Neural variability: structure, sources, control, and data augmentation

Akash Umakantha

PhD Thesis
December 1, 2021

Joint Program in Neural Computation and Machine Learning
Neuroscience Institute; Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis committee:
Matthew Smith (CMU, co-chair)
Byron Yu (CMU, co-chair)
Rob Kass (CMU)
Bruno Averbeck (NIH/NIMH)

Submitted in partial fulfillment of the requirements for Doctor of Philosophy

Keywords: correlated neuronal variability, dimensionality reduction, multi-electrode recordings, brain computer interfaces, probabilistic graphical models, deep learning, data augmentation, convolutional neural network, vision transformer

Abstract

Variability is an important aspect of neural systems, both in the brain and in artificial networks. In the brain, neurons respond differently from trial to trial, even to repeated presentations of the exact same stimulus and this variability is often correlated across neurons. Previous work has posited that shared trial-to-trial variability (i.e., correlated neuronal variability) is behaviorally relevant and could have important implications for computations and information encoding. In the first three sections of this thesis, I present work to further the understanding of shared variability in the brain. To better understand the structure of shared variability, we related pairwise neuronal correlations to population dimensionality reduction methods. To investigate volitional control of shared variability in non-motor brain areas, we designed a brain computer interface for prefrontal cortex. Finally, to elucidate sources of variability, we developed a method called pCCA-FA to partition local (i.e., single brain area) and global (i.e., brain-wide) factors that contribute to shared variability. Variability also plays an important role in learning, in both the brain and in artificial neural networks (i.e., deep learning). Data augmentation increases the size, quality, and variability of datasets for improved training of deep learning models. In the final section, we empirically evaluated how different augmentation setups perform for different model architectures for image classification. We introduced a new augmentation, called StyleAug, which outperforms other state-of-the-art augmentations for training vision transformers (ViTs).

Overall, this dissertation furthers the understanding of variability in both natural and artificial neural systems. For artificial neural networks, this work highlights that one should consider different types of training data variability (i.e., augmentations) for different model architectures. For neuroscience, this work advances the understanding of the structure of shared neuronal variability, its distinct sources, and to what degree it can be controlled.

Acknowledgements

There are many people that I wish to thank for their support and contributions, both direct and indirect, to this dissertation.

Firstly, I would like to thank my advisors Matthew Smith and Byron Yu. During the early years of my PhD, they had the patience to show me the ropes of how to analyze data, perform experiments, and communicate my findings. As I developed, they provided me the freedom to pursue various ideas that I had, helping me get back on track when those ideas did not work out or helping me develop them further when they did. Not only were they great mentors, but they created a warm, fun, and supportive environment for doing a PhD. I would also like to thank my thesis committee members Rob Kass and Bruno Averbeck for providing feedback on my work, for challenging me to improve my statistical rigor, and for helping me to view my work from different perspectives and better understand how it fits into the broader literature.

Thank you to Joao Semedo, Alireza Golestaneh, and Wan-Yi Lin–my co-workers and mentors during my internship at the Bosch Center for Artificial Intelligence (Chapter 5). They provided invaluable training in best practices for deep learning and computer vision, and were supportive and enthusiastic of my various interests and ideas. I would also like to thank Filipe Condessa and Zico Kolter for their feedback on my work at Bosch.

Thank you to Braden Purcell and Thomas Palmeri–my mentors from my undergraduate institution Vanderbilt University. They introduced me to the world of academic research and sparked my interests in computational and systems neuroscience.

Thank you to Pittsburgh's wonderful scientific communities. For neuroscience, thank you to the Program in Neural Computation, Center for the Neural Basis of Cognition, Neuroscience Institute, and SCABBY and Brain Group journal clubs. Thank you to Carnegie Mellon's world class Machine Learning Department and community. I have had the privilege to meet, work with, and learn from some of the most brilliant minds in both neuroscience and machine learning. I would also like to thank Melissa Stupka and Beck Clark, coordinators for the Program in Neural Computation, and Diane Stidle, coordinator for the Machine Learning Department, for their incredible support and timely responses to my many questions.

I am deeply indebted to the many members of the Smith and Yu labs that I have the pleasure to work with over the years. Thank you to: Katerina Acar, Deepshikha Acharya, William Bishop, MeeDm Bossard, Benjamin Cowley, Matt Golub, Evren Gokcen, Matt Hall, Jay Hennig, Deepa Issar, Richard Johnston, Sanjeev Khanna, Chris Ki, Tze Hui Koh, Darby Losey, Megan McDonnell, Yuki Minai, Rudina Morina, Asma Motiwala, Emilio Salazar-Gatzimas, Joao Semedo, Joana Soldado-Magranger, Samantha Schmitt, Adam Snyder, Liz Spencer, Pati Stan, Hillary Wehry, Ryan Williamson, and Shenghao Wu. I specifically want to thank Samantha Schmitt for training and guidance in performing animal experiments. I would also like to thank Adam Synder for mentoring me on my first rotation project in the two groups. This project transformed into a collaboration with Ben Cowley and Rudina Morina and became my first published manuscript (Chapter 2). I would also like to thank Ryan Williamson–my peer, collaborator, and co-author on two projects that are dear to my heart (Chapters 3 and 4). Finally, I would like to thank Kendra Noneman, Chris Ki, and Megan McDonnell whom I have had the pleasure of mentoring. Having the chance to mentor these three has been an invaluable learning opportunity for myself, and one of the most deeply rewarding experiences of doing my PhD.

Thank you to all of the friends who have helped make Pittsburgh my home. You have been a constant source of support and memories throughout these years–from playing sports, to board game nights, to nights out on the town. And thank you to my old friends throughout the world who have been my vacation and adventure buddies, and who I always look forward to chatting with on Zoom. To all my dear friends, you have helped make the good times better and the bad times less bad.

Finally, I owe everything to my wonderful family. Thank you to our family pets, Aki and Buddy, for your playfulness and companionship especially during the difficult COVID pandemic. To my mother and father, thank you for inspiring me, for encouraging me to pursue my interests, and for being an unwavering source of love and support throughout my life.

Contents

List of Figures

1 Introduction

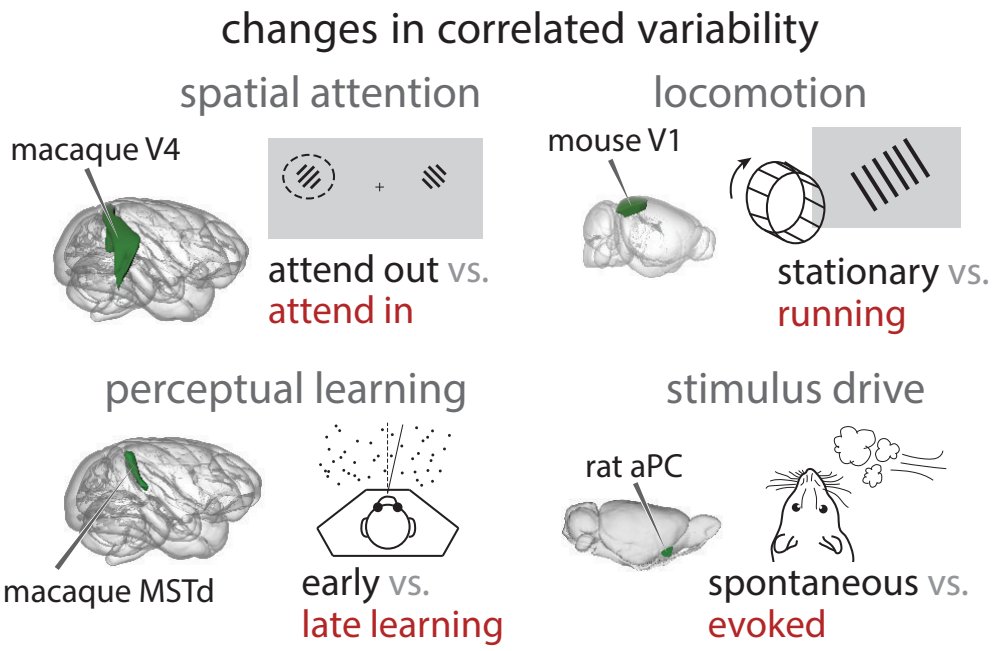


Figure 1: Correlated neuronal variability and changes in it have been widely observed. Four highlighted experiments in which correlated variability (e.g., r_{sc} mean) has been observed to change: spatial attention (macaque area V4 [1–3]), perceptual learning (macaque dorsal medial superior temporal area [4]), locomotion (mouse area V1 [5]), and stimulus drive (rat anterior piriform cortex [6]).

1.1 Neuronal variability in the brain

Neurons often respond differently even to repetitions of the same stimulus or task condition. These variable neuronal responses can be correlated across neurons from trial to trial, and is often measured using spike count correlations (r_{sc}, also referred to as noise correlation [7]). Correlated neuronal variability has been widely observed to change across the conditions of an experiment. For example, changes in spike count correlation have been observed with changes in attention [1, 2, 8–13], perceptual learning [4, 14], task difficulty [9], locomotion [5], stimulus drive [6, 15–19], decisions [20], task context [21], anesthesia [22], adaptation [23], and more (Fig. 1). Correlation also depend on properties of the neurons themselves, including their physical distance [17, 24–29], tuning properties [16, 24, 30, 31], time scales of activity [2, 16, 17, 32], and neuron type [3, 33].

A major reason that neuronal correlations have been the focus of many studies is that could have important implications for information coding and behavior. Because a neuronal responses are variable (i.e., "noisy"), it may be difficult to encode stimulus information using just a single neurons. Pooling across many neurons in a population should average out the noise and allow for better encoding. However, early work demonstrated that even small noise correlations can substantially limit the information encoded by a population of neurons [34, 35].

Subsequent work has noted that it is not only the amount or magnitude of these correlations, but importantly whether the noise is additive or multiplicative [36], and the alignment of noise correlations with signal correlations (i.e., how neurons covary with respect to different stimuli) [37, 38] that can impact information coding. When the noise and signal are aligned,

larger noise correlations limit information encoding; but when signal and noise are orthogonal, the presence of larger noise correlations can actually improve information encoding. Recently, experimental evidence has suggested the noise covariability can indeed interfere with signal and reduce information in the neuronal population [39, 40].

An improved understanding of the nature and characteristics of correlated neuronal variability will help elucidate how the brain encodes and processes information. In this dissertation, I present three research directions (summarized below) that advance our understanding of shared trial-to-trial variability in the brain. First, we bridged between pairwise correlations and dimensionality reduction to elucidate the structure of shared variability (Chapter 2). Second, we designed a brain computer interface (BCI) to investigate to what degree shared neuronal variability in prefrontal cortex (PFC) can be controlled (Chapter 3). Third, we recorded from PFC in both hemispheres of the brain and developed a model, called pCCA-FA, to partition the global and local sources of shared variability (Chapter 4).

Chapter 2 (structure): Bridging pairwise neuronal correlations and dimensionality reduction.
Two commonly used approaches to study interactions among neurons are spike count correlation, which describes pairs of neurons, and dimensionality reduction, applied to a population of neurons. Although both approaches have been used to study trial-to-trial neuronal variability correlated among neurons, they are often used in isolation and have not been directly related. In this section, we first established concrete mathematical and empirical relationships between pairwise correlation and metrics of population-wide covariability based on dimensionality reduction. Applying these insights to macaque V4 population recordings, we found that the previously reported decrease in mean pairwise correlation associated with attention stemmed from three distinct changes in population-wide covariability. Overall, our work builds the intuition and formalism to bridge between pairwise correlation and population-wide covariability and presents a cautionary tale about the inferences one can make about population activity by using a single statistic, whether it be mean pairwise correlation or dimensionality.

Chapter 2 is based on work that is available in a published article:
Umakantha A*, Cowley BR*, Morina R*. Snyder AC, Smith MA†, Yu BM† (2021). *Bridging neuronal correlations and dimensionality reduction.* Neuron, 109, 2740–2754.e12. (* and † denote equal contribution). DOI link. Simulation code. Code to compute metrics.

Chapter 3 (control): Stabilizing neuronal activity in prefrontal cortex using a brain computer interface.
Previous studies have shown that neuronal activity can drift slowly over time, and these slow drifts are thought to reflect slow changes in internal state (e.g., arousal, impulsivity, or engagement [41, 42]). We sought to assess to what degree these shared neuronal fluctuations were under volitional control in prefrontal cortex (PFC). We designed a novel brain computer interface (BCI) paradigm that required subjects to keep PFC neuronal activity close to the activity observed at the beginning of a session (i.e., the target activity). We showed that subjects were successfully able to use the BCI to reduce neuronal distance to the target. Furthermore, we found that neuronal activity drifted less on BCI trials than on non-BCI trials, demonstrating volitional control over PFC neuronal variability.

Chapter 3 is based on work that is part of a working manuscript:
Williamson RC*, Umakantha A*, Ki CS*, Smith MA†, Yu BM†. *Stabilizing neuronal activity in prefrontal cortex using a brain computer interface.* (* and † denote equal contribution).

Chapter 4 (sources): Local and global sources of coordinated neuronal variability in prefrontal cortex.
Previous work has shown that brain-wide signals (e.g., arousal or impulsivity [41]) contribute to how neurons co-fluctuate. In this section, we explore to what extent neuronal variability is shared across hemispheres (i.e., global) versus shared only within a brain area in one hemisphere (i.e., local), and the behavioral correlates of each type of variability. To ask this question, we simultaneously recorded from prefrontal cortex (PFC) in both hemispheres of the brain during a working memory task. We developed a probabilistic graphical model, called pCCA-FA, that allowed us to partition shared variability into across-hemisphere and within-hemisphere components. Surprisingly, we found that across-hemisphere shared variability was substantial, and often larger then within-hemisphere shared variability. Furthermore, the across-hemisphere latent neural activity was predictive of pupil size, which is thought to be associated with global cognitive phenomena such as arousal or wakefulness. Within-hemisphere latent activity did not predict pupil size. This suggests that across-hemispheres shared variability reflects global cognitive processes, while within-hemisphere shared variability might reflect local processes.

Chapter 4 is based on work that is part of a working manuscript:
Umakantha A*, Williamson RC*, Smith MA†, Yu BM†. *Coordinated variability of prefrontal cortex activity reflects global and local processes.* (* and † denote equal contribution).

1.2 Variability in artificial neural systems

Chapters 2-4 consider shared variability in natural neural systems (i.e., the brain); variability is also an important component of modern artificial neural networks (i.e., deep learning). Like in the brain, internal variability is important for deep learning both as a component of the models

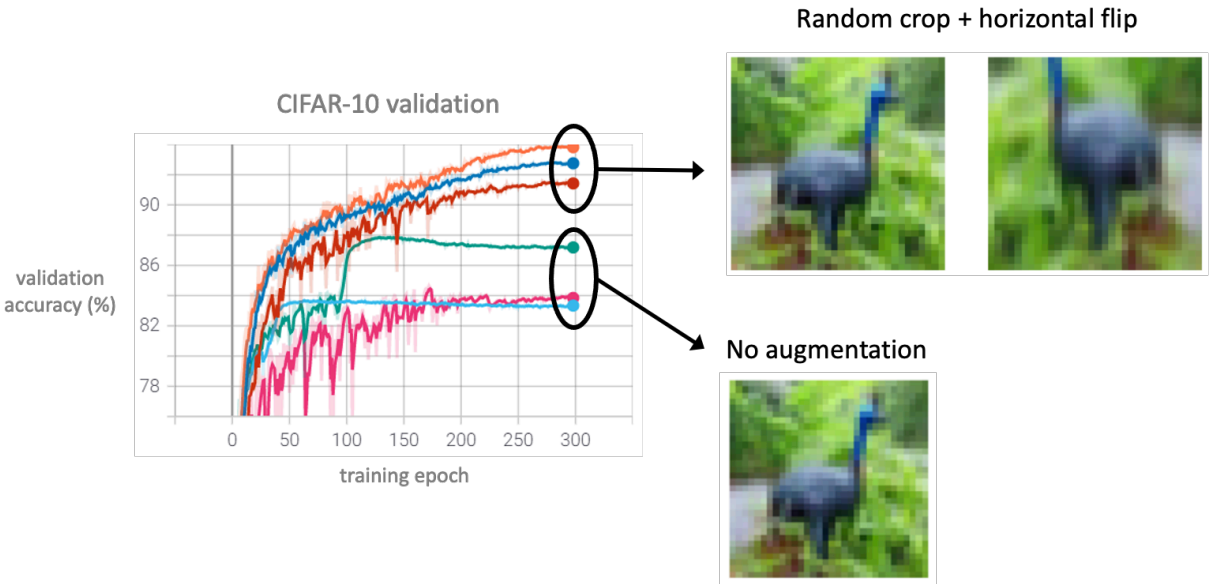


Figure 2: Data augmentation improves training of deep learning models. Six CNNs (ResNet-18) models with randomly initialized weights were trained on the CIFAR-10 dataset. Three models are trained on the raw image data without augmentation (no augmentation), while three models are trained with very basic image augmentations (random crop, random horizontal flip). The training curves for models trained with augmentation reach a higher level, are more stable, and do not asymptote.

themselves (e.g., stochastic generative models like variational autoencoders [43]), and also in regularization techniques when training models (e.g., dropout [44], stochastic depth [45]).

Another important source of variability is external to the deep learning models themselves. As humans, our brains/neural networks are constantly experiencing the external world, learning, and updating our beliefs and internal models (i.e., our synaptic weights or "parameters"). However, deep learning models can only learn the features and relationships in the dataset used to train them, limiting their robustness and generalization to unseen data. How can we increase the amount and variability of the training data that deep learning models learn from? One class of techniques is data augmentation–transformations of training data to increase the size, quality, and variability of datasets. Data augmentation plays a critical role in the learning of large, robust, and performant neural network models. (Fig. 2). However, the interaction between which data augmentation strategies work best for different model architectures is not known. In this section, we empirically evaluated different data augmentations and strategies for different deep learning architectures in the image classification task. Inspired by human visual perception, we also introduced a new data augmentation which outperforms other state-of-the-art augmentations for one of the model architectures.

Chapter 5 (data augmentation): How to augment your ViTs? Consistency loss and StyleAug, a random style transfer augmentation

The Vision Transformer (ViT) architecture has recently achieved competitive performance across a variety of computer vision tasks. One of the motivations behind ViTs is the use of weaker inductive biases, when compared to more traditional convolutional neural networks (CNNs), but this makes ViTs more difficult to train. They require very large training datasets, heavy regularization, and strong data augmentations. The data augmentation strategies used to train ViTs have largely been inherited from CNN training, despite the significant differences between the two architectures. In this work, we empirically evaluated how different data augmentation strategies performed on CNNs (e.g., ResNet) versus ViT architectures for image classification. We introduced a new data augmentation, called StyleAug, which performs style transfer from a training image to another randomly chosen image in the mini-batch. Combined with a consistency loss, StyleAug improves ViT validation accuracy, robustness to corruptions, shape bias, and transfer learning performance. We also found that, in addition to the classification loss, using a consistency loss between multiple augmentations of the same image was especially helpful when training ViTs.

Chapter 5 is based on work that will be available in an *arXiv* preprint:
Umakantha A, Semedo JD, Golestaneh SA, Lin WS. *How to augment your ViTs? Consistency loss and StyleAug, a random style transfer augmentation.*

2 [Structure] Bridging pairwise neuronal correlations and dimensionality reduction

The first studies of shared trial-to-trial neuronal variability often measured the Pearson correlation in spike counts between pairs of neurons, and typically only recorded from two or a few neurons simultaneously at a time. With recent advances in recording technology (e.g., Utah arrays and Neuropixel probes), it is possible to simultaneously record from tens or even hundreds of neurons. This has allowed the use of statistical techniques such as dimensionality reduction and graphical models to characterize neuronal population covariance structure. While both pairwise correlations and dimensionality reduction have been used to measure shared trial-to-trial neuronal variability, the relationship between the two has not been characterize. In this chapter, I present our work that bridges between the two perspectives and literatures to further our understanding of the structure of shared neuronal variability.

2.1 Introduction

Many studies of shared neuronal variability compute the average spike count correlation (r_{sc}, also known as noise correlation [7]) over pairs of recorded neurons for different experimental conditions, periods of time, neuron types, etc. A decrease in this mean correlation is commonly attributed to a reduction in the size (or gain) of shared co-fluctuations [35, 46–49], e.g., a decrease

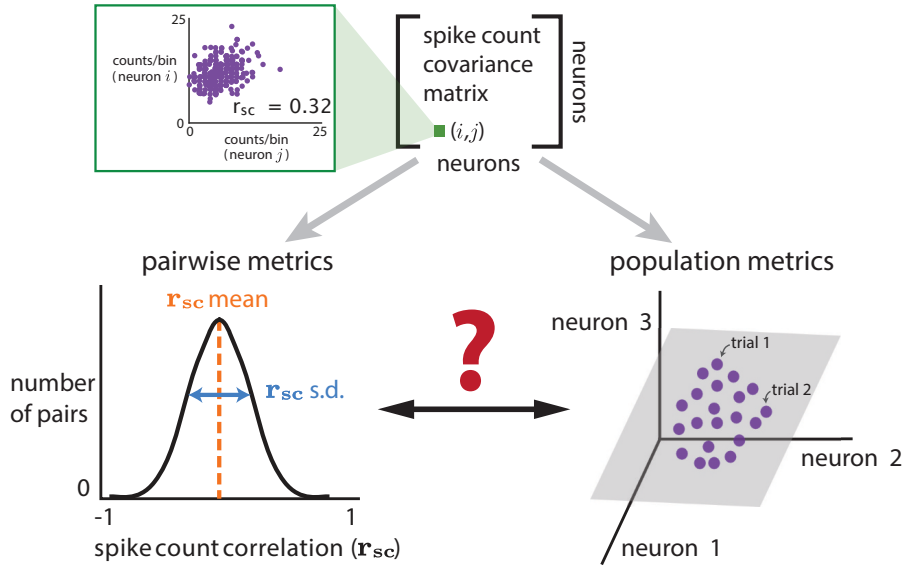


Figure 3: How do statistics computed on spike count correlations between pairs of neurons relate to how the entire population co-fluctuates? Pairwise (r_{sc}) and population (dimensionality reduction) metrics both arise from the same spike count covariance matrix, but the precise relationship between these two sets of metrics is not known. Top row: Each element of the spike count covariance matrix corresponds to the covariance across responses to repeated presentations of the same stimulus for two simultaneously-recorded neurons (e.g., neurons i and j, left inset). Bottom row: Pairwise metrics (left) typically summarize the distribution of spike count correlation with the mean (r_{sc} mean); in this work, we propose additionally reporting the standard deviation (r_{sc} s.d.). Population metrics (right) of the spike count covariance matrix are identified by applying dimensionality reduction to the population activity (e.g., gray plane depicts a low-dimensional space describing how neurons covary). By understanding the relationship between pairwise and population metrics, we can better interpret how changes in pairwise statistics correspond to changes in population metrics, and vice-versa.

in the strength of "common shared input" that drives each neuron in the population. However, other distinct changes at the level of the entire neuronal population can manifest as the same decrease in mean pairwise correlation. For example, a common input that drives the activity of all neurons up and down together could be altered to drive some neurons up and other neurons down. Alternatively, that first common input signal might remain the same, but a second input signal could be introduced that drives some neurons up and others down. It is difficult to differentiate these distinct possibilities using a single summary statistic, such as mean spike count correlation.

