered Kernel Alignment

Introduction

Quantifying similarity in neural representations (i.e. high-dimensional activation patterns) is a central interest of the cognitive computational neuroscience community (Barrett et al., 2019; Schrimpf et al., 2018). However, the methodologies used to quantify representational similarity are complex and diverse. For example, Klabunde et al. (2023) catalogued over thirty methods for quantifying similarity. It is difficult for practitioners to choose among this large menu of options, many of which give different numerical outputs (Soni et al., 2024).

It is important for the community to recognize cases where superficially distinct methods are, in fact, mathematically identical. For example, Harvey et al. (2024) showed that Procrustes shape distance scores (Ding et al., 2021; Williams et al., 2021) are, after a simple transformation, equivalent to Bures similarity scores (Tang et al., 2020). Here, we derive a similar result for Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008) and Centered Kernel Alignment (CKA; Cortes et al., 2012; Kornblith et al., 2019).

RSA is a mainstay of the cognitive computational neuroscience community that is rooted in work from psychology (Shepard & Chipman, 1970) and philosophy (Churchland, 1986). CKA is a more recent method developed in the deep learning community, which is also massively popular (>1600 citations at the time of writing). Many perceive RSA and CKA to be distinct methods because they are used by different research communities with limited cross-citation. Further, most implementations of RSA compare *representational distance matrices* (RDMs), while CKA involves comparing *kernel matrices*. Here, we will show that these differences are mostly superficial. Please note that an earlier version of this abstract appeared at a NeurIPS workshop in 2024.

Results

Assume that we are given two sets of neural responses from networks X and Y across M stimulus conditions. Let $x_1,\ldots,x_M \in \mathbb{R}^{N_X}$ denote the responses from system X and let $y_1,\ldots,y_M \in \mathbb{R}^{N_Y}$ denote the responses from system Y.

In RSA, we often use a distance function d to compute and compare RDMs. Let $D^X_{ij} = d(x_i,x_j)$ denote the RDM from system X and let $D^Y_{ij} = d(y_i,y_j)$ denote the RDM from system Y. A popular choice of distance function is the squared Euclidean distance, $d(x,x') = \|x-x'\|_2^2 = (x-x')^\top(x-x')$.

In CKA, we use a positive definite kernel function k to compute and compare kernel matrices. Let $K^X_{ij} = k(x_i,x_j)$ denote the kernel matrix from system X and let $K^Y_{ij} = k(y_i,y_j)$ denote the kernel matrix from system Y. Most commonly, practitioners use a linear kernel, $k(x,x') = x^\top x'$.

A similarity score between system X and system Y can be obtained by correlating the either RDMs or kernel matrices. A popular choice is the cosine similarity, written as:

$$S(A,B) = \frac{\mathrm{Tr}[AB]}{\|A\|_F\|B\|_F} = \frac{\mathrm{vec}(A)^\top \mathrm{vec}(B)}{\|\mathrm{vec}(A)\|_2\|\mathrm{vec}(B)\|_2} \quad (1)$$

for any two matrices A and B with appropriate dimensions. Some versions of RSA use other similarity functions, such as Spearman's rank correlation. These variants are difficult to mathematically analyze and we will not discuss them here.

Typical implementations of RSA compute the similarity between system X and system Y as $S(D^X, D^Y)$. In CKA, the similarity score is computed as $S(CK^XC, CK^YC)$ where C is a *centering matrix*, $C = I - \frac{1}{M}11^\top$. The matrices CK^XC and CK^YC are called *centered kernel matrices*, because matrix multiplying by C from the left and right has the effect of constraining the rows and columns to sum to zero.

In summary, typical implementations of RSA and CKA differ in two key respects. First, RSA uses RDMs while CKA uses kernel matrices. Second, RSA compares the raw RDMs while CKA compares the matrices after a centering operation. The punchline of this work is that the first difference is entirely superficial, and the two methods become equivalent if the centering operation is incorporated into RSA.

Main Proposition. *Let k be a positive definite kernel function associated with kernel matrices:*

$$K^X_{ij} = k(x_i,x_j) \quad \textit{and} \quad K^Y_{ij} = k(y_i,y_j) \quad (2)$$

Further, let D^X and D^Y be RDMs defined as:

$$D^X_{ij} = K^X_{ii} + K^X_{jj} - 2K^X_{ij} \quad \textit{and} \quad D^Y_{ij} = K^Y_{ii} + K^Y_{jj} - 2K^Y_{ij} \quad (3)$$

Then, the centered cosine similarity scores between these matrices agree:

$$S(CD^XC, CD^YC) = S(CK^XC, CK^YC) \tag{4}$$

This result can be proven by simple algebraic manipulations, which are not included here due to space constraints. Essentially identical results have been published by mathematicians studying different problems (e.g., Sejdinovic et al., 2013); however, the connection to RSA and CKA appears to be overlooked, or at least underappreciated, by the neuroscience and deep learning communities. To appreciate the significance of this result we state several implications below.