Distinguishing among these changes to the population-wide covariability might be possible by considering additional statistics that measure how the entire population of neurons co-fluctuates together. In particular, one may use dimensionality reduction to compute statistics that characterize multiple distinct features of population-wide covariability [50]. Dimensionality reduction has been used to investigate decision-making [51–54], motor control [55, 56], learning [14, 57, 58], sensory coding [59, 60], spatial attention [13, 46, 49, 61], interactions between brain areas [62–65], and network models [66–68], among others. As with mean spike count correlation, the statistics computed from dimensionality reduction can also change with attention [46, 49], stimulus drive [13, 69, 70], motor output [71], and anesthesia [26]. However, unlike mean spike count correlation (henceforth referred to as a "pairwise metric") which averages across pairs of neurons, the statistics computed from dimensionality reduction (henceforth referred to as "population metrics") consider the structure of population-wide covariability (Fig. 3). Although dimensionality reduction is often applied to trial-averaged activity (removing trial-to-trial variability), here we focus on using dimensionality reduction to study trial-to-trial variability (around the trial-averaged mean). An example of a commonly reported population metric is dimensionality [46, 56, 66–68, 70, 72, 73]. Dimensionality is used to assess whether the number of population co-fluctuation patterns (possibly reflecting the number of common inputs) changes across experimental conditions. Thus, population metrics could help to distinguish among the distinct ways in which population-wide covariability can change, especially those that lead to the same change in mean spike count correlation.

Both pairwise and population metrics aim to characterize how neurons covary, and both can be computed from the *same* spike count covariance matrix (Fig. 3). Still, studies rarely report both, and the relationship between the two is not known. In this study, we establish the relationship between pairwise metrics and population metrics both analytically and empirically using simulations. We find that changes in mean spike count correlation could correspond to several distinct changes in population metrics including: 1) the strength of shared variability (e.g., the strength of a common input), 2) whether neurons co-fluctuate together or in opposition (e.g., how similarly a common input drives each neuron in the population), or 3) the dimensionality (e.g., the number of common inputs). Furthermore, we show that a rarely-reported statistic–the standard deviation of spike count correlation–provides complementary information to the mean spike count correlation about how a population of neurons co-fluctuates. Applying this understanding to recordings in area V4 of macaque visual cortex, we found that the previously-reported decrease in mean spike count correlation with attention stems from multiple distinct changes in population-wide covariability. Overall, our results demonstrate that common ground exists between the literatures of spike count correlation and dimensionality reduction and provides a cautionary tale for attempting to draw conclusions about how a population of neurons covaries using one, or a small number of, statistics. Our framework builds the intuition and formalism to navigate between the two approaches, allowing for a more interpretable and richer description of the interactions among neurons.

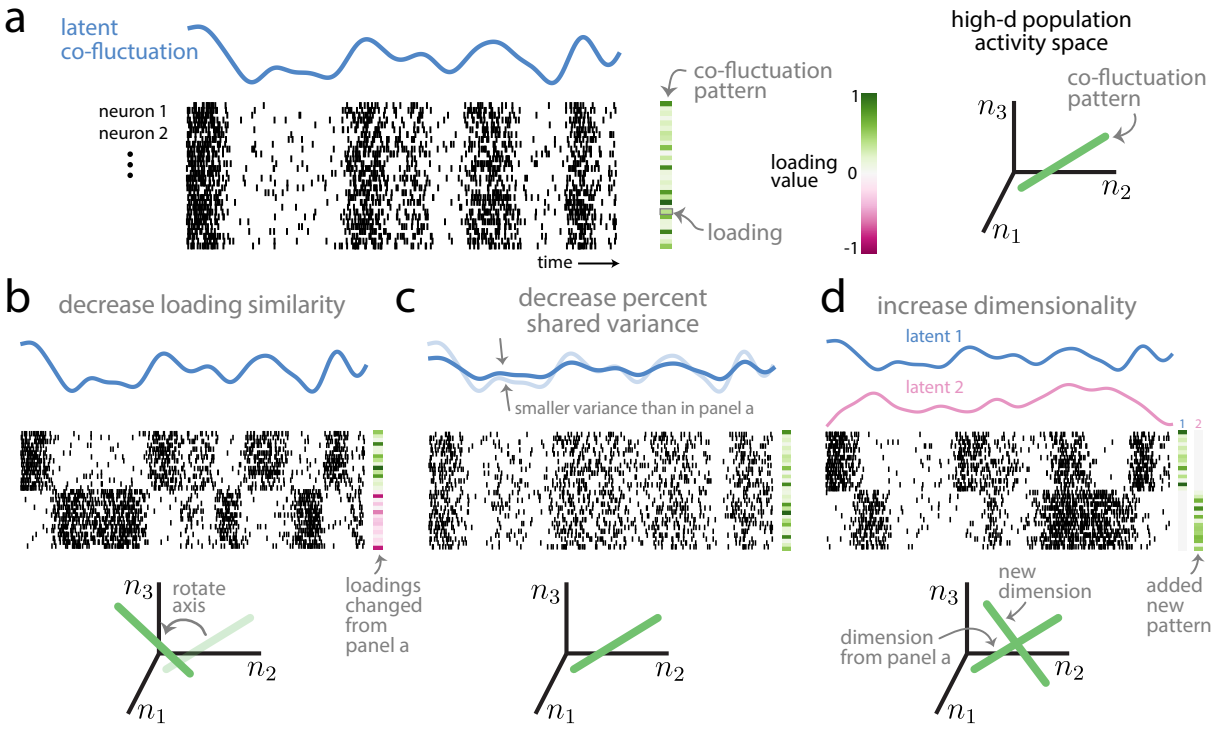


Figure 4: Intuition about population metrics. **a**. Population activity (population raster, where each row is the spike train for one neuron over time) is characterized by a latent co-fluctuation (blue) and a co-fluctuation pattern made up of loadings (green squares). Each neuron's underlying firing rate is a product of the latent and that neuron's loading (which may either be positive or negative). One may also view population activity through the lens of the population activity space (right plot), where each axis represents the activity of one neuron (n_1, n_2, n_3 represent neuron 1, neuron 2, and neuron 3). In this space, a co-fluctuation pattern corresponds to an axis whose orientation depends on the pattern's loadings (right plot, blue line). **b**. Population activity with a lower loading similarity than in panel **a**. The loadings have both positive and negative values (i.e., dissimilar loadings), leading to neurons that are anti-correlated (cf. top rows with bottom rows of population raster). Changing the loading similarity will rotate a pattern's axis in the population activity space (bottom plot, 'rotate axis'). **c**. Population activity with a lower %sv than in panel **a**. The strength of co-fluctuation is smaller than that in panel **a**. This leads to a lower %sv, as the latent co-fluctuation shows smaller amplitude changes over time. Changing %sv leads to no changes of the co-fluctuation pattern (bottom plot, axis is same as that in panel **a**). **d**. Population activity with a dimensionality of 2, compared to a dimensinality of 1 in panel **a**. Adding a new dimension leads to a new latent (orange line) and a new co-fluctuation pattern ('new dimension'). Each neuron's underlying firing rate is expressed as a weighted combination of the latents, where the weights correspond the neuron's loadings in each co-fluctuation pattern. Here, each dimension corresponds to a distinct subset of neurons (top rows vs. bottom rows); in general, this need not be the case, as each neuron typically has nonzero weights for both dimensions. In the population activity space (bottom plot), the activity varies along the two axes (i.e., a 2-d plane) defined by the two co-fluctuation patterns.

2.2 Defining pairwise and population metrics

We first define the metrics that we will use to summarize 1) the distribution of spike count correlations (i.e., pairwise metrics) and 2) dimensionality reduction of a population covariance matrix (i.e., population metrics). For pairwise metrics, we consider the mean and standard deviation (s.d.) of r_{sc} across all pairs of neurons, which summarize the r_{sc} distribution (Fig. 3, bottom left panel). For population metrics, we consider loading similarity, percent shared variance (abbreviated to %sv), and dimensionality (described below). These metrics each describe some aspect

of population-wide covariability and thus represent natural, multivariate extensions of r_{sc}.

To illustrate these three population metrics, consider the activity of a population of neurons over time (Fig. 4*a*, spike rasters). If the activity of all neurons goes up and down together, we would find the pairwise spike count correlations between all pairs of neurons to be positive. A more succinct way to characterize this population activity is to identify a single time-varying *latent co-fluctuation* that is shared by all neurons (Fig. 4*a*, blue line). The extent to which neurons are coupled to this latent co-fluctuation is indicated by a *loading* for each neuron. In this example, because the latent co-fluctuation describes each neuron's activity going up and down together, the loadings have the same sign (Fig. 4*a*, green rectangles). We refer to the latent co-fluctuation's corresponding set of loadings as a *co-fluctuation pattern*. A co-fluctuation pattern can be represented as a direction in the population activity space, where each coordinate axis corresponds to the activity of one neuron (Fig. 4*a*, right panel, green direction embedded in black coordinate axes).

The first population metric is *loading similarity*, a value between 0 and 1 that describes to what extent the loadings differ across neurons within a co-fluctuation pattern. A loading similarity close to 1 indicates that the loadings have the same sign and are of similar magnitude (Fig. 4*a*, green squares). A loading similarity close to 0 indicates that many of the loadings differ, either in magnitude, sign, or both (Fig. 4*b*, green and pink squares). In this case, some neurons may have positive loadings and co-fluctuate in the same direction as the latent (Fig. 4*b*, top rows of neurons show high firing rates when blue line is high and low firing rates when blue line is low), while other neurons may have negative loadings and co-fluctuate in opposition to the latent (Fig. 4*b*, bottom rows of neurons show low firing rates when blue line is high and high firing rates when blue line is low). One can view changing the loading similarity as rotating the direction of a co-fluctuation pattern in population activity space (Fig. 4*b*, bottom plot).

The second population metric is *percent shared variance* or %sv, which measures the percentage of spike count variance explained by the latent co-fluctuation. This percentage is computed per neuron, then averaged across all neurons in the population [66]. A %sv close to 100% indicates that the activity of each neuron is tightly coupled to the latent co-fluctuation, with a small portion of variance that is independent to each neuron (Fig. 4*a*). A %sv close to 0% indicates that neurons fluctuate almost independently of each other and their activity weakly adheres to the time course of the latent co-fluctuation (Fig. 4*c*). By changing %sv, one does not change the co-fluctuation pattern in population activity space (Fig. 4, green lines are the same in panels *a* and *c*) but rather the strength of the latent co-fluctuation (Fig. 4*c*, blue line has smaller amplitude than in panel *a*).

The third population metric is *dimensionality*. The variable activity of neurons may depend on multiple common inputs, e.g., top-down signals like attention and arousal [41, 46] or spontaneous and uninstructed behaviors [74, 75]–and these common inputs may differ in how they modulate neurons. This may result in two or more dimensions of the population activity (Fig. 4*d*, blue and orange latent co-fluctuations). For illustrative purposes, each dimension might correspond to a single group of tightly-coupled neurons (Fig. 4*d*, neurons in top rows have non-zero loadings for pattern 1, whereas neurons in bottom rows have non-zero loadings for pattern 2). However, in general, each neuron can have non-zero loadings for multiple patterns. In this work, we define dimensionality as the number of co-fluctuation patterns (or dimensions) needed to explain the shared variability among neurons. We use the term *dimension* to refer either to a latent co-fluctuation or its corresponding co-fluctuation pattern, depending on context. In population activity space, adding a new dimension adds a new axis along which neurons covary (Fig. 4*d*, green lines).

2.3 Varying population metrics to assess changes in pairwise metrics.

Given that both pairwise and population metrics are computed from the same spike count covariance matrix (Fig. 3), a relationship should exist between the two. We establish this relationship by deriving mathematical links and carrying out empirical simulations. In simulations, we assessed how systematically changing one of the population metrics (e.g., increasing loading similarity, Fig. 5*a*), changes the spike count covariance matrix (Fig. 5*b*), and the corresponding r_{sc} distribution (Fig. 5*c*), which we summarized using its mean and standard deviation (Fig. 5*d*). The covariance matrix was parameterized in a way that allowed us to create covariance matrices given a set of population metrics. Thus, our simulation procedure does not simulate neuronal activity, but rather creates covariance matrices which are consistent with the specified population metrics.

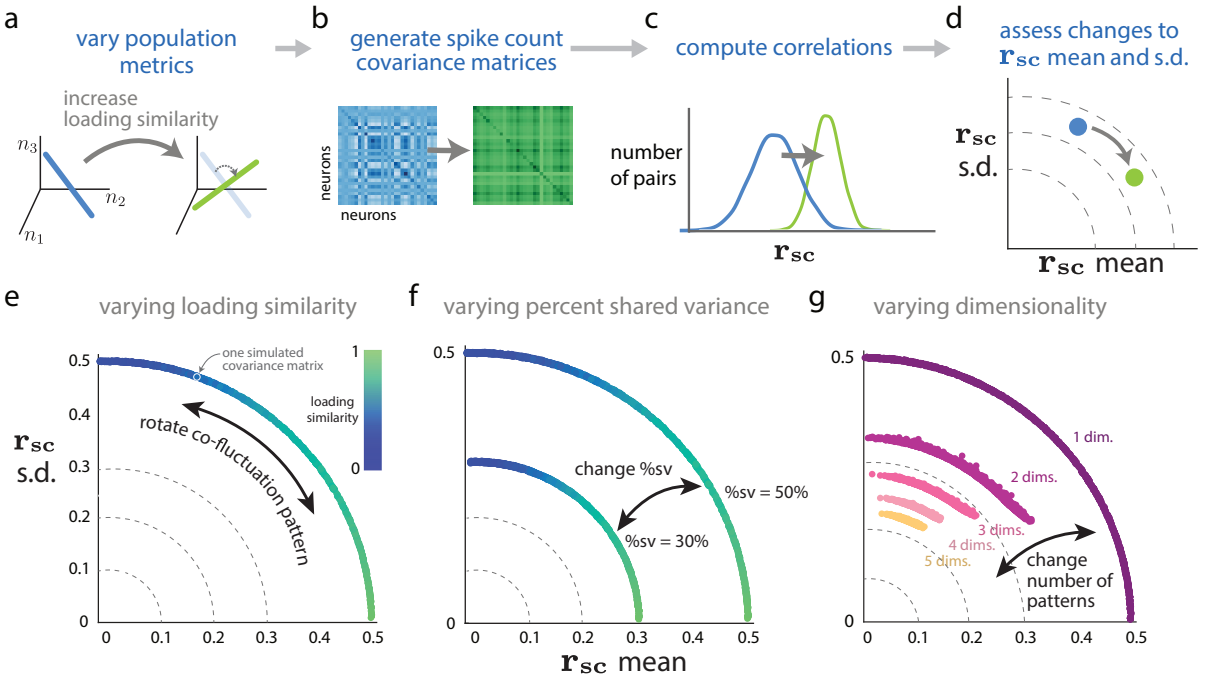


Figure 5: Relationship between population metrics and pairwise metrics. Panels **a-d** describe the simulation procedure to assess how systematic changes in population metrics lead to changes in pairwise metrics. **a**. We first systematically varied one of the population metrics while keeping the others fixed. For example, we can increase the loading similarity from a low value (left, blue) to a high value (right, green), while keeping %sv and dimensionality fixed. **b**. Then, we constructed covariance matrices corresponding to each value of the population metric in panel **a**, without generating synthetic data. **c**. For each covariance matrix from panel **b**, we directly computed the correlations (i.e., the r_{sc} distributions). **d**. We computed r_{sc} mean and r_{sc} s.d. from the r_{sc} distributions in panel **c** and then assessed how the change in a given population metric from panel **a** changed pairwise metrics. In this case, the increase in loading similarity increased r_{sc} mean and decreased r_{sc} s.d. (blue dot to green dot). **e**. Varying loading similarity with a fixed %sv of 50% and dimensionality of 1. Each dot corresponds to the r_{sc} mean and r_{sc} s.d. of one simulated covariance matrix with specified population metrics (dots are close together and appear to form a continuum). The color of each dot corresponds to the loading similarity, where a value of 1 indicates that all loading weights have the same value. **f**. Varying %sv. The same setting as in panel **e**, except we consider two different values of percent shared variance (50% and 30%). **g**. Varying dimensionality (i.e., number of co-fluctuation patterns) while sweeping loading similarity between 0 and 1 and keeping %sv fixed at 50%. In this simulation, the relative strengths of each dimension uniform across dimensions (i.e., flat eigenspectra).

Loading similarity has opposing effects on r_{sc} mean and s.d.

We first asked how the loading similarity of a single co-fluctuation pattern (i.e., one dimension) affected r_{sc} mean and s.d. Intuitively, a high loading similarity indicates that the activity of all neurons increases and decreases together (Fig. 4*a*), resulting in values of r_{sc} that are all positive and similar in value. Thus, r_{sc} mean would be large and positive and r_{sc} s.d. would be close to 0 (Fig. 5*e*, green dots near horizontal axis). On the other hand, a low loading similarity indicates that when some neurons increase their activity, others decrease their activity (Fig. 4*b*). Thus, r_{sc} values would be both positive (for pairs that change their activity in the same direction) and negative (for pairs that change their activity in opposition), resulting in large r_{sc} s.d. and r_{sc} mean close to 0 (Fig. 5*e*, blue dots near vertical axis). By gradually changing the loading similarity, we observed an arc-like trajectory in the r_{sc} mean versus r_{sc} s.d. plot (Fig. 5*e*). In Math Note A, we derive the analytical relationship between loading similarity and r_{sc}. In Math Note B, we show mathematically why the r_{sc} mean versus r_{sc} s.d. relationship follows a circular arc.

Decreasing %sv reduces r_{sc} mean and s.d.

We next asked how %sv, which measures the percentage of each neuron's variance that is shared with other neurons in the population, affected r_{sc} mean and s.d. In previous work, the r_{sc} mean is often interpreted as the amount of shared variability in a population of neurons [7]. In simulations, we found that r_{sc} mean and %sv were closely linked when loading similarity was high, but were unrelated when loading similarity was low. For example, when loading similarity was high and %sv decreased from 50% to 30%, we observed a proportionally-sized decrease in r_{sc} mean from 0.5 to 0.3 (Fig. 5*f*, green dots from outer arc to inner arc). On the other hand, when loading similarity was low and %sv decreased from 50% to 30%, r_{sc} mean changed very little and remained close to 0 (Fig. 5*f*, blue dots from outer arc to inner arc). Importantly, this illustrates that r_{sc} mean and %sv are not the same—it is possible for a population of neurons with high %sv to have smaller r_{sc} mean than a population with lower %sv (Fig. 5*f*, blue dots in outer arc have smaller r_{sc} mean than green dots in inner arc).

To understand this further, we derived the precise mathematical relationship between r_{sc}, %sv, and the loadings for a pair of neurons (Math Note A):

$$\rho_{ij} = \sqrt{\phi_i \phi_j}\ \mathrm{sign}(w_i w_j) \tag{1}$$

where ρ_{ij} is the r_{sc} between neurons i and j, ϕ_i is the %sv of neuron i (expressed as a proportion), and w_i is the loading of neuron i in the co-fluctuation pattern. Equation (1) shows that ρ_{ij} depends on %sv, but is also influenced by loading similarity. If all loadings have the same sign (i.e., loading similarity is high), then $\mathrm{sign}(w_i w_j)$ is always +1, and $\rho_{ij} = \sqrt{\phi_i \phi_j}$. In this case, r_{sc} mean (the average across all ρ_{ij}) is a good representation of %sv. However, if many loadings have opposite signs (i.e., low loading similarity), then some $\mathrm{sign}(w_i w_j)$ will be +1 and others will be -1. Even if %sv (and thus $|\rho_{ij}|$) is large, many correlations will have opposite signs, and averaging over them results in r_{sc} mean close to 0. In this case, r_{sc} mean is not a good representation of %sv.

More precisely, the %sv corresponds to the magnitude of r_{sc} values (i.e., each $|\rho_{ij}|$), as opposed to the r_{sc} mean. When loading similarity is low and %sv decreases, each $|\rho_{ij}|$ still becomes smaller—positive correlations become less positive and negative correlations become less negative. However, the reduction in %sv is not reflected by r_{sc} mean, but rather by a decrease in r_{sc} s.d. (Fig. 5*f*, blue dots from outer arc to inner arc). More generally, by considering *both* r_{sc} mean and s.d. together, we observed that reducing the %sv decreased the distance to the origin in the the r_{sc} mean versus r_{sc} s.d. plot (Fig. 5*f*, arc for %sv=30% closer to origin than arc for

%sv=50%). We showed mathematically that the %sv population metric can be estimated using the distance of pairwise metrics from the origin (Math Note B):

$$\%\text{sv} \approx \sqrt{(r_{\text{sc}} \text{ mean})^2 + (r_{\text{sc}} \text{ s.d.})^2}$$

These findings highlight the pitfalls of considering a single statistic (e.g., r_{sc} mean) on its own and the benefits of considering multiple statistics (e.g., both r_{sc} mean and s.d.) when trying to draw conclusions about how neurons covary. By considering r_{sc} mean and s.d. together, one can insight into the loading similarity (Fig. 5*e*) and the %sv (Fig. 5*f*) of a neuronal population. Thus far, we have only considered the specific case where activity co-fluctuates along a single dimension in the firing rate space. We next considered how pairwise metrics change in the more general case where neuronal activity co-fluctuates along multiple dimensions.

Adding more dimensions tends to reduce r_{sc} mean and s.d.

We sought to assess how dimensionality (i.e., the number of co-fluctuation patterns) is related to pairwise metrics. In simulations, we increased the number of co-fluctuation patterns (compare Fig. 4*a* to *d*; see Methods), while sweeping loading similarity and fixing the total %sv. We found that increasing dimensionality tended to reduce r_{sc} mean and s.d. (Fig. 5*g*, dots for larger dimensionalities lay closer to the origin than dots for smaller dimensionalities).

It seems counterintuitive that adding a new way in which neurons covary reduces the magnitude of r_{sc}. The intuition is that if multiple distinct (i.e., orthogonal) dimensions exist, then a neuron pair interacts in opposing ways along different dimensions. For example, consider two neurons with loadings of the same sign in one co-fluctuation pattern, and opposite sign in the second pattern. If only the first dimension exists, the two neurons would go up and down together and be positively correlated. If only the second dimension exists, the two neurons would co-fluctuate in opposition and be negatively correlated. When both dimensions exist, the positive correlation from the first dimension and the negative correlation from the second dimension offset, and the resulting correlation between the neurons would be smaller than if only the first dimension were present. We formalize the above intuition in Math Note C. We also show analytically that increasing dimensionality tends to move points closer to the origin in the r_{sc} mean versus r_{sc} s.d. plot (i.e., decrease r_{sc} mean and s.d.; Math Note D).

An increase in dimensionality does not imply that *both* r_{sc} mean and r_{sc} s.d. necessarily decrease. For example, in the case where the first dimension has high loading similarity, adding more dimensions means it is less likely for r_{sc} s.d. to be 0 (Fig. 5*g*, compare dot closest to horizontal axis for '1 dim.' to that for '2 dims.'). The intuition is that if the first dimension has a loading similarity of 1, the loading weights for all neurons are the same and thus r_{sc} values between all pairs are the same, resulting in r_{sc} s.d. of 0. Adding an orthogonal dimension to this pattern necessarily means adding a pattern with low loading similarity (Math Note E), making it less likely for r_{sc} across all pairs to be the same. Therefore, r_{sc} s.d. is unlikely to be 0 for two dimensions (Fig. 5*g*, the smallest r_{sc} s.d. for '2 dims.' is around 0.2). Still, in Figure 5*g* the dots for '2 dims.' are closer to the origin than the dots for '1 dim', implying that even if r_{sc} s.d. increases with an increase in dimensionality, the r_{sc} mean must decrease to a larger extent (Math Note D). As another example, in the case where the first dimension has low loading similarity, adding a second dimension with high loading similarity would increase r_{sc} mean. The r_{sc} s.d. would decrease to a larger extent than the increase in r_{sc} mean such that the dot for two dimensions is closer to the origin than that for one dimension (Math Note D).

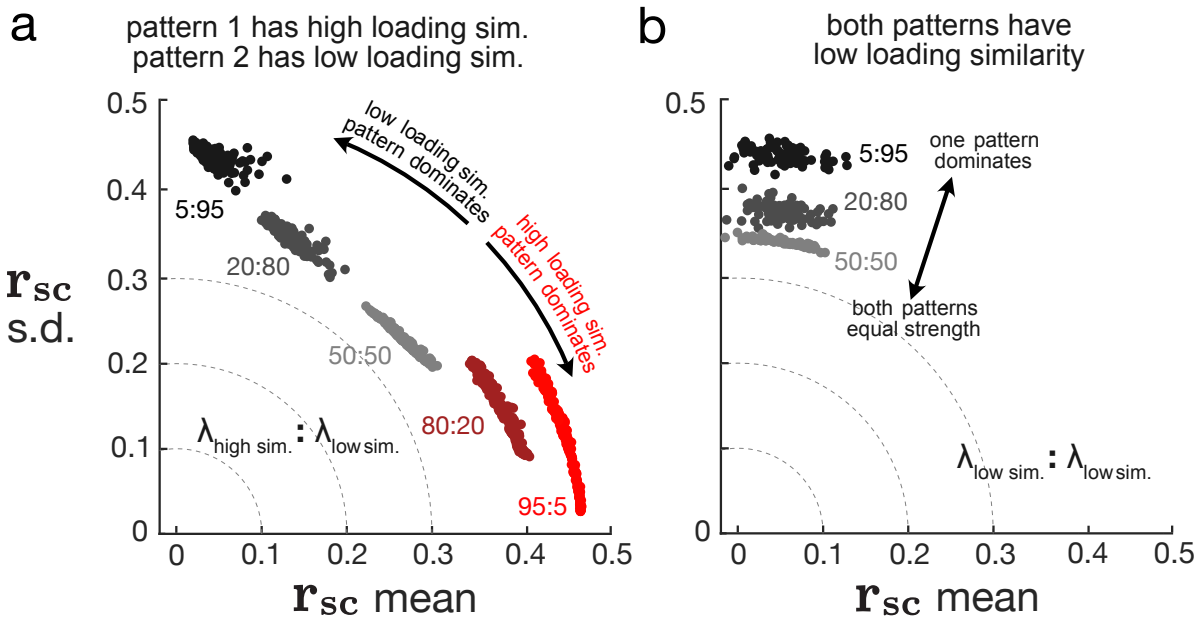


Figure 6: Relative strengths of dimensions affect r_{sc} distributions. With dimensionality of 2, we systematically varied the relative strengths of the two dimensions with a fixed total %sv of 50%. We considered two scenarios: 1) one dimension has high loading similarity and the other dimension has low loading similarity (panel **a**) and 2) both dimensions have low loading similarity (panel **b**). Each dot represents one simulated covariance matrix and r_{sc} distribution. The color of the dots indicate different relative strengths between the two dimensions, and numbers next to each cloud of dots indicate the ratio between the relative strength associated with each dimension. For example, in panel **a**, red dots correspond to the high loading similarity dimension being 19 times stronger (95:5) than the low loading similarity dimension. Black dots correspond to the low loading similarity dimension being 19 times stronger (5:95) than the high loading similarity dimension. In panel **b**, since both patterns have low loading similarity, clouds for 80:20 and 95:5 are very similar to clouds for 20:80 and 5:95 respectively and are thus omitted for clarity. See also Fig S1.

The relative strength of each dimension impacts pairwise metrics.

In the previous simulation (Fig. 5*g*), we assumed that each dimension explained an equal proportion of the overall shared variance (e.g., for two dimensions, each dimension explained half of the shared variance; see Methods). However, it is typically the case for recorded neuronal activity that some dimensions explain more shared variance than others; in other words, neuronal activity co-fluctuates more strongly along some patterns than others [49, 57, 66, 67, 71, 73, 76]. We sought to assess the influence of the relative strength of each dimension on pairwise metrics.

We reasoned that stronger dimensions would play a larger role than weaker dimensions in determining the r_{sc} distribution and pairwise metrics. Extending equation (1) to multiple dimensions, we show that the r_{sc} between a pair of neurons can be expressed as the sum of a contribution from each constituent dimension (Math Note C). The stronger a dimension, the larger the magnitude of its contribution to r_{sc}, and thus the larger its impact on r_{sc} mean and s.d.

To test this empirically, we performed a simulation with two dimensions, while systematically varying the relative strength of each dimension. We considered two scenarios: (1) one dimension has a pattern with high loading similarity and one dimension has a pattern with low loading similarity (Fig. 6*a*), and (2) both dimensions have patterns with low loading similarity (Fig. 6*b*). Note that both dimensions cannot have patterns with high loading similarity because they would not be orthogonal (Math Note E).

In scenario (1) where one dimension's pattern has high loading similarity and the other has low loading similarity, r_{sc} mean and r_{sc} s.d. reflects the loading similarity of the dominant dimension (Fig. 6*a*). When the dimension with a high loading similarity pattern dominated, r_{sc} mean was large and r_{sc} s.d. was small (Fig. 6*a*, red dots are close to horizontal axis). When the dimension with a low loading similarity pattern dominated, r_{sc} mean was small and r_{sc} s.d. was large (Fig. 6*a*, black dots are close to vertical axis). When the two dimensions were of equal strength (i.e., neither dimension dominated), r_{sc} mean and r_{sc} s.d. were both intermediate values (Fig. 6*a*, light gray dots are between red and black dots). Thus, the dimensions along which neuronal activity co-fluctuates more strongly have a greater influence on pairwise metrics (Supplementary Fig. 1).

In scenario (2) where both dimensions have patterns of low loading similarity, r_{sc} mean was low and r_{sc} s.d. was high (Fig. 6*b*), similar to when there is one dimension with low loading similarity (Fig. 5*e*, blue dots). When we made one dimension stronger than the other, r_{sc} mean remained low and r_{sc} s.d. remained high (Fig. 6*b*, light gray dots and black dots are both close to vertical axis) because both patterns had low loading similarity. However, the radius of the arc increased (Fig. 6*b*, black dots farther from the origin than light gray dots), and was close to the arc that would have been produced with a single dimension (Fig. 5*g*, '1 dim.'). Thus, whereas changing the number of dimensions causes discrete jumps in the arc radius (Fig. 5*g*), changing the relative strength of each dimension allows for r_{sc} mean and r_{sc} s.d. to vary continuously between the arcs for different dimensionalities. Put another way, changing the relative strength of each dimension varies the "effective dimensionality" of population activity in a continuous manner. Neuronal activity for which one dimension dominates another (Fig. 6*b*, black dots) has a lower effective dimensionality than when both dimensions have equal strength (Fig. 6*b*, light gray dots).

2.4 Reporting only a single statistic provides an incomplete description of population covariability

Figure 7 summarizes the relationships that we have established between pairwise metrics and population metrics. Rotating a co-fluctuation pattern from a low loading similarity to a high loading similarity increases r_{sc} mean and decreases r_{sc} s.d. along an arc (Fig. 7, arrow outside pink arc). Decreasing %sv decreases both r_{sc} mean and s.d. (Fig. 7, arrow pointing toward origin), and increasing dimensionality also tends to decrease r_{sc} mean and s.d. (Fig. 7, pink to yellow shaded regions).

These results provide a cautionary tale that using a single statistic on its own provides an opaque description of population-wide covariability. For example, a change in r_{sc} mean could correspond to changes in loading similarity, %sv, dimensionality, or a combination of the three. Likewise, reporting dimensionality on its own would be incomplete because the role of a dimension in explaining population-wide covariability depends how much shared variance it explains and the loading similarity of its co-fluctuation pattern. For example, consider a decrease in dimensionality by 1. This would have little impact on population-wide covariability if the removed dimension explains only a small amount of shared variance, whereas it could have a large impact if the removed dimension explains a large amount of shared variance.

Considering multiple statistics together provides a richer description of population-wide covariability. For example, in the case where population activity co-fluctuates along a single dimension, r_{sc} mean and r_{sc} s.d. can be used together to approximate %sv (using distance from the origin) and deduce whether loading similarity is low (r_{sc} s.d. $>$ r_{sc} mean) or high (r_{sc} mean $>$ r_{sc} s.d.), whereas r_{sc} mean alone would not provide much information about %sv or loading similarity (cf. Fig. 7). In the next section, we further demonstrate using neuronal recordings how relating pairwise and population metrics using the framework we have developed (Fig. 7)

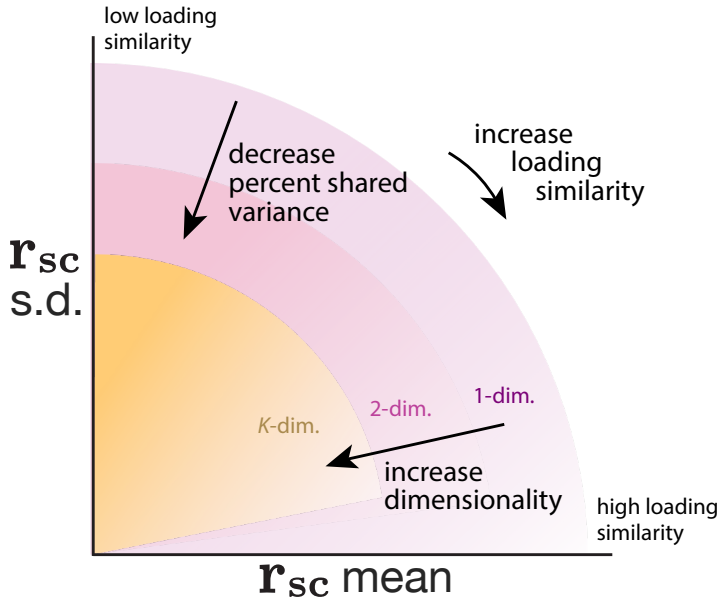


Figure 7: Summary of relationship between pairwise and population metrics. A change in r_{sc} mean and r_{sc} s.d. may correspond to changes in loading similarity, %sv, dimensionality, or a combination of the three. Shaded regions indicate the possible r_{sc} mean and r_{sc} s.d. values for different dimensionalities; increasing dimensionality tends to decrease r_{sc} mean and r_{sc} s.d. (shaded regions for larger dimensionalities become smaller). Within each shaded region, decreasing %sv decreases both r_{sc} mean and s.d. radially toward the origin. Finally, rotating co-fluctuation patterns such that the loadings are more similar (going from low to high loading similarity) results in moving clockwise along an arc such that r_{sc} mean increases and r_{sc} s.d. decreases. We also note two subtle trends. First, there are more possibilities for loading similarity to be low than high (Appendix E), suggesting that r_{sc} s.d. will generally tend to be larger than r_{sc} mean if neuronal activity varied along a randomly chosen co-fluctuation pattern (shading within each region is darker near the vertical axis than the horizontal axis). Second, this effect becomes exaggerated for higher-dimensional neuronal activity as many dimensions can have low loading similarity but only one dimension can have high loading similarity (Appendix E). Thus, it becomes progressively unlikely for r_{sc} s.d. to be 0 as dimensionality increases (shaded regions for larger dimensionalities lifted off the horizontal axis).

provides a richer description of how neurons covary than using a single statistic (e.g., r_{sc} mean) alone.

2.5 Case study: V4 neuronal recordings during spatial attention

When spatial attention is directed to the receptive fields of neurons in area V4 of macaque visual cortex, r_{sc} mean among those neurons decreases [1–3, 10, 77]. This decrease has often been attributed to a reduction in shared modulations among the neurons. However, we have shown both mathematically and in simulations that several distinct changes in population metrics (e.g., decrease in loading similarity, decrease in %sv, or an increase in dimensionality) could underlie this decrease in r_{sc} mean. Here, we sought to assess which aspects of population-wide covariability underlie, and how each of them contribute to, the overall decrease in r_{sc} mean.

We analyzed activity recorded simultaneously from tens of neurons in macaque V4 while the animal performed an orientation-change detection task [Fig. 8*a*; previously reported in 13]. To probe spatial attention, we cued the animal to the location of the stimulus that was more likely to change in orientation. As expected, perceptual sensitivity increased for orientation changes in the cued stimulus location (Fig. 8*a* inset, red dot above black dot). 'Attend-in' trials were

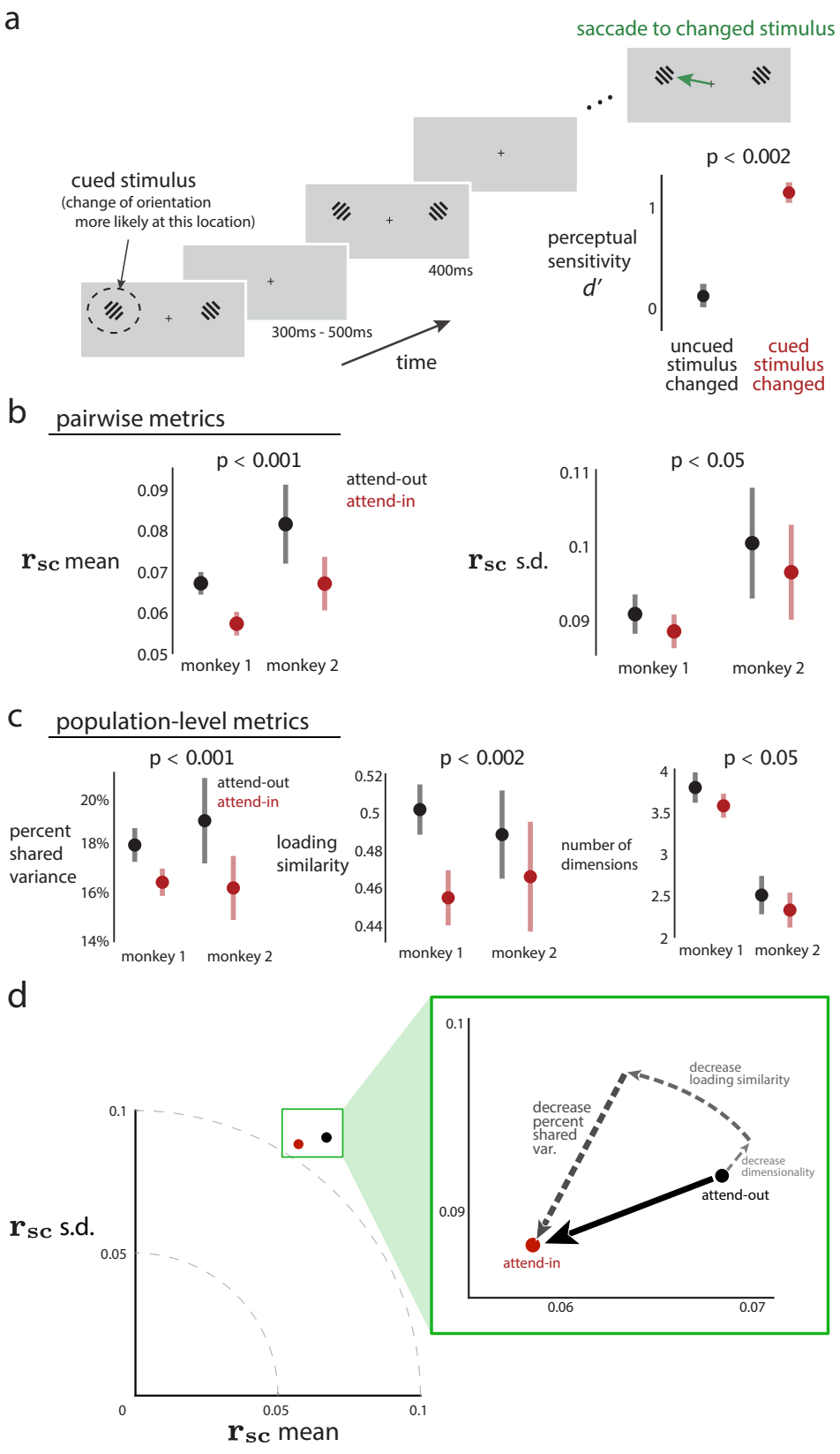

a
saccade to changed stimulus
cued stimulus
(change of orientation
more likely at this location)
400ms
300ms - 500ms
time
p < 0.002
perceptual
sensitivity
d′
1
0
uncued
stimulus
changed
cued
stimulus
changed
b
pairwise metrics
p < 0.001
attend-out
attend-in
r_sc mean
0.09
0.08
0.07
0.06
0.05
monkey 1
monkey 2
p < 0.05
r_sc s.d.
0.11
0.1
0.09
monkey 1
monkey 2
c
population-level metrics
p < 0.001
attend-out
attend-in
percent
shared
variance
20%
18%
16%
14%
monkey 1
monkey 2
p < 0.002
loading
similarity
0.52
0.5
0.48
0.46
0.44
monkey 1
monkey 2
p < 0.05
number of
dimensions
4
3.5
3
2.5
2
monkey 1
monkey 2
d
r_sc s.d.
0.1
0.05
0
0.05
0.1
r_sc mean
0.1
0.09
0.06
0.07
decrease
loading similarity
decrease
percent
shared
var.
decrease
dimensionality
attend-out
attend-in

Figure 8 ***(previous page)*****: An observed decrease in r_{sc} mean of macaque V4 neurons during a spatial attention task corresponds to changes in multiple population metrics. a**. Experimental task design. On each trial, monkeys maintained fixation while Gabor stimuli were presented for 400 ms (with 300-500 ms in between presentations). When one of the stimuli changed orientation, animals were required to saccade to the changed stimulus to obtain a reward. At the beginning of a block of trials, we performed an attentional manipulation by cuing animals to the location of the stimulus that was more likely to change for that block (dashed circle denotes the cued stimulus and was not presented on the screen). The cued location alternated between blocks. Animals were more likely to detect a change in stimulus at cued rather than uncued locations (inset in bottom right, $p < 0.002$ for both animals; data for monkey 1 is shown). During this task, we recorded activity from V4 neurons whose receptive fields (RFs) overlapped with one of the stimulus locations. **b**. r_{sc} mean (left panel) and r_{sc} s.d. (right panel) across recording sessions for two animals. Black denotes 'attend-out' trials (i.e., the cued location was outside the recorded V4 neurons' RFs), and red denotes 'attend-in' trials (i.e., the cued location was inside the RFs). Data was pooled across both animals to compute p-values reported in titles for comparison of attend-out (black) and attend-in (red). For individual animals, r_{sc} mean was lower for attend-in than attend-out ($p < 0.001$ for each animal). r_{sc} s.d. was also lower for attend-in than attend-out ($p < 0.05$ for monkey 1, and $p = 0.148$ for monkey 2). **c**. Population metrics identified across recording sessions for two animals (same data as in **b**). Black denotes attend-in trials, red denotes attend-out trials. Data was again pooled across animals to compute p-values reported in titles for comparing attend-out and attend-in. %sv was lower for attend-in than attend-out ($p < 0.001$ for monkey 1 and $p < 0.02$ for monkey 2). Loading similarity was lower for attend-in than attend-out ($p < 0.001$ for monkey 1 and $p = 0.162$ for monkey 2). Dimensionality was lower for attend-in than attend-out ($p = 0.113$ for monkey 1 and $p = 0.174$ for monkey 2). In panels **a-c**, dots indicate means and error bars indicate 1 s.e.m., both computed across recording sessions. **d**. Summary of the real data results. Attention decreases both r_{sc} mean and r_{sc} s.d. (black dot to red dot). These decreases in pairwise metrics correspond to a combination of decreases in %sv, loading similarity, and dimensionality (dashed arrows).

those in which the cued stimulus location was inside the aggregate receptive fields (RFs) of the recorded V4 neurons, whereas 'attend-out' trials were those in which the cued stimulus location was in the opposite visual hemifield.