Corollary 1. *Linear CKA is equivalent to squared Euclidean distance RSA with centering.*

This follows from our main proposition since the squared Euclidean distance can be written $d(x,x') = x^\top x + x^\top x' - 2x^\top x'$. We see that this coincides with eq. (3) with a linear kernel function $k(x,x') = x^\top x'$. We note that linear CKA and squared Euclidean distance RSA are arguably the most popular variants of each respective method.

Corollary 2. *Nonlinear CKA is equivalent to a form of topological RSA with centering (Lin & Kriegeskorte, 2024).*

In topological RSA, a nonlinear function $\phi(\cdot)$, called the geo-topological transform, is applied elementwise to the RDM. The key attributes of $\phi(\cdot)$ are that (a) it monotonically increases and (b) it saturates at some maximal value (see **Fig. 1a**). Applying ϕ then has the effect of preserving distances close to zero and flattening out large distance scores, which accentuates the topological features of the neural representations (Lin & Kriegeskorte, 2024).

Let $\widetilde{D}^X = \phi(D^X)$ and $\widetilde{D}^Y = \phi(D^Y)$ denote the transformed RDMs. If we incorporate the centering operation into topological RSA, the similarity score becomes $S(C\widetilde{D}^XC, C\widetilde{D}^YC)$. Our main proposition implies that we can view this as a form of CKA if we can identify kernel matrices $\widetilde{K}^X$ and $\widetilde{K}^Y$ that satisfy:

$$\widetilde{D}^X_{ij} = \widetilde{K}^X_{ii} + \widetilde{K}^X_{jj} - 2\widetilde{K}^X_{ij} \quad \text{and} \quad \widetilde{D}^Y_{ij} = \widetilde{K}^Y_{ii} + \widetilde{K}^Y_{jj} - 2\widetilde{K}^Y_{ij} \tag{5}$$

It is easy to define nonlinear kernel functions $k(\cdot,\cdot)$ that satisfy this relationship. For example, Kornblith et al. (2019) studied the RBF kernel with lengthscale parameter $\ell > 0$:

$$k(x,x') = \exp(-\|x-x'\|_2^2/\ell) \tag{6}$$

Plugging this choice of kernel into eqns. 2 and 3 yields:

$$\widetilde{D}^X_{ij} = 2 - 2k(x_i,x_j) \quad \text{and} \quad \widetilde{D}^Y_{ij} = 2 - 2k(y_i,y_j) \tag{7}$$

The definition of the kernel in eq. 6 implies that the elements of $\widetilde{D}^X$ and $\widetilde{D}^Y$ are monotonically increasing and saturating functions of the distance, with the lengthscale parameter $\ell > 0$ controlling the steepness of the transformation (see **Fig. 1b**). Thus, CKA with nonlinear kernels—which is explored in works by, e.g., Alvarez (2022) and Kornblith et al. (2019)—is closely related to a recently proposed extension of RSA.

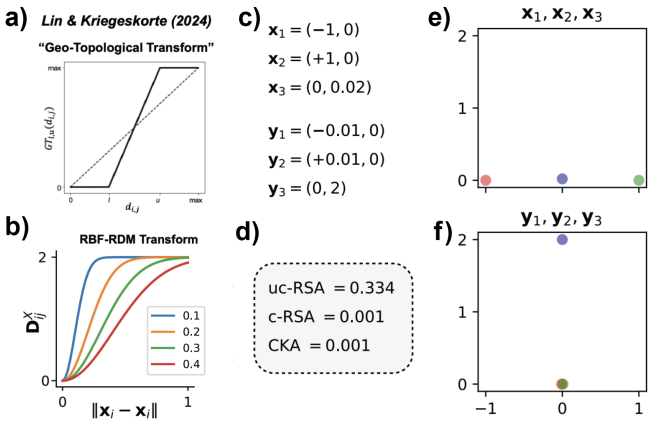

Figure 1: (a) Example geo-topological transform function, ϕ. (b) Effective geo-topological transform applied by nonlinear RBF CKA (colors correspond to different lengthscale parameters, ℓ). (c-f) Example 2D neural responses to 3 stimuli. Uncentered RSA (uc-RSA) is inflated above zero, while CKA and centered RSA (c-RSA) scores correctly indicate that the responses are maximally dissimilar.

Discussion

CKA and distance-based RSA are perhaps the two most popular approaches for quantifying similarity in neural population codes. We have shown that these are actually equivalent if one incorporates a mean centering transformation $D \mapsto CDC$ into RSA and uses a cosine similarity comparison criterion.