For pairwise metrics, r_{sc} mean decreased when attention was directed into the RFs of the V4 neurons (Fig. 8*b*, left panel), consistent with previous studies [1–3, 10, 13]. We further found that r_{sc} s.d. was lower for attend-in trials than for attend-out trials, an effect not reported previously (Fig. 8*b*, right panel).

The decrease in both r_{sc} mean and r_{sc} s.d. could arise from several different types of distinct changes in population-wide covariability. To compute the population metrics, we applied factor analysis (FA) separately to attend-out and attend-in trials (see Methods). FA is the most basic dimensionality reduction method that characterizes shared variance among neurons [50], and is consistent with how we created covariance matrices in Figures 5 and 6. We found three distinct changes in population metrics. First, neuronal activity during attend-in trials had lower %sv than during attend-out trials (Fig. 8*c*, left), consistent with previous interpretations that attention reduces the strength of shared modulations [46, 48, 49, 76]. Second, we also found lower loading similarity for attend-in trials than attend-out trials for the dominant dimension (i.e., the dimension that explains the largest proportion of the shared variance; Fig. 8*c*, middle; see also Supplementary Fig. 2*b*). This implies that, with attention, neurons in the population co-fluctuate in a more heterogeneous manner (i.e., more pairs of neurons co-fluctuate in opposition, and fewer pairs co-fluctuate together). Third, we found that dimensionality was slightly lower for attend-in than attend-out trials (Fig. 8*c*, right). Thus, on average, a smaller number of distinct shared signals were present when attention was directed into the neurons' RFs. The small change in dimensionality is consistent with the relative strength of each dimension (i.e., eigenspectrum shape) being similar for attend-in and attend-out (Supplementary Fig. 2*a*). Taken together, this

collection of observations of both pairwise and population metrics leads to a more refined view of how attention affects population-wide covariability.

The pairwise (Fig. 8*b*) and population (Fig. 8*c*) metrics are computed based on the same recorded activity and each represents a different view of population activity. The central contribution of our work is to provide a framework by which to understand these two perspectives and five different metrics in a coherent manner. Using the relationships between pairwise and population metrics we have established in the r_{sc} mean versus r_{sc} s.d. space, we can decompose the decrease in r_{sc} mean and s.d. into: 1) a small decrease in dimensionality (Fig. 8*d*, small dashed arrow), 2) a decrease in loading similarity (Fig. 8*d*, medium dashed arrow), and 3) a substantial decrease in %sv (Fig. 8*d*, large dashed arrow). We quantify these contributions in Supplementary Fig. 3. The r_{sc} mean and s.d. decreased despite the decrease in dimensionality (which alone would have tended to increase r_{sc} mean and s.d.) because of the larger contributions of loading similarity and %sv to pairwise metrics in these V4 recordings. We have also applied the same analysis to population recordings in visual area V1 [78, available on CRCNS.org] and found that, although r_{sc} mean and s.d. both decreased (like in the V4 recordings), the population metrics changed in a different way compared to the V4 recordings (Supplementary Fig. 4). Together, these analyses demonstrate the need for considering both pairwise and population metrics together when studying correlated variability, with a bridge that allows one to navigate between the two.

2.6 Discussion

Coordinated variability in the brain has long been linked to the neural computations underlying a diverse range of functions, including sensory encoding, decision making, attention, learning, and more. In this study, we sought to relate two major bodies of work investigating the coordinated activity among neurons: studies that measure spike count correlation between pairs of neurons (r_{sc}) and studies that use dimensionality reduction to measure population-wide covariability. We considered three population metrics and established analytically and empirically that: 1) increasing loading similarity corresponds to increasing r_{sc} mean and decreasing r_{sc} s.d., 2) decreasing percent shared variance (%sv) corresponds to decreasing both r_{sc} mean and s.d., and 3) increasing dimensionality tends to decrease r_{sc} mean and s.d. Applying this understanding to recordings in macaque V4, we found that the previously-reported decrease in mean spike count correlation associated with attention stemmed from a decrease in %sv, a decrease in loading similarity, and decrease in dimensionality. This analysis revealed that attention involves multiple changes in how neurons interact that are not well captured by a single statistic alone. Overall, our work demonstrates that common ground exists between the literatures of spike count correlation and dimensionality reduction approaches, and builds the intuition and formalism to navigate between them.

Our work also provides a cautionary tale for attempting to summarize population-wide covariability using one, or a small number of, statistics. For example, reporting only r_{sc} mean is incomplete because several distinct changes in population-wide covariability can correspond to the same change in r_{sc} mean. In a similar vein, reporting only dimensionality is incomplete because it does not indicate how strongly the neurons covary, nor their co-fluctuation patterns. For this reason, we recommend reporting several different pairwise and population metrics (e.g., the five used in this study along with the eigenspectrum of the shared covariance matrix), as long as they can be reliably measured from the data available. This not only allows for a deeper and more complete understanding of how neurons covary, but also it allows one to make tighter connections to previous literature that uses the same metrics. Future work may seek to revisit previous results of correlated neuronal variability that are based on a single statistic (e.g., r_{sc} mean), and reinterpret them within a framework that considers multiple perspectives and

statistics of population-wide covariability, such as that presented here.

There are some situations where it is not feasible to reliably measure population statistics, such as recording from a small number of neurons in deep brain structures [79, 80], or when the number of trials is small relative to the number of neurons recorded. In such situations, the r_{sc} can be measured between pairs of neurons recorded in each session and then averaged across sessions to obtain the r_{sc} mean. Based on our findings, we recommend that studies which report r_{sc} mean also report r_{sc} s.d. because the latter provides additional information about population-wide covariability. For example, in the special case of one latent dimension (typically not known in advance for real data), measuring r_{sc} mean and r_{sc} s.d. allows one to estimate the loading similarity and %sv (cf. Fig. 5*e-f*). In general, even when there is more than one latent dimension in the population, r_{sc} s.d. provides value in situating the data in the r_{sc} mean versus r_{sc} s.d. plot. Changes in r_{sc} mean and s.d. can then inform changes in population metrics based on the relationships established in this work (cf. Fig. 8*d*).

The reason that our work, and many previous studies, have focused on trial-to-trial variability is that it has important implications for information coding. Early work on information-limiting correlations typically focused on r_{sc} mean [e.g., 1, 7, 34, 35], which reflects the strength of shared variability among neurons. Recent theoretical work [81, 82, 84] and experimental evidence [14, 41, 85, 86] has shown that it is not only the strength of shared trial-to-trial variability but also the directions of shared variability relative to stimulus tuning (Fig. 9*a*) that need to be considered for information coding. These properties of shared trial-to-trial variability are precisely what are measured by the population metrics used here. In particular, the %sv measures how strongly trial-to-trial variability is shared among neurons (Fig. 9*b*), loading similarity measures the direction(s) of variability (Fig. 9*c*), and dimensionality measures how many different directions of variability exist in the data (Fig. 9*d*). By considering these three population metrics together, along with the way in which mean population responses vary across conditions (i.e., the stimulus-encoding directions), we can more incisively characterize how trial-to-trial variability impacts information coding than by using r_{sc} mean alone. Understanding how patterns of shared variability are related to (e.g., align with or are orthogonal to) patterns of stimulus encoding and downstream readouts will be likely critical for understanding information coding in the brain.

We considered three population metrics — dimensionality, percent shared variance (%sv), and loading similarity — that summarize the structure of population-wide covariability and are rooted in well-established concepts in existing literature. First, dimensionality has been used to describe how neurons covary across conditions [i.e., an analysis of trial-averaged firing rates; 52, 55, 70, 83, 87], as well as how neurons covary from trial to trial [46, 57, 66–68, 72, 73, 88, 89]. We focused on the latter in our study to connect with the r_{sc} literature, which also seeks to understand the shared trial-to-trial variability between neurons. To focus on the shared variability among neurons, we used factor analysis (FA) to measure dimensionality. Another commonly-used dimensionality reduction method, principal components analysis (PCA), although appropriate for studying trial-averaged activity, does not distinguish between variability that is shared among neurons and variability that is independent to each neuron. Second, investigating the loading similarity has provided insight about whether shared variability among neurons arises from a shared global factor which drives neurons to increase and decrease their activity together [26, 46, 47, 49, 66, 90] or whether the co-fluctuations involve a more intricate pattern across the neuronal population [13, 41, 91]. Third, we have previously reported %sv for area V1 [66], area M1, and network models [66, 89]. Conceptually, %sv and r_{sc} mean are both designed to capture the strength of shared variability in a population of neurons. Thus, we might initially think that there should be a one-to-one correspondence between the two quantities. Indeed, if the population activity is described by one co-fluctuation pattern with a high loading similarity, there is a direct relationship between %sv and r_{sc} mean (Fig. 5*f*). However, in general, %sv and r_{sc} mean do not have a one-to-one correspondence between them (Fig. 5*f*, moderate or low loading similarity).

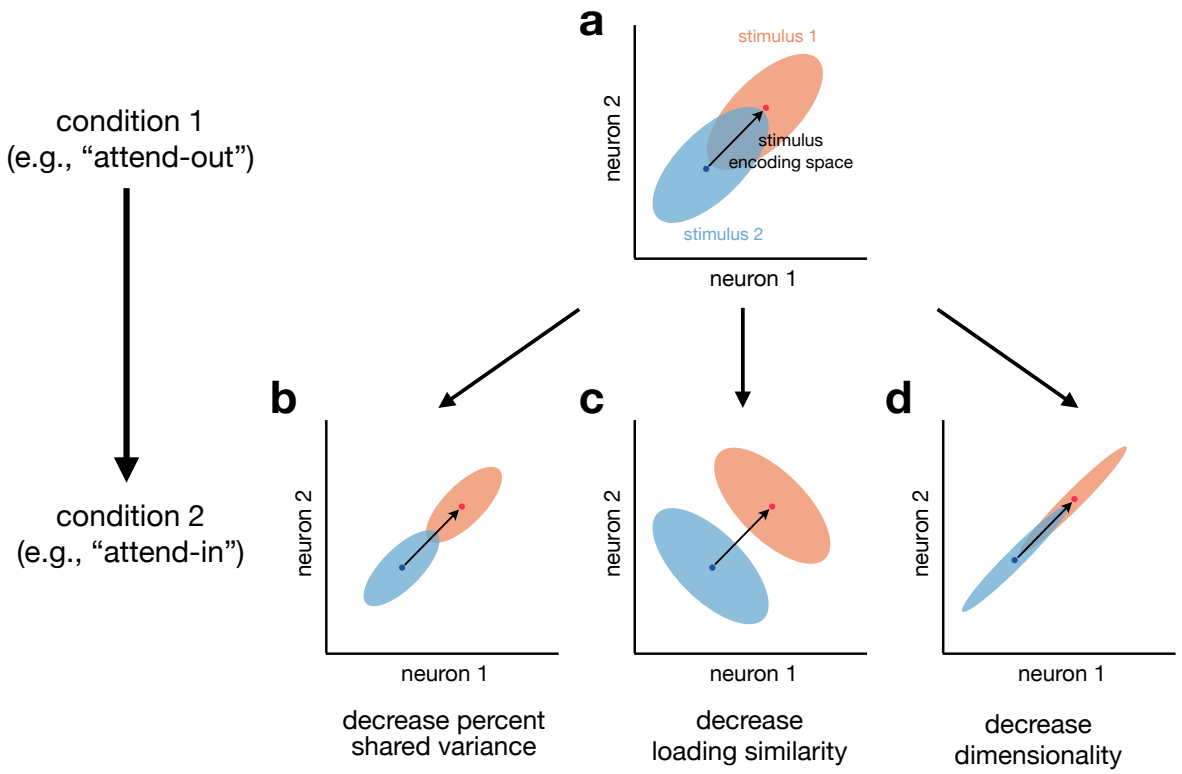


Figure 9: Population metrics and information coding. For illustrative purposes, we consider the responses of two neurons to two different stimuli. **a**. In “condition 1” (e.g., “attend-out” in our V4 analyses), the two neurons have positively correlated trial-to-trial variability (blue and orange clouds each have positive correlation) and a stimulus encoding space (black arrow) defined by the span of the trial-averaged responses (blue and orange dots). Then, we consider how changes in trial-to-trial neuronal variability (i.e., shapes of the clouds) from one experimental condition to another (e.g., spatial attention) can influence decoding of the two stimuli. For simplicity, we construct examples in which the stimulus encoding space remains constant between the two conditions. We illustrate here the changes in population metrics that we observed in our V4 data (Fig. 8*d*). **b**. First, a decrease in percent shared variance (both clouds are smaller in size) results in more accurate decoding of the population responses to the two stimuli (the blue and orange ellipses are less overlapping here than in panel **a**). **c**. Second, a decrease in the loading similarity of the strongest dimension (both clouds have been rotated to have negative correlation) also leads to an improvement in decoding performance. In this case, the improvement stems from the fact the stimulus encoding space (black arrow) and the strongest dimension of trial-to-trial variability (negative correlation) are misaligned [81, 82]. **d**. Third, a decrease in dimensionality (the less dominant dimension has been squashed for both clouds) could either improve or have no impact on decoding performance. Here, the dimension that was squashed (negative correlation direction) was orthogonal to the stimulus encoding dimension (black arrow), leading to no impact on decoding performance. In general, all else being equal, higher-dimensional trial-to-trial variability [distinct from high-d signal; 83] is more likely to overlap with stimulus encoding dimensions and thus limit the amount of information encoded.

We focus here on studying trial-to-trial activity fluctuations that are shared between neurons. Many studies have considered the source of these shared fluctuations in the context of pairwise correlations [7]. Most commonly, pairwise correlations have been suggested to originate through common input [34, 35]. However, there are in fact numerous mechanisms that can shape the trial-by-trial shared variability of neuronal populations, including neuromodulation [92, 93], coupled inhibition, or distinct patterns of neuronal connectivity [49, 66–68]. These mechanisms likely produce distinct signatures in population metrics, such as %sv, loading similarity, and

dimensionality. The framework that we have developed here can be applied to spiking network models with different underlying mechanisms of shared cortical variability to identify signatures in population metrics [49, 66–68]. We can then assess whether any of those signatures are present in neuronal recordings to gain insight into the underlying mechanisms of shared variability in the brain.

Although pairwise correlation and dimensionality reduction have most commonly been computed based on spike counts, several studies have also computed these metrics on neuronal activity recorded using other modalities, such as calcium imaging [51, 73, 85, 94]. The relationships that we established here between pairwise and population metrics are properties of covariance matrices in general and do not rely on or assume recordings of neuronal spikes. Thus, the intuition built here can be applied to other recording modalities.

Our work here focused on studying interactions within a single population of neurons. Technological advances are enabling recordings from multiple distinct populations simultaneously, including neurons in different brain areas, neurons in different cortical layers, or different neuron types [e.g., 95, 96]. Studies are dissecting the interactions between these distinct populations using pairwise correlation [3, 12, 78] and dimensionality reduction [41, 62–65, 89, 97]. As we have shown here for a single population of neurons, considering a range of metrics from both the pairwise correlation and dimensionality reduction perspectives, and understanding how they relate to one another, will provide rich descriptions of how different neuronal populations interact.

2.7 Methods

Spike count covariance matrix

Both pairwise metrics and population metrics are computed directly from the spike count covariance matrix Σ of size $n \times n$ for a population of n neurons. Each entry in Σ is the covariance between the activity of neuron i and neuron j:

$$\Sigma_{ij} = \mathrm{cov}(x_i, x_j) = \mathrm{E}[(x_i - \mu_i)(x_j - \mu_j)] \tag{2}$$

where x_i and x_j represent the activity of neurons i and j, respectively, and μ_i and μ_j represent the mean activity of neurons i and j, respectively. The variance of the ith neuron is equal to Σ_{ii}.

Pairwise metrics

We computed the spike count correlation (r_{sc}) between neurons i and j directly from the spike count covariance matrix:

$$\rho_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}} \tag{3}$$

We then summarized the distribution of r_{sc} values across all pairs of neurons in the population with two pairwise metrics: the r_{sc} mean and r_{sc} standard deviation (s.d.).

Population metrics

The metrics we use for characterizing population-wide covariability are based on factor analysis [FA; 49, 51, 66, 69, 72, 88, 89], a dimensionality reduction method. We chose FA because it is the most basic dimensionality reduction method that explicitly separates variance that is shared among neurons from variance that is independent to each neuron. This allows us to relate the population metrics provided by FA to spike count correlation, which is designed to measure shared variability between pairs of neurons. One might consider using principal component analysis (PCA), but it does not distinguish shared variance from independent variance. Thus,

FA is more appropriate than PCA for studying the shared variability among a population of neurons.

Decomposing the spike count covariance matrix
FA decomposes the spike count covariance matrix Σ into a low-rank shared covariance matrix, which captures the variability shared among neurons in the population, and an independent variance matrix, which captures the portion of variance of each neuron unexplained by the other neurons (Supplementary Fig. 5*a*):

$$\Sigma = \Sigma_{\text{shared}} + \Psi \tag{4}$$

where $\Sigma_{\text{shared}} \in \mathbb{R}^{n \times n}$ is the shared covariance matrix for n neurons, and $\Psi \in \mathbb{R}^{n \times n}$ is a diagonal matrix containing the independent variance of each neuron. The low-rank shared covariance matrix can be expressed using the eigendecomposition as (Supplementary Fig. 5*a*):

$$\Sigma_{\text{shared}} = U \Lambda U^T \tag{5}$$

where $U \in \mathbb{R}^{n \times d}$ and $\Lambda \in \mathbb{R}^{d \times d}$, with $d < n$. The rank (i.e., dimensionality) of the shared covariance matrix, d, indicates the number of latent variables. Each column of U is an eigenvector and represents a co-fluctuation pattern containing the loading weights of each neuron (i.e., how much each neuron contributes to that dimension). The matrix Λ is a diagonal matrix where each diagonal element is an eigenvalue and represents the amount of variance along the corresponding co-fluctuation pattern (e.g., in Fig. 4 panel *a* has larger eigenvalue than panel *c*).

Based on this matrix decomposition, we defined the three metrics that describe the population-wide covariability:

- **Loading similarity**: the similarity of loading weights across neurons for a given co-fluctuation pattern. Scalar value between 0 (the weights are maximally dissimilar, defined precisely below) and 1 (all weights are the same).
- **Percent shared variance (%sv)**: the percentage of each neuron's variance that is explained by other neurons in the population. Percentage between 0% and 100%.
- **Dimensionality**: the number of dimensions (i.e., co-fluctuation patterns). Integer value.

We give the precise definitions of these population metrics below and in Supplementary Fig. 5*b*.

Loading similarity
We sought to define loading similarity such that, for a given co-fluctuation pattern, if the weights for all neurons are the same, we would measure a loading similarity of 1. When the weights are as different as possible, we would measure a loading similarity of 0. We define the loading similarity based on the variance across the n weights (for n neurons) in a co-fluctuation pattern $\mathbf{u_k}$. The smallest possible variance is 0; the largest possible variance, for a unit vector $\mathbf{u_k}$, is $1/n$ (Math Note F). Thus, we define loading similarity for a co-fluctuation pattern $\mathbf{u_k} \in \mathbb{R}^n$ as:

$$\text{loading similarity}(\mathbf{u_k}) = 1 - \frac{\text{var}(\mathbf{u_k})}{\max_{\mathbf{v_k}} \text{var}(\mathbf{v_k})} = 1 - \frac{\text{var}(\mathbf{u_k})}{1/n} \tag{6}$$

where the loading similarity is computed on unit vectors (i.e., $\mathbf{u_k}$ has a norm of 1). The notation $\text{var}(\mathbf{u_k})$ denotes that the variance is being taken across the n elements of the vector $\mathbf{u_k}$. The denominator of equation (6) acts as a normalizing factor, bounding the loading similarity value between 0 and 1.

The loading similarity distinguishes between a co-fluctuation pattern along which all neurons in the population have the same weight in which case they change their activity up and down together (Fig. 4*a*; loading similarity of 1), from one in which weights are different and some neurons increase their activity when others decrease their activity (Fig. 4*b*; loading similarity of 0). The loading weights we use here are closely related to 'population coupling' [90] and 'modulator weights' [46]. For some types of shared fluctuations, these weights are similar across neurons in a population [i.e., high loading similarity; 46, 49, 90]. For other types of shared fluctuations, the weights vary substantially across neurons in the population [i.e., low loading similarity; 41].

We show in Math Note E why, if one dimension has high loading similarity, the other dimensions must have low loading similarity. The reason is that co-fluctuation patterns are defined to be mutually orthogonal. If one co-fluctuation pattern has all weights close to the same value (i.e., high loading similarity), then all other co-fluctuation patterns must have substantial diversity in their weights (i.e., low loading similarity) to satisfy orthogonality.

Percent shared variance

The percent shared variance (%sv) measures the percentage of each neuron's spike count variance that is explained by other neurons in the population [66, 89]. Equivalently, we can think of %sv in terms of latent co-fluctuations. Because latent co-fluctuations capture the shared variability among neurons, the %sv measures how much of each neuron's variance is explained by the latent co-fluctuations. The activity of neurons may be tightly linked to the latent co-fluctuation (e.g., Fig. 4*a*), in which case a large percentage of each neuron's variance is shared with other neurons, or may only be loosely linked to the latent co-fluctuation (e.g., Fig. 4*c*), in which case a small percentage of each neuron's variance is shared with other neurons. Mathematically, we define the %sv for a neuron i:

$$\%\text{sv for neuron i} = \frac{\Sigma_{shared,ii}}{\Sigma_{ii}} \cdot 100\% = \frac{s_i}{s_i + \psi_i} \cdot 100\% \tag{7}$$

where s_i is the i^{th} entry along the diagonal of the shared covariance matrix (Supplementary Fig. 5*a*, Σ_{shared}), and ψ_i is the i^{th} entry along the diagonal of the independent covariance matrix (Supplementary Fig. 5*a*, Ψ). A %sv of 0% indicates that the neuron does not covary with (i.e., is independent of) other neurons in the population, whereas a %sv of 100% indicates that the neuron's activity can be entirely accounted for by the activity of other neurons in the population. To compute %sv for an entire population of neurons, we averaged the %sv of the individual neurons. All %sv values reported in this study are the %sv for the neuronal population.

Dimensionality

Dimensionality refers to the number of latent co-fluctuations needed to describe population-wide covariability. For example, the population-wide covariability can be described by one latent co-fluctuation (Fig. 4*a*) or by several latent co-fluctuations (Fig. 4*d*). In the population activity space, dimensionality corresponds to the number of axes along which the population activity varies (see Fig. 4*d*, bottom inset). Mathematically, the dimensionality is the rank of the shared covariance matrix (i.e., the number of columns in U, Supplementary Fig. 5*a*).

Creating the spike count covariance matrices with specified population metrics

To relate pairwise and population metrics, we created spike count covariance matrices of the form in equation (4) with specified population metrics. Importantly, we did not simulate spike counts, nor fit a factor analysis model to simulated data. Rather, we created covariance matrices using (4) and computed pairwise correlations directly from the entries of the covariance matrix,

as shown in (3). Across simulations (Figs. 5 and 6), we simulated with $n = 30$ neurons and set independent variances (i.e., diagonal elements of Ψ in equation (4)) to 1.

Specifying co-fluctuation patterns to obtain different loading similarities
Each co-fluctuation pattern $\mathbf{u_k}$ is a vector with $n = 30$ entries (one entry per neuron). We generated a single co-fluctuation pattern by randomly drawing 30 independent samples from a Gaussian distribution with a mean of 2.5. We choose a nonzero mean so that we could obtain co-fluctuation patterns with loading similarities close to 1 when drawing from the Gaussian distribution (i.e., a mean of 0 would have resulted in almost all co-fluctuation patterns having a loading similarity close to 0). To get a range of loading similarities between 0 and 1, we used different standard deviations for the Gaussian. For a small standard deviation value, all entries in the co-fluctuation pattern are close to 2.5, resulting in a high loading similarity. For larger standard deviations, some loading weights are positive and some negative, with large variability in their values, resulting in co-fluctuation patterns with low loading similarity. We increased the Gaussian standard deviation from 0.1 to 5.5 with increments of size 0.1. For each increment, we generated 50 patterns and normalized them to have unit norm. In total, we created a set of 2,750 random patterns.

The following procedure describes the construction of shared covariance matrices with one co-fluctuation pattern. We chose a single pattern $\mathbf{u_1} \in \mathbb{R}^{30\times 1}$ (i.e., U has only 1 column) from the set of 2,750. We constructed the shared covariance matrix by computing $U\Lambda U^T$, where Λ was chosen to achieve a desired percent shared variance (see below). The covariance matrix was then computed according to equation (4). We created a covariance matrix, yielding a spread of loading similarities between 0 and 1 (Fig. 5*e-f*). In the next section, we describe the procedure for creating a covariance matrix with more dimensions.

Specifying the percent shared variance
To achieve a given %sv, either the independent variance or the amount of shared variability (i.e., the eigenvalues) of each dimension can be adjusted. In the main text, we set the independent variance of each neuron to $\Psi_i = 1$, and changed the total amount of shared variability by multiplying each eigenvalue (each diagonal element in Λ from equation (5)) by the same constant value, a. To obtain a specified %sv, we identified a by searching through a large set of possible values (from 10^{-4} to 10^3 with step size 10^{-3}). We allowed for a tolerance of $\epsilon = 10^{-3}$ between the desired %sv and the %sv that was achieved after scaling the eigenvalues by a. In other analyses, we allowed the independent variances to be different across neurons (e.g., drawn from an exponential distribution), and the relationships between pairwise and population metrics were qualitatively similar to those in the main text.

Increasing dimensionality
To assess how changing dimensionality affects pairwise metrics, we created covariance matrices whose shared covariance matrix comprised more than 1 dimension. To create a shared covariance matrix with d dimensions, we randomly chose d patterns from the set of 2750 we had generated above (see 'Specifying co-fluctuation patterns to obtain different loading similarities'). We then orthogonalized the chosen patterns using the Gram-Schmidt process to obtain d orthonormal (i.e., orthogonal and unit length) co-fluctuation patterns $U \in \mathbb{R}^{30\times d}$. We formed the shared covariance matrix using $U\Lambda U^T$, where $\Lambda \in \mathbb{R}^{d\times d}$ is a diagonal matrix containing the eigenvalues (i.e., the strength of each dimension; see 'Specifying the relative strengths of each dimension' below). We repeated this procedure to produce 3,000 sets of d orthonormal patterns (i.e., 3,000 different U matrices), each of which was used to create a shared covariance matrix. The spike count covariance was computed according to equation (4).

Specifying the relative strengths of each dimension
In simulating shared covariance matrices with more than one dimension, we chose the relative

strength of each dimension by specifying the eigenspectrum (diagonal elements of Λ in equation (5)). We worked with three sets of eigenspectra. First, a flat eigenspectrum had eigenvalues that were all equal (Fig. 5*g*). Second, for two dimensions, we varied the ratio of the two eigenvalues between 95:5, 80:20, 50:50, 20:80, and 5:95 (Fig. 6). Third, we considered an eigenspectrum in which each subsequent eigenvalue falls off according to an exponential function (Supplementary Fig. 1). Only the relative (and not the absolute) eigenvalues (i.e., the shape of the eigenspectrum) affect the results, because the eigenspectrum was subsequently scaled to achieve a desired %sv (see 'Specifying the values of percent shared variance').

Analysis of V4 neuronal recordings from a spatial attention task

Electrophysiological recordings

We analyzed data from a visual spatial attention task reported in a previous study [77]. Briefly, we implanted a 96-electrode "Utah" array (Blackrock Microsystems; Salt Lake City, UT) into visual cortical area V4 of an adult male rhesus macaque monkey (data from two monkeys were analyzed; in our study, monkey 1 corresponds to "monkey P" and monkey 2 corresponds to "monkey W" from [77]). After recording electrode voltages (Ripple Neuro.; Salt Lake City, UT), we used custom software to perform off-line spike sorting [98, freely available at `https://github.com/smithlabvision/spikesort`]. This yielded 93.2 $\pm$ 8.9 and 61.9 $\pm$ 27.4 candidate units per session for monkey 1 and 2, respectively. Experiments were approved by the Institutional Animal Care and Use Committee of the University of Pittsburgh and were performed in accordance with the United States National Research Council's Guide for the Care and Use of Laboratory Animals.

To further ensure the isolation quality of recorded units, we removed units from our analyses according to the following criteria. First, we removed units with a signal-to-noise ratio of the spike waveform less than 2.0 [98]. Second, we removed units with overall mean firing rates less than 1 Hz, as estimates of r_{sc} for these units tends to be poor [7]. Third, we removed units that had large and sudden changes in activity due to unstable recording conditions. For this criterion, we divided the recording session into ten equally-sized blocks and for each unit computed the difference in average firing rate between adjacent blocks. We excluded units with a change in average firing rate greater than 60% of the maximum firing rate (where the maximum is taken across the ten equally-sized blocks). Fourth, we removed an electrode from each pair of electrodes that were likely electrically-coupled. We identified the coupled electrodes by computing the fraction of threshold crossings that occurred within 100 μs of each other for each pair of electrodes. We then removed the fewest number of electrodes to ensure this fraction was less than 0.2 (i.e., pairs with an unusually high number of coincident spikes) for all pairs of electrodes. Fifth, we removed units that did not sufficiently respond to the visual stimuli used in the experiment. Evoked spike counts (i.e., a neuron's response after stimulus presentation) were taken between 50 ms to 250 ms after stimulus onset, and spontaneous spike counts (i.e., a neuron's response during a blank screen) were taken in a 200 ms window that ended 50 ms before stimulus onset. For each unit, we computed a sensitivity measure d' between evoked and spontaneous activity:

$$d' = \frac{\mu_{\text{evoked}} - \mu_{\text{spontaneous}}}{\sqrt{\frac{1}{2}(\sigma^2_{\text{evoked}} + \sigma^2_{\text{spontaneous}})}}$$

for mean spike counts μ_{evoked} and $\mu_{\text{spontaneous}}$ and spike count variances σ^2_{evoked} and $\sigma^2_{\text{spontaneous}}$. We removed units with $d' < 0.5$ from analyses, as these units had spontaneous and evoked responses that were difficult to distinguish.

After applying these five criteria, 44.5 ± 11.3 and 18.8 ± 6.7 units per session (mean ± s.d. over sessions) remained for monkeys 1 and 2, respectively. Although these remaining units likely contained both single-unit and multi-unit activity, we refer to each unit as a neuron for simplicity. In this study, we restricted analyses to sessions with at least 10 neurons remaining after applying the above criterion (23 sessions for monkey 1, and 14 sessions for monkey 2).

Visual stimulus change-detection task

Animals were trained to perform a change-detection task with a spatial attention cue to the location of the visual stimulus that was more likely to change [13]. In the visual change-detection task (Fig. 8*a*), animals fixated a central dot while Gabor stimuli were presented in two locations on a computer screen. One location was chosen to be within the aggregate receptive fields (RFs) of the recorded V4 neurons (mapped prior to running the experiment), and the other location was placed at the mirror symmetric location in the opposite hemifield. Animals maintained fixation while a sequence of Gabor stimuli were presented. Each drifting Gabor stimulus (oriented at either 45° or 135°) was presented for 400 ms, followed by a blank screen presented for a random interval (between 300 and 500 ms). The sequence continued, with a fixed probability for each presentation, until one of the two stimuli changed orientation when presented (i.e., the 'target'). Upon target presentation, animals were required to make a saccade to the target to earn a juice reward. We manipulated spatial attention in the experiment by cueing the more probable target location in blocks. At the beginning of each block, the cue was denoted by presenting only one Gabor stimulus at the more probable target location (90% likely), and requiring animals to detect orientation changes at this location for 5 trials. Consistent with the results of previous studies, we found that animals had greater perceptual sensitivity for orientation changes at the cued (i.e., attended) location than the uncued location (Fig. 8*a*, inset in the bottom right) and shorter reaction times [13].

Data processing and computing spike counts

We first separated the trials into two groups: (1) "attend in" trials, for which the cued stimulus was inside the recorded neurons' RFs and (2) "attend out" trials, for which the cued stimulus was outside the RFs. Since the initial orientation of the stimulus at the cued location could be one of two values (i.e., 45° or 135°), we further divided trials, resulting in a total of 4 groups of trials per session (attend in & 45°, attend out & 45°, attend in & 135°, attend out & 135°). Each combination of cued location and stimulus orientation was treated as an independent sample. The same neurons were used for each of the 4 groups within each session, ensuring a fair comparison between the attend-in and attend-out conditions.

We analyzed all stimulus presentations for which the target stimulus did not change. For each stimulus presentation, we took spike counts in a 200 ms window starting 150 ms after stimulus onset. For each of the 4 groups, we formed a spike count matrix $X \in \mathbb{R}^{n \times t}$, containing the spike counts of the n recorded neurons for the t trials belonging to that group. These spike count matrices were then used to compute both the pairwise and population metrics (described below). For all analyses (Fig. 8), we excluded recording sessions with fewer than 10 neurons. Additionally, because population metrics depend on the number of trials [66], for each session we equalized the number of trials across the 4 groups by randomly subsampling from groups with larger numbers of trials.

Computing pairwise metrics for V4 spike counts

We computed pairwise metrics on each combination of attention state ('attend in' and 'attend out') and stimulus orientation. We computed the correlation matrix for X as described above in 'Pairwise metrics' and then computed r_{sc} mean and r_{sc} s.d. For each attention state, we averaged the r_{sc} mean and r_{sc} s.d. over sessions and different stimulus orientations.

Computing population metrics for V4 spike counts

We fit the parameters of a factor analysis model (see Supplementary Fig. 5*a*) to each spike count matrix X (as described above) using the expectation-maximization (EM) algorithm. For each session, this was performed separately for each attention state and stimulus orientation. Using the FA parameters, we then computed the three population metrics (Supplementary Fig. 5*b*). For dimensionality, we first found the number of dimensions d that maximized the cross-validated data likelihood. We fit an FA model with d dimensions, and then found the number of dimensions required to explain 95% of the shared variance, termed d_{shared} [66]. We report d_{shared} because it tends to be a more reliable estimate of dimensionality than the number of dimensions that maximizes the cross-validated data likelihood. We computed %sv as described by equation (7). We report the loading similarity as defined in equation (6) for the co-fluctuation pattern that explained the most shared variability (i.e., the eigenvector with the largest eigenvalue; see Supp. Fig. 1 for why the loading similarity of this dimension is most informative), since it contributes most to describing the population-wide covariability. For 'attend in' and 'attend out' conditions, we averaged the population metrics across sessions and stimulus orientations.

Much of our work focuses on systematically changing a single population metric and assessing changes in pairwise metrics (Fig. 5*a*-*d*). When analyzing neuronal recordings, one needs to fit factor analysis to the recordings in order to estimate the population metrics. When estimating the population metrics together, it could be the case that changes in one population metric impacts or biases the estimation of another population metric. We characterized these estimation errors in Supplementary Fig. 6. Moreover, in Supplementary Fig. 7, we show that our main findings are the same when estimating population metrics from Poisson simulated data, which resembled realistic neuronal activity.

Statistics

We employed paired permutations tests for all statistical comparisons of pairwise metrics and population metrics between 'attend-in' and 'attend-out' conditions (Fig. 8*b*-*c*). First, for a given metric, we computed its value separately for each stimulus type (i.e., 45° or 135°), condition (i.e., attend-in or attend-out), and session. We then averaged the difference between attend-in and attend-out across stimulus types and sessions. To compute a null distribution, we randomly permuted the pair of attend-in and attend-out labels for each stimulus type and condition combination and recomputed the average difference. We ran $10,000$ permutations to obtain a null distribution of $10,000$ samples. We computed p-values as the proportion of samples in the null distribution that were more extreme than the average difference in the data, corresponding to $p < 0.0001$ as the highest attainable level of significance in our statistical analyses.

2.8 Math Notes

A Relationship between correlation, loading similarity, and %sv (one latent dimension)

We establish here the mathematical relationship between r_{sc}, loading similarity, and %sv. This will provide the formalism for understanding why decreasing %sv decreases both r_{sc} mean and s.d. (Fig. 5*f*), that a high loading similarity corresponds to large r_{sc} mean and low r_{sc} s.d. (Fig. 5*e*), and that a low loading similarity corresponds to small r_{sc} mean and large r_{sc} s.d. (Fig. 5*e*).

Let n be the number of neurons, and let $\mathbf{w}$ be the co-fluctuation pattern (i.e., loading vector $[w_1, w_2, ..., w_n]^T \in \mathbb{R}^{n\times 1}$), $\lambda \in \mathbb{R}_+$ be the strength of the co-fluctuation pattern (i.e., eigenvalue of the shared covariance matrix), and $\Psi \in \mathbb{R}^{n\times n}$ be a diagonal matrix specifying the independent variance of each neuron $(\psi_1, \psi_2, ..., \psi_n)$. Then the covariance matrix of the population activity is (see Methods and Supplementary Fig. 5):

$$\Sigma = \Sigma_{shared} + \Psi = \mathbf{w}\lambda\mathbf{w}^T + \Psi$$

From this, we observe that $\Sigma_{ij} = \Sigma_{shared,ij} = \lambda w_i w_j$ on the off-diagonal entries (i.e., if $i \neq j$). Along the diagonals, $\Sigma_{shared,ii} = \lambda w_i^2$ and $\Sigma_{ii} = \lambda w_i^2 + \psi_i$. The correlation (i.e., r_{sc} if Σ is a spike count covariance matrix) between neurons i and j can be written as:

$$\begin{aligned} \rho_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}} &= \frac{\lambda w_i w_j}{\sqrt{(\lambda w_i^2 + \psi_i)(\lambda w_j^2 + \psi_j)}} \\ &= \sqrt{\frac{\lambda w_i^2}{\lambda w_i^2 + \psi_i}}\sqrt{\frac{\lambda w_j^2}{\lambda w_j^2 + \psi_j}}\ \text{sign}(w_i w_j) \\ &= \sqrt{\phi_i \phi_j}\ \text{sign}(w_i w_j) \end{aligned} \tag{8}$$

where ϕ_i and ϕ_j represent the %sv (as proportions) for neurons i and j, respectively, and $\text{sign}(w_i w_j) = +1$ if $w_i w_j > 0$ or -1 if $w_i w_j < 0$. The last line follows from the fact that %sv is defined in equation (7) as:

$$\phi_i = \frac{\Sigma_{shared,ii}}{\Sigma_{ii}} = \frac{\lambda w_i^2}{\lambda w_i^2 + \psi_i} \tag{9}$$

Equations (8) and (9) provide a basis for understanding the relationships between r_{sc}, %sv, and loading similarity. The r_{sc} mean and s.d. are computed across all pairs of neurons ρ_{ij}, for $i < j$.

For establishing a relationship between pairwise metrics and %sv, consider decreasing the overall %sv of the population, while keeping the loadings w_i fixed. This corresponds to decreasing λ in equation (9), which implies ϕ_i for each neuron decreases, and thus the product $\sqrt{\phi_i \phi_j}$ decreases for all pairs. The magnitude of each ρ_{ij} decreases (i.e., each ρ_{ij} moves closer to 0). As such, decreasing %sv of the population decreases the distance of a point from the origin in the r_{sc} mean versus r_{sc} s.d. plot, all else being equal (Fig. 5*f*).

For establishing a relationship between pairwise metrics and loading similarity, consider two extreme cases: 1) when loading similarity is 1 (as high as possible) 2) when it is 0 (as low as possible). We first assume that each neuron has the same independent variance ψ_i for simplicity, as we did in Figure 5. A loading similarity of 1 corresponds to each $w_i = +\frac{1}{\sqrt{n}}$ or each $w_i = -\frac{1}{\sqrt{n}}$. In either case, $\text{sign}(w_i w_j)$ is always +1. Furthermore, ϕ_i is the same for every neuron and $\sqrt{\phi_i \phi_j}$ = %sv (i.e., the %sv of the population, expressed as a proportion) for every pair of neurons. Thus, all ρ_{ij} = %sv for all pairs of neurons i and j. In this case, r_{sc} mean = %sv and r_{sc} s.d. = 0. If the independent variances ψ_i are different across neurons, we can still get each $\text{sign}(w_i w_j) = +1$ and each ϕ_i to be the same by setting each $w_i = +\sqrt{\psi_i}$ or each $w_i = -\sqrt{\psi_i}$. This would also result in ρ_{ij} = %sv for all pairs of neurons i and j, and thus r_{sc} mean = %sv and r_{sc} s.d. = 0. In this case, the loading similarity is still high (all w_i are the same sign; we can show that load. sim.> 0.5), but not equal to 1.

Now, consider a scenario in which half the loadings are $+\frac{1}{\sqrt{n}}$ and the other half are $-\frac{1}{\sqrt{n}}$ (and assume again that ψ_i are the same for every neuron). This is one way to obtain a loading similarity of 0. In this case, ϕ_i are still the same for every neuron, so $\sqrt{\phi_i \phi_j}$ = %sv for all pairs. However, $\text{sign}(w_i w_j) = -1$ for $\left(\frac{n}{2}\right)^2 = \frac{n^2}{4}$ pairs, and $\text{sign}(w_i w_j) = +1$ for $2 \times \binom{n/2}{2} = \frac{n^2}{4} - \frac{n}{2}$ pairs. We can show that r_{sc} mean $= -\frac{\%\text{sv}}{n-1}$ and, by using equation (10) from Math Note B below, r_{sc} s.d. = %sv $\sqrt{1 - \frac{1}{(n-1)^2}}$. Thus, for a large number of neurons n, this case (where loading similarity=0) corresponds to small negative r_{sc} mean (close to 0), and large r_{sc} s.d. (close to the

%sv). As an example, for 30 neurons and %sv=50%, this corresponds to r_{sc} mean = -0.0172 and r_{sc} s.d. = 0.4997.

With this analysis, we have established that for one latent dimension:

- Decreasing %sv decreases the magnitudes of correlations (i.e., each ρ_{ij} closer to 0). r_{sc} mean and s.d. both decrease (as seen empirically in Fig. 5*f*).
- Starting from a loading similarity near 1, a decrease in loading similarity involves flips in the signs of some correlations (i.e., some ρ_{ij} become $-\rho_{ij}$). r_{sc} mean decreases but r_{sc} s.d. increases (as seen empirically in Fig. 5*f*).
- Both r_{sc} mean and %sv measure shared variance among neurons, but they are not always equal. Equations (8) shows that the two quantities are equal if all sign$(w_i w_j)$ are the same (i.e., when loading similarity is high). However, in general r_{sc} mean and shared variance (%sv) are not the same—e.g., when loading similarity is low, or when there are multiple dimensions (Math Note C).

In this section, we consider the extremes of loading similarity. In the next section, we analyze how gradual changes in loading similarity affect r_{sc} mean and s.d. for a fixed %sv.

B Circular arc in r_{sc} mean versus r_{sc} s.d. plot for one latent dimension and fixed %sv

We establish here mathematically that gradually varying the loading similarity for one latent dimension and fixed %sv results in an arc-like relationship between r_{sc} mean and r_{sc} s.d., and that the radius of the arc is approximately equal to the %sv (Fig. 5*e-f*).

We use the same notation as in Math Note A. Let $E[.]$ and $Var(.)$ denote the mean and variance across all neurons or all pairs of neurons, depending on context. In particular, we are interested in $E[\rho] = r_{\text{sc}}$ mean, $\sqrt{Var(\rho)} = r_{\text{sc}}$ s.d., where the expectation and variance are computed across ρ_{ij} for all pairs of neurons in a given population (i.e., the upper triangle of the correlation matrix, ρ_{ij} for $i > j$).