This begs the question: Is it a good idea to mean center RDMs in RSA analysis? We defer a complete analysis to future work, but remark that centering has the effect of rescaling RSA similarity scores to range from zero (least similar) to one (most similar). In **Fig. 1c-d** we enumerate a minimal example of 2D neural responses to 3 conditions, which yield a linear CKA score of zero and an uncentered-RSA score $\approx 1/3$. **Fig. 1e-f** visualizes these responses as three points in 2D space. In fact, one cannot arrange a set of 3 points in 2D space that make the uncentered RSA score go below $1/3$. Thus, without centering, RSA scores can be inflated above zero, which one might argue is undesirable. Incorporating centering "fixes" this problem and makes RSA with cosine RDM similarity precisely equivalent to CKA. However, further research into the desirability of centering is needed.

In conclusion, we have shown that two of the most influential frameworks for quantifying similarity in neural representations—CKA and RSA—are close to equivalent. This under-appreciated equivalence can greatly simplify comparisons of neural representations. For example, Cortes et al. (2012) derive error bounds on how many sampled stimuli, M, are needed to accurately estimate CKA. Our work shows how their mathematical analysis can be immediately applied to RSA. Likewise, statistical frameworks developed for RSA (e.g. Diedrichsen et al., 2021; Schütt et al., 2023) can be immediately adapted and applied to CKA-based analysis.

References

Alvarez, S. A. (2022). Gaussian rbf centered kernel alignment (cka) in the large-bandwidth limit. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(5), 6587–6593.

Barrett, D. G., Morcos, A. S., & Macke, J. H. (2019). Analyzing biological and artificial neural networks: Challenges with opportunities for synergy? *Current opinion in neurobiology*, *55*, 55–64.

Churchland, P. M. (1986). Some reductive strategies in cognitive neurobiology. *Mind*, *95*(379), 279–309. Retrieved March 24, 2025, from http://www.jstor.org/stable/2254072

Cortes, C., Mohri, M., & Rostamizadeh, A. (2012). Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, *13*, 795–828.

Diedrichsen, J., Berlot, E., Mur, M., Schütt, H. H., Shahbazi, M., & Kriegeskorte, N. (2021). Comparing representational geometries using whitened unbiased-distance-matrix similarity. *Neurons, Behavior, Data analysis, and Theory*, *5*(3), 1–31.

Ding, F., Denain, J.-S., & Steinhardt, J. (2021). Grounding representation similarity through statistical testing. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems* (pp. 1556–1568, Vol. 34). Curran Associates, Inc.

Harvey, S. E., Larsen, B. W., & Williams, A. H. (2024, 15 Dec). Duality of bures and shape distances with implications for comparing neural representations. In M. Fumero, E. Rodolá, C. Domine, F. Locatello, K. Dziugaite, & C. Mathilde (Eds.), *Proceedings of unireps: The first workshop on unifying representations in neural models* (pp. 11–26, Vol. 243). PMLR.

Klabunde, M., Schumacher, T., Strohmaier, M., & Lemmerich, F. (2023). Similarity of neural network models: A survey of functional and representational measures. *arXiv preprint arXiv:2305.06329*.

Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019, September). Similarity of neural network representations revisited. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning* (pp. 3519–3529, Vol. 97). PMLR.

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

Lin, B., & Kriegeskorte, N. (2024). The topology and geometry of neural representations. *Proceedings of the National Academy of Sciences*, *121*(42), e2317881121.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., et al. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007.

Schütt, H. H., Kipnis, A. D., Diedrichsen, J., & Kriegeskorte, N. (2023). Statistical inference on representational geometries (J. T. Serences & T. E. Behrens, Eds.). *eLife*, *12*, e82566.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., & Fukumizu, K. (2013). Equivalence of distance-based and rkhs-based statistics in hypothesis testing [Full publication date: October 2013]. *The Annals of Statistics*, *41*(5), 2263–2291. http://www.jstor.org/stable/23566550

Shepard, R. N., & Chipman, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive psychology*, *1*(1), 1–17.

Soni, A., Srivastava, S., Khosla, M., & Kording, K. P. (2024). Conclusions about neural network to brain alignment are profoundly impacted by the similarity measure. *bioRxiv*, 2024–08.

Tang, S., Maddox, W. J., Dickens, C., Diethe, T., & Damianou, A. (2020). Similarity of neural networks with gradients. *arXiv preprint arXiv:2003.11498*.

Williams, A. H., Kunz, E., Kornblith, S., & Linderman, S. (2021). Generalized shape metrics on neural representations. *Advances in Neural Information Processing Systems*, *34*, 4738–4750.