Let c be the distance of a point (corresponding to one instance of the population activity covariance matrix) from the origin in the r_{sc} mean versus r_{sc} s.d. plot (i.e., $c = \sqrt{(r_{\text{sc}} \text{ mean})^2 + (r_{\text{sc}} \text{ s.d.})^2}$). We want to know whether c is the same for all population activity covariance matrices with one latent dimension and fixed %sv. This would correspond to point being equidistant from the origin, and thus a circular arc. We can write c as:

$$\begin{aligned} c^2 &= (r_{\text{sc}} \text{ mean})^2 + (r_{\text{sc}} \text{ s.d.})^2 \\ &= E[\rho]^2 + Var(\rho) \\ &= E[\rho]^2 + E\left[\rho^2\right] - E[\rho]^2 \\ &= E\left[\rho^2\right] \end{aligned}$$

Thus, the squared distance (i.e., squared radius) is equal to $E\left[\rho^2\right]$, the mean of ρ_{ij}^2 across all pairs in the population. Let m be the number of pairs (i.e., $m = \binom{n}{2} = \frac{n(n-1)}{2}$). Now, using equations (8) and (9) derived in Math Note A:

$$
\begin{aligned}
E\left[\rho^2\right] &= \frac{1}{m}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\rho_{ij}^2 \\
&= \frac{1}{m}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{(\lambda w_i^2)(\lambda w_j^2)}{\left(\lambda w_i^2+\psi_i\right)\left(\lambda w_j^2+\psi_j\right)} \\
&= \frac{1}{m}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\phi_i\phi_j
\end{aligned}
$$

where ϕ_i and ϕ_j are the %sv of neurons i and j (expressed as proportions), as defined in Math Note A. We can show that $2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\phi_i\phi_j = \sum_{i=1}^{n}\sum_{j=1}^{n}\phi_i\phi_j - \sum_{i=1}^{n}\phi_i^2$. Intuitively, if we have a symmetric matrix Φ with entries $\Phi(i,j) = \phi_i\phi_j$, and we want to find the sum of the off-diagonal elements ($2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\phi_i\phi_j$), then we can take the sum of all elements and subtract the diagonal elements ($\sum_{i=1}^{n}\sum_{j=1}^{n}\phi_i\phi_j - \sum_{i=1}^{n}\phi_i^2$). Using this equivalence, it follows:

$$
\begin{aligned}
E\left[\rho^2\right] &= \frac{1}{m}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\phi_i\phi_j \\
&= \frac{1}{2m}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\phi_i\phi_j - \sum_{i=1}^{n}\phi_i^2\right) \\
&= \frac{1}{2m}\left(\sum_{i=1}^{n}\phi_i\sum_{j=1}^{n}\phi_j - \sum_{i=1}^{n}\phi_i^2\right) \\
&= \frac{1}{2m}\left(n^2E\left[\phi\right]^2 - \sum_{i=1}^{n}\phi_i^2\right) \\
&= \frac{1}{n-1}\left(nE\left[\phi\right]^2 - E\left[\phi^2\right]\right) \\
&= \frac{1}{n-1}\left(nE\left[\phi\right]^2 - Var\left(\phi\right) - E\left[\phi\right]^2\right) \\
&= \frac{1}{n-1}\left((n-1)E\left[\phi\right]^2 - Var\left(\phi\right)\right) \\
&= E\left[\phi\right]^2 - \frac{1}{n-1}Var\left(\phi\right) \\
&= (\%sv)^2 - \frac{1}{n-1}Var\left(\phi\right) \qquad (10)
\end{aligned}
$$

This provides an equation for the squared radius (i.e., squared distance from the origin) of a point in the r_{sc} mean versus r_{sc} s.d. plot. In the above derivation, $E\left[\phi\right]$ and $Var\left(\phi\right)$ are taken across the percent shared variance of each neuron in the population ϕ_i. Thus, $E\left[\phi\right]$ is equal to our population metric %sv. Now, we will bound $Var\left(\phi\right)$, which by definition is greater than or equal to 0. Since $0 \leq \phi_i \leq 1$, one instance where the maximum variance occurs is when there are an equal number of $\phi_i = 0$ and $\phi_i = 1$ (and $E\left[\phi\right] = 0.5$). Then,

$$
\begin{aligned}
Var(\phi) &= \frac{1}{n}\sum_{i=1}^{n}(\phi_i - 0.5)^2 \\
&= \frac{1}{n}\left(\frac{n}{2}(1-0.5)^2 + \frac{n}{2}(0-0.5)^2\right) \\
&= \frac{1}{n}(0.25n) \\
&= 0.25
\end{aligned}
$$

So $0 \leq Var(\phi) \leq 0.25$. For a small number of neurons n, the second term is non-negligible. For example, for a model with 6 neurons and %sv = 50%, the radius of the data points may vary between 0.4472 and 0.5. As the number of neurons increases, the second terms becomes negligible, and data points lie approximately along an arc with radius equal to %sv. For example, for 30 neurons as in our simulations and a %sv of 50%, the radius only varies between 0.4913 and 0.5.

To summarize, equation (10) computes the distance from the origin of a point for a given population of neurons. For a fixed %sv, $Var(\phi)$ can be the same or differ across many simulation runs. If $Var(\phi) = 0$ or is the same across runs, then the points will lie perfectly along an arc, with radius specified by equation (10). However, if $Var(\phi)$ is different across runs, the distances of each point from the origin will differ slightly, so they will lie close to, but not exactly along, an arc.

With this analysis, we have shown that in the case of one latent dimensions:

- A point (i.e., corresponding to a given population of neurons, simulated or real) on the r_{sc} mean versus r_{sc} s.d. plot has distance from the origin (i.e., radius) less than or equal to %sv.
- If the %sv for individual neurons (ϕ_i) are all the same (see Math Note A), then the radius equals %sv.
- As the number of neurons increases, the radius becomes asymptotically closer to %sv.

C Relationship between correlation, loading similarity, and %sv (multiple latent dimensions)

In Math Note A, we established a mathematical relationship between r_{sc}, loading similarity, and %sv in the case of one latent dimension. Here, we generalize equation (8) to include multiple dimensions in order to better understand the relationship between r_{sc} and dimensionality. We demonstrate here that the general relationships between r_{sc}, %sv, and loading similarity for one latent dimension also hold true for multiple latent dimensions. For multiple latent dimensions, the relative strengths of each dimension is an important consideration—a stronger dimension plays a bigger role in determining the r_{sc} distribution. Finally, we consider the relationship between dimensionality itself and r_{sc}. We will discover below that increasing dimensionality tends to decrease the magnitude of r_{sc} values.

First, consider the case of two latent dimensions. Again, let n be the number of neurons, let $\mathbf{w}$ be the co-fluctuation pattern (i.e., loading vector $[w_1, w_2, ..., w_n]^T \in \mathbb{R}^{n\times 1}$) with eigenvalue λ_w, let $\mathbf{v}$ be another pattern orthogonal to $\mathbf{w}$ ($[v_1, v_2, ..., v_n]^T \in \mathbb{R}^{n\times 1}$; $\mathbf{v} \perp \mathbf{w}$), with eigenvalue λ_v, and let $\Psi \in \mathbb{R}^{n\times n}$ be a diagonal matrix specifying the independent variance of each neuron $(\psi_1, \psi_2, ..., \psi_n)$. Then the covariance is $\Sigma = \Sigma_{shared} + \Psi = \Sigma_w + \Sigma_v + \Psi = \mathbf{w}\lambda_w\mathbf{w}^T + \mathbf{v}\lambda_v\mathbf{v}^T + \Psi$. On the off-diagonals entries (i.e., if $i \neq j$), $\Sigma_{ij} = \lambda_w w_i w_j + \lambda_v v_i v_j$. Along the diagonals, $\Sigma_{shared,ii} = \Sigma_{w,ii} + \Sigma_{v,ii} = \lambda_w w_i^2 + \lambda_v v_i^2$ and $\Sigma_{ii} = \lambda_w w_i^2 + \lambda_v v_i^2 + \psi_i$.

Because the shared covariance matrix Σ_{shared} can be expressed as a sum of two component matrices $\Sigma_w + \Sigma_v$, we can express the %sv of neuron i (ϕ_i) as

$$\begin{aligned}\phi_i = \frac{\Sigma_{shared,ii}}{\Sigma_{ii}} &= \frac{\Sigma_{w,ii}}{\Sigma_{ii}} + \frac{\Sigma_{v,ii}}{\Sigma_{ii}} \\ &= \frac{\lambda_w w_i^2}{\lambda_w w_i^2 + \lambda_v v_i^2 + \psi_i} + \frac{\lambda_v v_i^2}{\lambda_w w_i^2 + \lambda_v v_i^2 + \psi_i} \\ &= \phi_i^{(w)} + \phi_i^{(v)}\end{aligned}$$

where $\phi_i^{(w)}$ is the %sv variance of neuron i explained by dimension $\mathbf{w}$ and $\phi_i^{(w)}$ is the %sv variance of neuron i explained by dimension $\mathbf{v}$.

With this decomposition of ϕ_i, and following similar steps as in equation (8):

$$\rho_{ij} = \sqrt{\phi_i^{(w)}\phi_j^{(w)}}\,\mathrm{sign}(w_i w_j) + \sqrt{\phi_i^{(v)}\phi_j^{(v)}}\,\mathrm{sign}(v_i v_j) \tag{11}$$

where %sv values (ϕ) are represented as proportions. Equation (11) relates r_{sc}, %sv, and loading similarity for the case of two latent dimensions. Next, we compare these relationships for one versus two latent dimensions.

We will show that, for two latent dimensions, the relative strength of each dimension (i.e., the ratio $\lambda_w : \lambda_v$) is an important consideration. For two latent dimensions, decreasing the overall %sv by decreasing both $\phi^{(w)}$ and $\phi^{(v)}$ equally (e.g., $\lambda_w = \lambda_v$ and both decrease equally) pushes each ρ_{ij} closer to 0–r_{sc} mean and s.d. will decrease. This is similar to what happens for one latent dimension when %sv is decreased. On the other hand, even if the overall %sv is held constant, but $\phi^{(w)}$ increases relative to $\phi^{(v)}$ (i.e., increase the strength of $\mathbf{w}$ relative to $\mathbf{v}$), pairwise correlations could change. Each ρ_{ij} will largely be determined by $\phi^{(w)}$ and $\mathbf{w}$—r_{sc} mean and s.d. will be more similar to what they would be if only $\mathbf{w}$ existed (Fig. 6*a*). In other words, each ρ_{ij} for two latent dimensions is the sum of the ρ_{ij} that would have been produced by each of the two constituent dimensions on their own. The dimension with larger relative strength λ will have larger ϕ; the stronger dimension will play a larger role in determining each value of ρ_{ij} and thus the resulting r_{sc} distribution.

Using this logic, we can deduce that increasing the loading similarity of one of the dimensions would increase r_{sc} mean and decrease r_{sc} s.d. for the same reasons as for one latent dimension (Math Note A). Doing so for a relatively stronger dimension would result in larger changes in r_{sc} than doing so for a relatively weaker dimension.

We have shown how having multiple latent dimensions can affect the relationship between r_{sc}, %sv, and loading similarity. Now, we show that dimensionality itself and r_{sc} are related—larger dimensionality tends to decrease r_{sc} mean and s.d. To see this, we can generalize equation (11) for $d < n$ orthogonal latent dimensions $\mathbf{u_1}, \ldots, \mathbf{u_d} \in \mathbb{R}^n$.

$$\rho_{ij} = \sum_{k=1}^{d} \sqrt{\phi_i^{(u_k)}\phi_j^{(u_k)}}\,\mathrm{sign}(u_{k_i} u_{k_j})$$

Considering the sign of one term, ρ_{ij} could have the same sign for $\mathrm{sign}(u_{k_i} u_{k_j})$ across all dimensions $\mathbf{u_1}, \ldots, \mathbf{u_d}$; in this case, larger dimensionality acts to increase the correlation between neurons i and j (ρ_{ij}) above the level corresponding to a single dimension. However, because the loading vectors $\mathbf{u_1}, \ldots, \mathbf{u_d}$ are orthogonal, a pair of neurons i and j is likely to have many $\mathrm{sign}(u_{k_i} u_{k_j})$ of opposite sign across dimensions; in this case, larger dimensionality pushes the

correlation between neurons i and j (ρ_{ij}) closer to 0. Thus, we would expect the magnitude of correlations to decrease as more dimensions are added (i.e., a tendency for r_{sc} mean and s.d. to decrease; Fig. 5g). In the next section, we show this relationship mathematically.

D Increasing dimensionality decreases arc radius

We establish here that increasing dimensionality results in a decrease in the radius of the arc in the r_{sc} mean versus r_{sc} s.d. plot (Fig. 5g). We extend the math for an arc for one latent dimension (Math Note B) to multiple latent dimensions. We will refer to the one latent dimension as the '1-d case' and multiple (k) latent dimensions as the 'k-d case'.

We use the same notation as in Math Note C. Consider the distance c of a point (corresponding to one instance of the population activity covariance matrix) from the origin in the r_{sc} mean versus r_{sc} s.d. plot. From Math Note B, $c^2 = E[\rho^2]$. For this 2-d case, the correlation between neurons i and j is $\rho_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}} = \frac{\lambda_w w_i w_j + \lambda_v v_i v_j}{\sqrt{\left(\lambda_w w_i^2 + \lambda_v v_i^2 + \psi_i\right)\left(\lambda_w w_j^2 + \lambda_v v_j^2 + \psi_j\right)}}$. Thus we can write ρ_{ij}^2 as:

$$
\begin{aligned}
\rho_{ij}^2 &= \frac{(\lambda_w w_i w_j + \lambda_v v_i v_j)^2}{\left(\lambda_w w_i^2 + \lambda_v v_i^2 + \psi_i\right)\left(\lambda_w w_j^2 + \lambda_v v_j^2 + \psi_j\right)} \\
&= \frac{\lambda_w^2 w_i^2 w_j^2 + \lambda_w \lambda_v 2 w_i w_j v_i v_j + \lambda_v^2 v_i^2 v_j^2}{\left(\lambda_w w_i^2 + \lambda_v v_i^2 + \psi_i\right)\left(\lambda_w w_j^2 + \lambda_v v_j^2 + \psi_j\right)} \\
&= \phi_i \phi_j - \frac{\lambda_w \lambda_v \left(w_i^2 v_j^2 - 2 w_i w_j v_i v_j + w_j^2 v_i^2\right)}{\left(\lambda_w w_i^2 + \lambda_v v_i^2 + \psi_i\right)\left(\lambda_w w_j^2 + \lambda_v v_j^2 + \psi_j\right)} \\
&= \phi_i \phi_j - \frac{\lambda_w \lambda_v (w_i v_j - w_j v_i)^2}{\left(\lambda_w w_i^2 + \lambda_v v_i^2 + \psi_i\right)\left(\lambda_w w_j^2 + \lambda_v v_j^2 + \psi_j\right)}
\end{aligned}
$$

where the % shared variance of neuron i in this 2-d case is $\phi_i = \frac{\Sigma_{shared,ii}}{\Sigma_{ii}} = \frac{\lambda_w w_i^2 + \lambda_v v_i^2}{\lambda_w w_i^2 + \lambda_v v_i^2 + \psi_i}$.

Then letting m is the number of pairs in the population, and following similar steps to (10) in Math Note B, we arrive at:

$$
\begin{aligned}
E[\rho^2] &= \frac{1}{m} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \rho_{ij}^2 \\
&= (\%sv)^2 - \frac{1}{n-1} Var\left(\phi\right) - \frac{1}{m} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{\lambda_w \lambda_v (w_i v_j - w_j v_i)^2}{\left(\lambda_w w_i^2 + \lambda_v v_i^2 + \psi_i\right)\left(\lambda_w w_j^2 + \lambda_v v_j^2 + \psi_j\right)}
\end{aligned}
\tag{12}
$$

Not including the negative sign in front, note that this final term is non-negative (given that λ_w and λ_v are non-negative, as for any covariance matrix). Thus, comparing the final line in equation (12) to the final line from equation (10), we observe that the distance of the point for the 2-d case in the r_{sc} mean versus r_{sc} s.d. plot is necessarily smaller than or equal to the distance for the corresponding 1-d case.

More generally, for a k-dimensional case we can show that:

$$E[\rho^2] = (\%sv)^2 - \frac{1}{n-1} Var\left(\phi\right) - \frac{1}{m} \sum_{w,v} \left[\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{\lambda_w \lambda_v (w_i v_j - w_j v_i)^2}{\left(\lambda_w w_i^2 + \lambda_v v_i^2 + \psi_i\right)\left(\lambda_w w_j^2 + \lambda_v v_j^2 + \psi_j\right)} \right] \tag{13}$$

where the sum $\sum_{w,v}$ is taken over all unique pairs of loading vectors (w, v). Indeed, as more latent dimensions are subsequently added, the radius of the r_{sc} mean versus r_{sc} s.d. plot decreases (Fig. 5*g*). Intuitively, this final term accounts for how population activity covaries along many different dimensions in the high-d firing rate space. As more *orthogonal* dimensions are added, population activity is further pulled in different directions in the high-d space, more interaction terms come into play, and the magnitude of correlations is further decreased. This tends to decrease both r_{sc} mean and r_{sc} s.d., explaining why the radius of the arc in the r_{sc} mean versus r_{sc} s.d. plot tends to decrease as dimensionality increases.

We note that r_{sc} mean and r_{sc} s.d. do not necessarily *both* need to decrease. For example, consider a pattern with a loading similarity of 1; loading weights for all neurons would have the same value, r_{sc} across all pairs would be the same value, and thus r_{sc} s.d. would be 0 (see Math Note A). When a second pattern of necessarily low loading similarity (see Math Note E) is added, r_{sc} values across pairs of neurons would differ, and r_{sc} s.d. would be larger than 0. Therefore, r_{sc} s.d. can increase when going from the 1-d case to the 2-d case. However, the corresponding decrease in r_{sc} mean would be larger in magnitude than the increase in r_{sc} s.d., resulting in an overall decrease in arc radius (Fig. 5*g*, 1 to 2 dimensions, data points closest to the horizontal axis).

The third term in equation (13) can also help explain variability of the radius ($E[\rho^2]$) across different random instantiations with the same population metrics (Figs. 5*g* and 6). Consider a fixed %sv. For the 1-d case, the radius is determined by the first two terms of the above equation, and any variability in radius will be caused by different values of $Var(\phi)$ across different instantiations. For the 2-d case, the third term also plays a factor in determining the radius, and this term varies across different random instantiations, typically to a larger degree than the second term for large numbers of neurons n (see Math Note B). Thus, the 2-d and k-d cases have greater variability in $E[\rho^2]$ than 1-d cases (Fig. 5*g*, Fig. 6). Other subtle factors can affect the variability of $E[\rho^2]$. For example, variability in $E[\rho^2]$ can increase or decrease depending on the relative strengths of each dimension and their corresponding loading similarities (Fig. 6 and Supplementary Fig. 1). This can be explained by the third component of equation (13), in particular by the terms involving λ_w and λ_v.

E Properties of loading similarities across different co-fluctuation patterns

We asked whether there was a relationship between the loading similarities of different co-fluctuation patterns in the same model. In our simulations and V4 data analysis, we ensured that we obtain unique co-fluctuation patterns by constraining dimensions to be orthogonal. Thus, we might conjecture that if one pattern has high loading similarity (e.g., $[1, \ldots, 1]$), then another pattern in the same model necessarily has low loading similarity (e.g., $[1, -1, 1, -1, \ldots, -1, 1]$). Indeed, this is true because the sum across the loading similarities of each pattern in a model is at most 1. We show this property of loading similarity here.

Let $\mathbf{w}$ and $\mathbf{v}$ be vectors representing two co-fluctuation patterns in the same model. We use the notation $\mathbf{w} \cdot \mathbf{v}$ to refer to the element-wise product between $\mathbf{w}$ and $\mathbf{v}$, resulting in a vector that is the same size as $\mathbf{w}$ and $\mathbf{v}$. Furthermore, we use $E[\mathbf{w}]$, $Var(\mathbf{w})$, and $Cov(\mathbf{w})$

as shorthand to refer to computations across the elements of a vector (and *not* as operations on a random variable): e.g., $E[\mathbf{w}] = \frac{1}{n}\sum_{i=1}^{n} w_i$, and $Cov[\mathbf{w}, \mathbf{v}] = E[\mathbf{w} \cdot \mathbf{v}] - E[\mathbf{w}]E[\mathbf{v}] = \frac{1}{n}\sum_{i=1}^{n} w_i v_i - \left(\frac{1}{n}\sum_{i=1}^{n} w_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} v_i\right)$. Also, in this section we refer to the loading similarity of vector $\mathbf{w}$ as $ls(\mathbf{w})$ for shorthand.

We first show a constraint on loading similarities for a model with two co-fluctuation patterns (i.e. loading vectors for each dimension). Let n be the number of neurons and let $\mathbf{w}, \mathbf{v} \in \mathbb{R}^n$ be two loading vectors. As in our simulations and data analysis (see Methods), $\mathbf{w}$ and $\mathbf{v}$ are orthogonal unit vectors: $\sum_{i=1}^{n} w_i^2 = 1$, $\sum_{i=1}^{n} v_i^2 = 1$, and $\sum_{i=1}^{n} w_i v_i = 0$. Then, using these constraints,

$$\begin{aligned} Cov(\mathbf{w}, \mathbf{v}) &= E[\mathbf{w} \cdot \mathbf{v}] - E[\mathbf{w}]E[\mathbf{v}] \\ &= \frac{1}{n}\sum_{i=1}^{n} w_i v_i - E[\mathbf{w}]E[\mathbf{v}] \\ &= -E[\mathbf{w}]E[\mathbf{v}] \\ Var(\mathbf{w}) &= E[\mathbf{w} \cdot \mathbf{w}] - E[\mathbf{w}]^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} w_i^2 - E[\mathbf{w}]^2 \\ &= \frac{1}{n} - E[\mathbf{w}]^2 \end{aligned} \tag{14}$$

Because correlation is bounded between -1 and 1, we know that $|Cov(\mathbf{w}, \mathbf{v})| \leq \sqrt{Var(\mathbf{w})Var(\mathbf{v})}$. It follows that:

$$\begin{aligned} Cov^2(\mathbf{w}, \mathbf{v}) &\leq Var(\mathbf{w})Var(\mathbf{v}) \\ E[\mathbf{w}]^2 E[\mathbf{v}]^2 &\leq \left(\frac{1}{n} - E[\mathbf{w}]^2\right)\left(\frac{1}{n} - E[\mathbf{v}]^2\right) \\ 0 &\leq \frac{1}{n^2} - \frac{1}{n}\left(E[\mathbf{w}]^2 + E[\mathbf{v}]^2\right) \\ nE[\mathbf{w}]^2 + nE[\mathbf{v}]^2 &\leq 1 \\ ls(\mathbf{w}) + ls(\mathbf{v}) &\leq 1 \end{aligned} \tag{15}$$

The last step follows from the definition of loading similarity:

$$ls(\mathbf{w}) \equiv 1 - \frac{Var(\mathbf{w})}{1/n} = 1 - \frac{\frac{1}{n} - E[\mathbf{w}]^2}{1/n} = nE[\mathbf{w}]^2$$

The final inequality in equation (15) proves the intuition provided at the beginning of this section–if $ls(\mathbf{w})$ is large, then $ls(\mathbf{v})$ must be small (at most $1 - ls(\mathbf{w})$). More strongly, if $ls(\mathbf{w}) = 1$, then $ls(\mathbf{v}) = 0$.

Generally, for a model with d dimensions and patterns $\mathbf{u_1}, \ldots, \mathbf{u_d} \in \mathbb{R}^n$, we can show that $\sum_{i=1}^{d} ls(\mathbf{u_i}) \leq 1$. To see this, we can construct a matrix C with entries $c_{ij} = Cov(\mathbf{u_i}, \mathbf{u_j}) = -E[\mathbf{u_i}]E[\mathbf{u_j}]$ for $i \neq j$, and $c_{ii} = Var(\mathbf{u_i}) = \frac{1}{n} - E[\mathbf{u_i}]^2$ (derived from the constraints in equation (14)). Note that $C \in \mathbb{R}^{d \times d}$, with variances on the diagonal and covariances on off-diagonals, is a covariance matrix, which implies $det(C) \geq 0$. For a 3-d model,

$$det(C) = \frac{1}{n^2}\left(1 - nE[\mathbf{u_1}]^2 - nE[\mathbf{u_2}]^2 - nE[\mathbf{u_3}]^2\right) \geq 0$$

which implies $ls(\mathbf{u_1}) + ls(\mathbf{u_2}) + ls(\mathbf{u_3}) \leq 1$. In general, for a d-dimensional model (with $d \leq n$):

$$det(C) = \frac{1}{n^{d-1}} \left(1 - \left(\sum_{i=1}^{d} nE[\mathbf{u_i}]^2\right)\right) \geq 0$$
$$\sum_{i=1}^{d} ls(\mathbf{u_i}) \leq 1 \tag{16}$$

Equation (16) has several implications:

- If one knows the loading similarities of all dimensions $\mathbf{u_1}, \ldots, \mathbf{u_d}$ in a model, then the maximum possible loading similarity of any new dimension is $1 - \sum_{i=1}^{d} ls(\mathbf{u_i})$. It follows that two dimensions with high loading similarity cannot co-exist in the same model.
- If one dimension has $ls = 1$, then all other dimensions in the model (or that would be added to the model) necessarily have $ls = 0$. Note that there is only one possibility for a pattern to have $ls = 1$ (i.e., $\mathbf{u} = [\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}]^T$, such that $Var(\mathbf{u}) = 0$). This implies that there are many possibilities for a pattern to have $ls(\mathbf{u}) = 0$. More loosely, there are relatively few ways for a pattern to have high loading similarity, but many more ways for a pattern to have low loading similarity.

F Maximum variance of a unit vector

We defined loading similarity for a co-fluctuation pattern $\mathbf{u}$ (normalized to have norm 1) of n neurons to be $1 - \frac{var(\mathbf{u})}{1/n}$, where the variance is computed along the elements of $\mathbf{u}$. This value lies between 0 and 1 because the maximum variance across the elements of $\mathbf{u}$ is $1/n$. We now show this mathematically.

Let $\mathbf{u} \in R^n$ be a unit vector. Because $\mathbf{u}$ is a unit vector, $\sum_{i=1}^{n} u_i^2 = 1$. Using these facts:

$$\begin{aligned} Var(\mathbf{u}) &= E[\mathbf{u}^2] - E[\mathbf{u}]^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} u_i^2 - E[\mathbf{u}]^2 \\ &= \frac{1}{n} - E[\mathbf{u}]^2 \\ &\leq \frac{1}{n} \end{aligned}$$

This holds with equality when $E[\mathbf{u}] = 0$ (i.e., when the mean across the elements in a co-fluctuation pattern is 0). This implies that the smallest loading similarity is 0 (when $Var(\mathbf{u}) = 1/n$), and the largest loading similarity is 1 (when $Var(\mathbf{u}) = 0$).

3 [Control] Stabilizing neuronal activity in prefrontal cortex using a brain computer interface

The previous chapter bridged between two perspectives that measured the structure of shared trial-to-trial neuronal variability. One phenomena by which structured shared variability arises is slow drifts in neuronal population activity, which are thought to reflect slow changes in internal cognitive state. In this chapter, I present work in which we ask to what degree these slow shared fluctuations are under volitional control and can be stabilized.

3.1 Introduction

Previous studies have shown that neuronal activity varies slowly and in a coordinated manner over the course of a single experimental session (i.e., over several hours [41, 46]). The slow changes in neuronal activity are correlated with slow changes in pupil size [41, 42]. Thus, they are thought to reflect slow changes in internal states (e.g., arousal) and behaviors such as impulsivity (i.e., reaction times and false alarm rates [41]) and engagement (i.e., movement vigor [42]).

Our goal in this work was to test whether animals could volitionally modulate these slow fluctuations and stabilize neuronal activity over the course of hours (Fig. 10**a**). To ask this question, we trained two rhesus macaques to control a BCI that provided visual feedback about their prefrontal cortex (PFC) population activity. In particular, the size of an on-screen annulus was linked to the distance of the animal's neuronal activity from a "target" state. Thus, to successfully complete the BCI over the course of many trials, animals would need to 1) decrease their neuronal distance to the target and 2) keep neuronal activity relatively stable over the course of the session.

3.2 Designing a BCI to stabilize neuronal activity

We first designed a BCI that allowed animals to use PFC neuronal activity to manipulate visual computer feedback and obtain a reward (Fig. 10**b**). For visual feedback, we chose to use a centrally located annulus that expanded and contracted in size based on neuronal activity (Fig. 10**c**, green circle). The annulus was low-contrast and confined to a small window around the central fixation dot (2 degrees of eccentricity) to minimize the likelihood that it evoked responses in the PFC neurons from which we recorded. We defined a small annulus size that collapsed to the size of the fixation dot as the condition for reward.

For animals to use the BCI, they needed to understand the meaning of the annulus and associate a small annulus size with reward. We had them perform a memory guided saccade task during which we temporally linked a small annulus size with an upcoming reward. During this task, the annulus would gradually shrink to its smallest size immediately prior to presentation of the "go cue", after which animals had the opportunity to make a saccade and obtain a juice reward. On separate sessions after annulus training, we found that reaction times were faster on trials where the annulus was present than on trials where it was not (data not included here). This suggested that animals used the shrinking annulus to predict the timing of the go cue, and were better prepared to respond when the go cue was presented. With these results, we were confident that animals associated a small annulus with the go cue and an upcoming reward, and thus thought the small annulus a desirable state.

We next implanted a Utah array in dlPFC (area 8ar) and defined a brain-computer interface (BCI) by linking neuronal activity to the visual feedback of the annulus (Fig. 10**b**). Animals performed 60 calibration trials at the beginning of each BCI session, in which they fixated a central dot while an annulus gradually collapsed on the fixation dot. We used calibration neuronal activity (spike counts in 50 ms bins) to define a 4-dimensional latent state space using

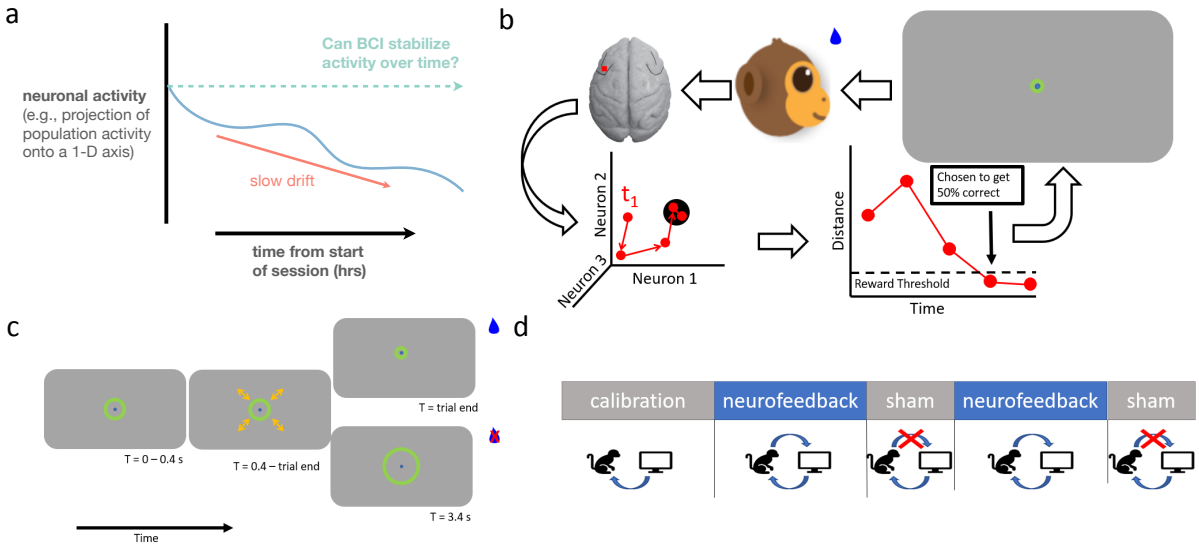


Figure 10: Neurofeedback experiment. **a**. Illustration of central question. Internal states and neuronal activity (curved blue line) can drift slowly over time (red arrow) in ways not directly related to a task at hand (e.g. due to arousal, impulsivity, satiation, etc). If we show these shifts in internal state to an animal using a Brain Computer Interface (BCI), can they use feedback to stabilize their neuronal activity over time (green dashed line)? **b**. Neuronal activity was recorded from "Utah" arrays implanted in prefrontal cortex in two rhesus macaques. The goal of the animals was to move neuronal activity (red) to a target window that was defined based on neuronal activity recorded at the beginning of each recording session. To provide feedback on the position of neuronal activity relative to the target state, distance between the current neuronal state and the target window was computed and then mapped to the radius of an annulus (green circle, upper right); a larger annulus corresponded to larger neuronal distance. To achieve reward, the animal needed to maintain neuronal activity within a predefined reward window (dashed line) for 400 ms. This reward threshold was defined using activity recorded earlier in the session in order to achieve reward on 50% of calibration trials **c**. Timeline of a BCI trial. The animal fixated a blue central circle centered on a computer monitor with a gray background and a green annulus. After 400 ms, the annulus provided continuous feedback about distance of neuronal activity from the target. If the reward criteria was met, the trial ended and the animal was rewarded. If the reward criteria was not reached within 3.4 seconds of fixation (i.e. 3 seconds of BCI control), the trial ended and the animal was not rewarded. **d**. Block structure. Each session started with a calibration block. Each trial in the block followed the sequence of events described in **c** except that the annulus shrunk monotonically until it reached the reward threshold at exactly 3.4 seconds after fixation. Neuronal activity from this calibration block was used to define the BCI mapping (factor analysis, target location, and reward threshold) between internal state of PFC activity and the visual feedback presented on the screen. After the calibration block, the task alternated between 100-trial neurofeedback blocks, with 90 BCI trials and 10 randomly interspersed "sham" trials, and sham blocks with 20 consecutive "sham" trials. "Sham" trials were used as control, or reference, trials. On "sham" trials, annulus feedback form previous sessions were replayed without any indication to the animal, meaning that the visual feedback and internal state of the animal were disassociated on these trials.

factor analysis. This procedure enabled us to capture important and intuitive dimensions of neuronal variability while discarding noise. We next defined a target state in the latent space as the average activity during the 60 calibration trials (Fig. 10**b**, large black dot in neuron state space). On BCI trials, neuronal distance from the target was mapped linearly to the size of the annulus on the screen–the smaller the neuronal distance, the smaller the size of the annulus. For a BCI trial to be rewarded, the distance of the animal's neuronal state to the target had to remain below a threshold (Fig. 10**b**, dashed line in distance vs. time plot) for 8 consecutive time bins, or 400 ms. animals therefore needed to make the annulus small to get a reward.

Thus, we designed a novel BCI paradigm in which we used visual feedback (i.e., annulus size) to show animals how far their neuronal activity was from a target state, defined as the initial activity at the beginning of the session. To obtain rewards in this BCI paradigm, animals needed to decrease neuronal distance to the target. Neuronal activity also needed to be relatively stable and close to the target throughout the session for animals to continue to get rewards during the experiment.

We used a combination of BCI trials and control trials to test whether animals were using the visual feedback in our BCI paradigm. After calibration of the BCI system, animals were given control of the visual feedback and performed BCI trials for a majority of the session. On BCI trials, animals had 3 seconds to achieve the target state and obtain a juice reward (Fig. 10**c**, top); otherwise the trial would end with no reward (Fig. 10**c**, bottom). The remainder of trials were controls to assess successful use of the BCI. We term them "sham trials" because we disassociated the visual feedback (i.e., annulus size) from the internal neuronal state. On sham trials, we replayed visual feedback from previous sessions where the animal received a reward at the last possible moment in the trial. We included two types of sham trials: BCI sham and block sham. BCI sham trials were interspersed among BCI trials but occurred rarely, meaning that animals would still be trying to control their internal neuronal state, though the feedback on the screen would not be helpful. BCI sham trials were used to assess to what extent animals were using moment-to-moment visual feedback to achieve the target. Block sham trials were isolated in a separate 20-trial block after every 90 BCI trials. Since block sham trials always lasted the full 3 seconds and were presented consecutively, we presumed that the animals' engagement decreased and they no longer tried to keep their internal neuronal state close to the target. Block sham trials, along with calibration trials, were used to assess chance level BCI performance if the animals had not been actively modulating neuronal activity.

During the BCI task, trials were organized into two types of blocks. The purpose of the first "neurofeedback" block type was to encourage the animal to use visual feedback; it consisted of 100 trials, 90 BCI and 10 BCI sham trials. The second "sham" block type was used as a control and consisted of the 20 sham trials. After the 60 calibration trials were completed, these two blocks alternated for the remainder of the session (Fig. 10**d**).

3.3 Neurofeedback reduced neuronal distance to the target

We wanted to know whether animals were able to modulate their neuronal activity on BCI trials to reach the target state. If this were the case, neuronal activity would be closer to the target (i.e., smaller distance) on BCI trials than on reference trials where BCI was not used.

We first assessed whether BCI distance to the target had decreased relative to the calibration trials that were used to define the BCI mapping. On BCI trials, we analyzed the activity on both corrects (target reached) and misses (target not reached). On calibration trials, we played neuronal activity through the BCI mapping. For calibration trials that did not reach the target, we analyzed all timepoints; for calibration trials that did reach the target, we only analyzed timepoints until the target was achieved. This guaranteed a fair comparison of distance between calibration and BCI trials. For each session, we computed the average neuronal distance to the

target on each trial and compared the distribution on BCI trials (Fig. 11**a**, gray distribution) to the distribution on calibration trials (Fig. 11**a**, green distribution). Across sessions and animals, distance on BCI trials was smaller than distance on calibration trials (Fig. 11**b**, dots fall below the equality diagonal).

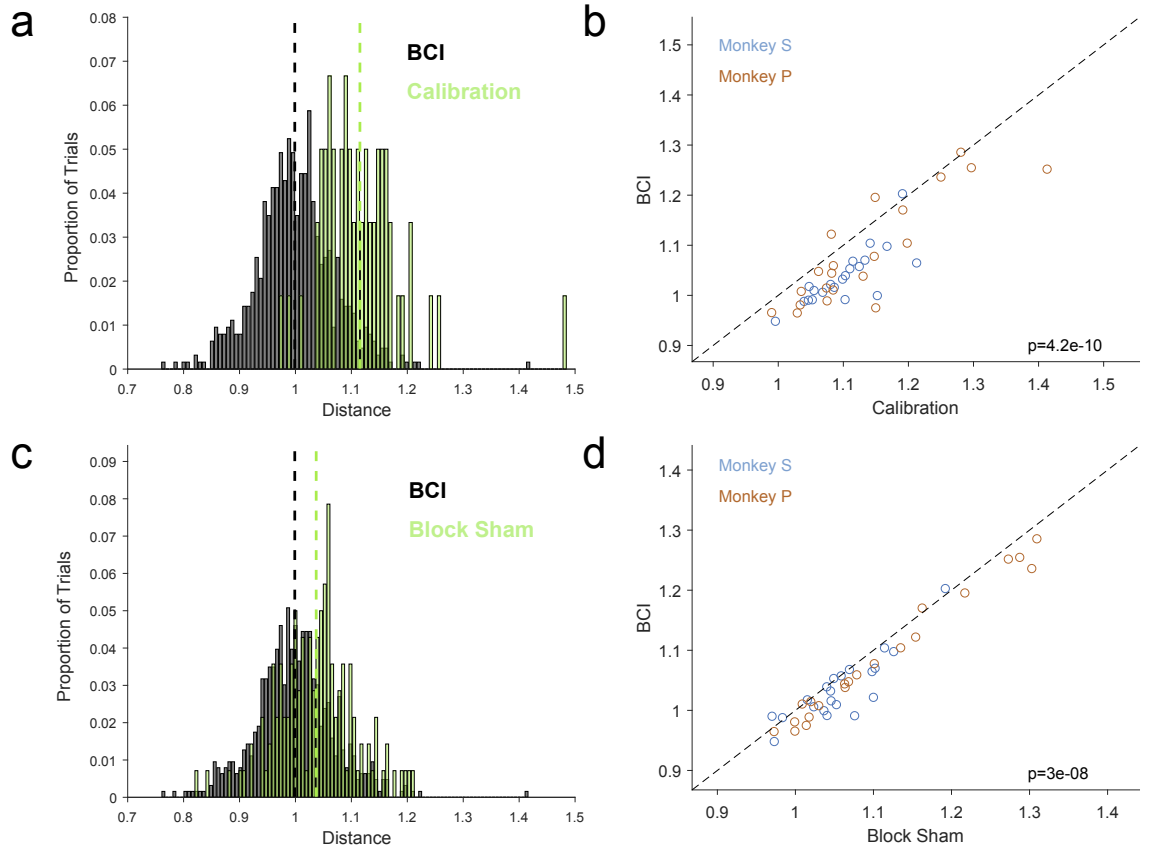


Figure 11: Distance decreases during neurofeedback a. Distribution of average distances on BCI trials (grey) and calibration trials (green) for an example session. Distances on calibration trials were obtained by playing neuronal activity through the BCI mapping offline. The mean of the each distribution is indicated by a dashed line of the corresponding color. **b**. On each session, we compared the mean BCI distance (y-axis) with the mean calibration distance (x-axis). Dashed black line indicates equality between the two values. Distance on BCI trials was significantly smaller than distance on calibration trials. **c-d**. Same as **a-b**, but for block sham trials instead of calibration trials. Block sham trials were presented in alternating blocks with BCI trials throughout the session, but feedback on block sham trials did not accurately reflect internal state. Thus, comparing BCI and block sham trials controlled for any changes in spiking statistics that might have occurred over the course of the session. Distance on BCI trials was significantly smaller than distance on block sham trials.

We next controlled for the possibility that the smaller distance on BCI trials relative to calibration corresponded to changes in neuronal activity over time. Calibration trials happened at the beginning of the session, and were only presented for a few minutes. BCI trials happened after calibration and were presented for the remainder of the session (typically several hours). Any uncontrolled changes in neuronal state (e.g. slow drift, Cowley et al., 2020) that occurred over the session might therefore bias our comparison between BCI and calibration. Thus, we also compared BCI trials (Fig. 11**c**, gray distribution) to block sham trials (Fig. 11**c**, green distribution), which were presented in alternation with "neurofeedback" blocks throughout the entire session. We played neuronal activity on block sham trials through the BCI and performed the analysis in the same manner as described above (Fig. 11**a-b**), and found that average neuronal

distance to the target was significantly smaller on BCI trials than block sham trials (Fig. 11**d**, dots fall below the equality diagonal). Together, these results showed that animals successfully decreased their neuronal distance to the target state on BCI trials.

3.4 Neurofeedback suppresses neuronal drift

Thus far, we have demonstrated that animals have used neurofeedback via our stabilization BCI to obtain a reward; they had smaller neuronal distance to the target when using BCI, as compared to when they were not using BCI. We next asked whether the successful decrease in neuronal distance on BCI trials also suppressed slow drifts in neuronal activity over time. To test this, we examined the change in distance over the course of individual blocks. To reduce noise, we first averaged the distance on trials with the same index across blocks within the same session. In other words, we took the distance from the first trial in each block and then averaged across blocks; we then repeated this process for each trial index with both BCI blocks and sham blocks. We then used linear regression to determine the slope of how neuronal distances changed on average during a BCI block (Figure 12*a*). We then aggregated results across sessions and found that there was no significant change in distance during BCI blocks in both animals (Figure 12, monkey P in panel *b*, monkey S in panel *c*). We next performed the same analysis for sham blocks and found that distance increased significantly in both animals (Figure 12), example session in panel *d*, aggregated slopes in panels *e* and *f*). Taken together, these results show that: 1) slow drift in PFC neuronal activity existed in our data (i.e., during the sham blocks when neural distance increased), but 2) neurofeedback via use of the stabilization BCI suppressed slow drift (i.e., on BCI blocks when neuronal distance did not increase).

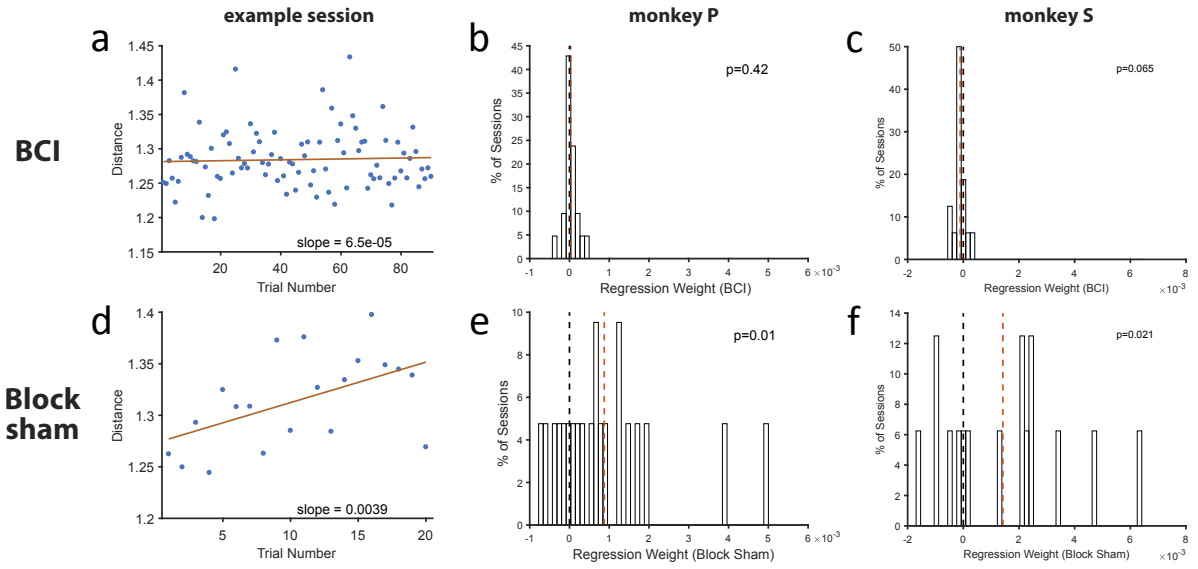


Figure 12: Neurofeedback suppresses neuronal drift. **a**. Changes in neural distance over the couse of a BCI block on an example session. We computed the average distance across blocks within each session for each trial index within the block. We fit a linear regression to measure the within-block change (i.e., the slope of the regression line) in distance across the block. **b**. Histogram of regression slopes on each session for monkey P. Dashed black line indicates 0 and dashed red line indicates the average slope across sessions. The slope was not significantly different from 0, implying that distance did not change significantly within BCI blocks. **c**. Same as **b**, but for monkey S. Slope was not significantly different from 0. **d**-**f**. Same as in **a**-**c**, except for sham blocks. The slopes on sham blocks were significantly positive for both animals, indicating that neuronal distance to the target increased during sham blocks.

3.5 Discussion and future directions

In this work, we designed a novel brain computer interface (BCI) for prefrontal cortex (PFC) with the goal to stabilize neuronal activity over time. To successfully obtain a reward, subjects had to keep their internal neuronal state (i.e., population firing rate vector) close to a target state defined at the beginning of each session. We showed that, by using the BCI, subjects: 1) were successfully able to reduce the distance of their internal neuronal state to the target state, and 2) suppressed slow neuronal drift.

Slow drift in neural activity has been linked to slow changes in pupil size in previous studies, which is often thought to reflect arousal and engagement [41, 42]. In our work, we showed that using the BCI suppressed slow neuronal drift (Fig. 12). We hypothesize that the decrease in slow drift associated with using our BCI (Fig. 12) might also correspond to a decrease in the slow fluctuations of pupil size. If true, this would support the interpretation that our BCI not only stabilizes neuronal activity, but also the animal's internal cognitive state (i.e., arousal, engagement, or wakefulness). Future work will test this hypothesis.

How precisely do animals reduce their neuronal distance to the target state on rewarded BCI trials? There are several strategies that subjects could have used to successfully obtain reward on BCI trials. For example, they could have kept their internal starting point (i.e., neuronal activity at the beginning of each trial) close to the target state–a result of control of slow-timescale (on the order of seconds to minutes) variability over many trials. Alternatively, subjects could have decreased neuronal variability around the target state *within* each trial–a result of control of fast-timescale variability (on the order of several tens or hundreds of milliseconds). Or subjects could have used a combination of both strategies. Our analyses thus far have shown that, on average, distance is smaller on BCI trials than on sham trials, providing evidence for strategy 1. Future work will test strategy 2: whether the within-trial variability of neuronal activity (e.g., spike count variance, spike count correlations, and population metrics) is different on BCI trials than sham trials. Answering these questions will elucidate what aspects of PFC neuronal variability (e.g., fast vs slow timescale, shared vs independent variance) are under volitional control.

3.6 Methods

Task: Overview and motivation

The subject was required to perform two tasks: a calibration task and a brain-computer interface (BCI) task. The data collected during the calibration task was used to fit the parameters that were subsequently used during the BCI task to map neural activity to feedback. The calibration task consisted of 60 trials with sham feedback. Neural activity from the calibration period was used to train the final BCI mapping of neural activity to annulus size. After calibration, the subject performed BCI trials in alternating BCI and sham blocks (Fig. 10*d*). The BCI block consisted of 90 BCI trials and 10 sham trials. The sham blocks consisted of all sham trials. The sham trials consist of feedback inconsistent with the current neural state, but were realistic feedback in that we replayed feedback from a trial in a previous session.

Task: Details

In the calibration task, the subject was required to passively fixate a blue dot at the center of a grey screen. After fixation, a green annulus appeared on the screen. During the first 20 trials, the annulus was fixed in size. During the subsequent 60 trials, the annulus moved on the screen after a 400 ms delay. The movement continued throughout a 2.5 to 3 s wait period, after which the annulus and the fixation dot were removed from the screen. Near the end of the wait period,

the annulus converged toward the center, indicating that the trial was about to end. The subject was rewarded after successfully maintaining fixation through the entire wait period.

The BCI task was identical to the calibration task except that the size of the annulus was controlled by the recorded neural activity. If the neural activity entered a state associated with a small annulus, then the trial would end and the subject would be rewarded. To receive a reward, the annulus needed to remain below a pre-determined threshold for 400 ms. The details of the calibration algorithm, the mapping of neural activity to the annulus size, and the setting of the threshold are described in the next section.

During the BCI task, trials were organized into blocks as follows. Each block consisted of a specific ratio of BCI trials and sham trials. Sham trials were defined as trials in which annulus size from missed trials from a previous session were played as feedback rather than the true feedback based on the current neural activity. Since these were missed trials and lasted for the maximum trial length, all sham trials were of the same duration. The first block consisted of 100 BCI trials. The purpose of this block was to encourage the subject to use the feedback, since all feedback presented was valid. The second block consisted of 100 trials, of which 90 BCI trials were valid and 10 trials were sham trials. The third block consisted of 20 sham trials. The second and third blocks alternated throughout the session after the first block was completed.

BCI calibration

One key decision point in our design was whether we would require the subject to stabilize neural activity in the full neural space or in a low-dimensional latent subspace. We identified two major problems with using the full neural space. First, the BCI would be highly sensitive to any array instability. If a single neuron fired at a low rate during the calibration period and then suddenly fired at a high rate later in the session, then the BCI feedback would become very difficult to control. A low-dimensional latent BCI mapping would be more robust to these instabilities. Second, assuming Poisson-like spiking variability, it can be shown that the optimal strategy in the full neural space is to reduce the firing rate of all neurons in the population. Intuitively this is because any neurons that happen to have a large spike count in a given bin will have a large adverse effect on the BCI performance. Reducing the global firing rate would reduce the probability of the detrimental high spike count instances. BCI mappings that allow a firing-rate reducing strategy are also highly subject to large scale array instabilities. For example, an instability that produces an average drop or rise in firing rate would result in a large increase or decrease in BCI performance, respectively. In contrast, since a low-dimensional latent consists of a linear combinations of units across the population, a high spike count for one unit may not adversely affect the mapping, depending on what the rest of the population is doing. For these reasons, we decided to require the subject to control neural activity in a factor analysis latent space [99]. Previous studies have similarly employed linear combinations of neural activity to address these issues [100, 101].

Calibration was performed as follows. We first performed a light sorting using a neural network sorter (described in more detail in "Neural network sorter" below) to remove noise (e.g., movement artifacts). We next binned spike counts into non-overlapping 50 ms bins beginning 400 ms after fixation to the end of the wait period. We aggregated spike counts across trials and applied factor analysis (FA; see "Factor analysis" below for details) to the aggregate spike count matrix to identify a subspace that explained population covariance structure [102–104]. All sessions used either a dimensionality of 4 or 5 for the latent subspace. After fitting FA, we computed the posterior mean of the latent variables and smoothed the latents using an exponential smoother with a time constant of 300 ms (i.e., 6 time bins). To determine the distance threshold that would achieve a reward, we computed the distance of the smoothed latents from the calibration mean. We then aggregated all distances and computed percentile

in 0.1 percentile increments. We swept percentile values to determine what percentile threshold would achieve reward on 50% of the calibration trials. The value of 50% was used to balance the need to motivate learning with the need to motivate the subject to continue performing the task. This also helped normalize the subject's initial BCI performance across sessions.

Factor analysis

As an additional denoising step prior to providing BCI feedback, we projected neural activity into a low dimensional subspace using factor analaysis, or FA [50, 99]. Factor analysis is defined as:

$$\mathbf{x} \sim \mathcal{N}(\mu, LL^T + \Psi) \tag{17}$$

where $\mathbf{x} \in \mathbb{R}^{n \times 1}$ is a vector of spike counts across the n simultaneously-recorded neurons, $\mu \in \mathbb{R}^{n \times 1}$ is a vector of mean spike counts, $L \in \mathbb{R}^{n \times m}$ is the loading matrix relating m latent variables to the neural activity, and $\Psi \in \mathbb{R}^{n \times n}$ is a diagonal matrix of independent variances for each neuron. In our BCI, the number of latent variables was always set to either 4 or 5, depending on the session. The model parameters μ, L, and Ψ were estimated using the expectation-maximization (EM) algorithm.

BCI feedback

To map neural activity to annulus radius, we performed a similar procedure as was done during calibration. Briefly, we sorted spikes from the previous 50 ms using our neural network sorter, projected the resulting spike count vector into the calibration-defined factor space, updated the exponential smoother, and then mapped the smoothed projection to a percentile value. This percentile value was then mapped to annulus size using a predefined affine transformation. Annulus feedback was updated every 50 ms.

Neural network sorter

To separate waveforms likely to be caused by neural spiking from waveforms caused by other electrical artifacts, we developed a neural network classifier that labeled spike waveforms as "neural" or "noise". The classifier was trained using array recordings from multiple animals in which the waveforms had been hand sorted. Classification required very little computation time, allowing for the classification of hundreds of waveforms in a few milliseconds. We therefore applied this algorithm, both during training of the BCI mapping, and also online during the BCI task to help ensure that activity going into the BCI was of neural origin. Details of this neural network sorter can be found in Issar et al. (2020) [105].

4 [Sources] Local and global sources of coordinated neuronal variability in prefrontal cortex

Chapters 2 and 3 focus on neuronal variability within one brain area in one hemisphere of cortex. However, one might imagine that neuronal variability in one area of cortex may be shared with another area (e.g., an input or output area), or be due to brain-wide signals that impact many areas (e.g., arousal, impulsivity). In this work, I present research that utilizes bilateral neuronal recordings and develops a new method to identify and separate global and local sources of shared neuronal variability.

4.1 Introduction

Variability in neural activity has been shown to have significant effects on the ability of groups of neurons to encode information about sensory inputs [37–40, 106], motor outputs [107, 108], decisions [106, 109, 110], attention [1, 46, 77], and other processes. This is especially true when variability is shared among neuron in a given population [111]. Most work studying neural variability has done so in populations of neurons confined to a single brain region. It is therefore unclear to what extent variability shared among neurons in the local populations previously studied was also shared with neurons in other brain regions. Here we leverage multi-area recordings to separate the study of variability shared among neurons in distant brain regions from variability shared only among neurons in a single brain region.

Neural activity observed within a brain area may be generated within an area, come from another area, or may be shared across many areas. For example global shared signals might reflect large changes in visual input (e.g., luminance shifts), tonic arousal changes [41, 112], spontaneous behaviors [113], or top-down feedback [114]. On the other hand, local shared signals might reflect local tuning similarity [115, 116], spatial scales of connectivity [117], or local computations [111]. It is important to be able to separate these local and global scales of shared variability in order to properly study these distinctive cognitive processes.

The majority of previous work that has focused on interactions between brain areas has largely utilized imaging [118, 119], local field potential [120], or EEG [121]. A few recent studies have investigated between-area interactions using spiking activity of tens of neurons in different areas, typically within the same hemisphere of the brain [41, 77, 122]. However, some research in monkey motor cortex [63] and ALM/premotor cortex in mice [123], has used spiking activity to study interactions between neurons in two different hemispheres of the brain. Other work has investigated across-hemisphere shared variability in V4 neurons during an attention task [1, 46, 124].

Ideally, neural processes with distinct mechanisms could be studied independently, however because multiple processes can influence groups of neurons it has not been obvious how to separate the neural signals that should be attributed to each process. One approach to studying shared variability is to use dimensionality reduction methods, such as factor analysis (FA), which allow for the separation of variability attributed to a single neuron from variability shared with other neurons in a population [99, 102, 104, 125]. However these methods do not provide a mechanism for separating variability shared among neurons in one population from variability shared between two distinct populations. Other dimensionality reduction approaches do consider shared interactions between brain areas [122, 126, 127]. One such example is probabilistic canonical correlation analysis (pCCA), which finds dimensions of maximum correlation to identify variability that is shared between two distinct brain areas. However this method does not separate the remaining variability shared among neurons in a single population from independent neural variability. Given the inability of these methods to separate shared variability into across-area and within-area components, another approach is needed to study these two types

of variability separately.

In this work we developed a novel method called pCCA-FA (i.e., a combination of probabilistic canonical correlation analysis and factor analysis), for separating within-area and across-area interactions. This method combines the advantages of FA and pCCA into a single probabilistic framework. We applied this method to bilateral multielectrode array recordings in prefrontal cortex during a standard visuo-spatial working memory task. We found that many pairs of neurons across hemispheres have large correlations (both positive and negative). To further partition within-area and across-area sources of shared variability, we developed a new model, called pCCA-FA, and applied it to our bilateral PFC population recordings. We found that both across-hemisphere and within-hemisphere interactions represented a large portion of shared variability. Furthermore, across-hemisphere latent projections predicted pupil size, a signal thought to be associated with global cognitive phenomena such as arousal or wakefulness. On the other hand, within-hemisphere latent projections were not predictive of pupil. Taken together, our results demonstrated that substantial shared variability exists between neuronal populations in different hemispheres of the brain and that this variability likely reflects global cognitive processes.

4.2 Simultaneous bilateral recordings of PFC population activity

In order to study both the global shared fluctuations across hemispheres of cortex and local shared fluctuations within a single brain area, we simultaneously recorded population activity from PFC in both hemispheres while subjects performed a working memory task (Fig. 13*a*; insets show rasters from the delay period of an example trial). Inspecting the rasters, we can observed that there is a shared increase in spiking activity across many neurons in both left and right PFC at around 600 ms. The key question in this work is: can we identify and partition the shared trial-to-trial co-fluctuations that are global and present across hemispheres from the shared co-fluctuations that are local and only present among the neuron in one of the brain areas (Fig. 13*b*)?

To answer this question, we first measured the spike count correlation (r_{sc}) distributions for pairs of neurons within the same PFC and for pairs of neurons across different hemispheres. There are many pairs of neurons in both the within-area r_{sc} distribution and across-hemisphere r_{sc} distribution with large magnitude (both positive and negative) and significant correlations (Fig. 13*c*). When we asses the mean r_{sc} of these distributions, a commonly-used metric [7], we found that within-area r_{sc} mean was larger than across-hemisphere r_{sc} mean (Supp. Fig. 8*a*). However, r_{sc} mean is a coarse metric that averages across the many large magnitude positive and negative correlations observed in Fig. 13*c* [104]. By dissecting r_{sc} further (instead of computing the mean r_{sc} across the distribution), we found that there is a relationship between the r_{sc} of a pair of neurons and their signal correlation (i.e., tuning to the target location in the working memory task; see Methods). This was true for both within-area and across-hemisphere pairs of neurons (Supp. Fig. 8*b*).

4.3 pCCA-FA partitions across-area and within-area shared variability

To better characterize the shared fluctuations of neurons within and across areas, we sought a computational method that would leverage activity across the entire population of recorded neurons to allow us to separate within and across-area shared trial-to-trial variability. One powerful approach to leveraging the activity of a population of neurons is dimensionality reduction, which seeks to explain population variability using a relatively small number of latent variables [50].

One commonly used dimensionality reduction method called factor analysis (FA) has been used to measure within-area shared variability [99, 102–104]. An important feature of FA is that it separates variability shared among neurons in the population from variability private to

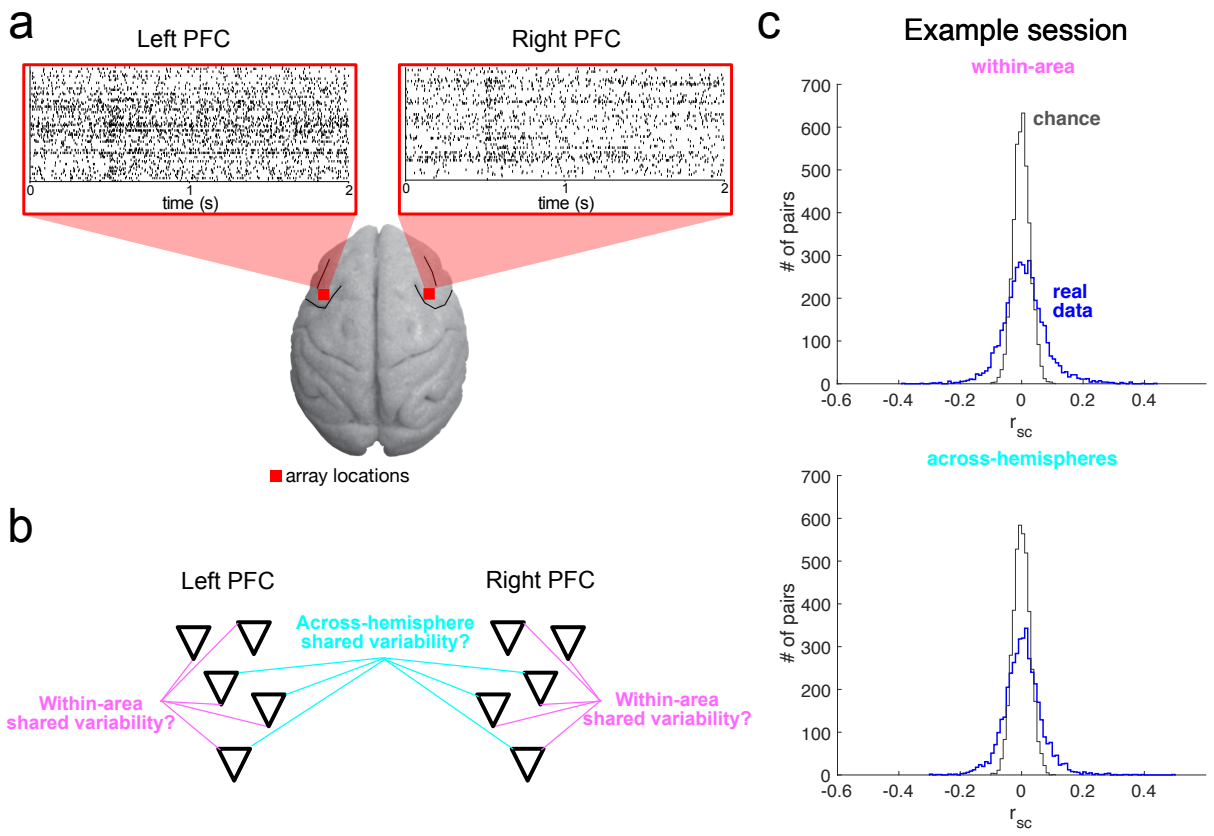


Figure 13: Trial-to-trial neuronal variability within vs. across areas. a. Recording setup. We recorded from PFC in both hemispheres using 96-channel Utah arrays while subjects performed a visual working memory task. Raster plots show spiking activity during the delay period of one example trial. **b**. The key question this study aims to answer: can we separate trial-to-trial variability that is shared among neuron across areas/hemispheres (cyan) from that which is shared among neurons within the same brain area (magenta)? **c**. The r_{sc} distributions for within-area pairs (top) and across-area pairs (bottom) in one example session. There are many pairs of neurons in both distributions with large and significant correlations (blue real data histogram extends beyond the gray chance distribution). Chance distributions are generated by computing r_{sc} distributions on data with randomly shuffled trials.

each neuron. However, FA does not partition within and across-hemisphere shared variability. Another dimensionality reduction method, probabilistic canonical correlation analysis (pCCA) has been used to find dimensions that have the most correlation between two brain areas [126]. However, pCCA does not partition within-area shared variability from variability independent to each neuron.

To facilitate the separation of across-area, within-area, and independent neural variability, we developed a new dimensionality reduction method called pCCA-FA (probabilistic canonical correlation analysis–factor analysis) to jointly model neural activity in each PFC with: 1) dimensions that capture trial-to-trial variability shared between neurons across areas (Fig. 14*a*; global, cyan) and 2) latent variables that are private to each area/hemisphere to capture trial-to-trial variability shared between neurons within the same area (Fig. 14*a*; local, magenta). The pCCA-FA model also accounts for variability that is independent to each individual neuron, which we term independent variance (Fig. 14*a*; black).

The pCCA-FA model is defined as a probabilistic graphical model (see Methods). One group of latent variables (Fig. 14*b*; z, defined by across-area global dimensions) contribute to shared

variability in both areas, while another group of latent variables (Fig. 14*b*; z_x, z_y, defined by within-area local dimensions) only contribute to shared variability in their respective brain areas (area X, or area Y respectively).

Through the lens of covariance matrix estimation and decomposition, pCCA-FA decomposes the full-rank covariance of the two PFC populations into a sum of 3 matrices: a low-rank across-area (e.g., global) covariance matrix, a low-rank within-area (e.g., local) covariance matrix, and a diagonal independent neuron covariance matrix (Fig 14*c*, top). The within-area covariance matrix is block diagonal, as it does not explain shared co-fluctuations between neurons across areas (i.e., it does not contribute to cross-covariance between two brain areas).

Using this decomposition, we investigated the characteristics of shared variability in across-area and within-area components. To assess the strength of shared variability, we computed the percent shared variance [%sv, 102, 104]. To do so, for each neuron we compute the amount of variance explained by a given component divided by the total variance of the neuron (Fig. 14*c*, lower left; see Methods). We then report the average %sv across neurons in a brain area. We also assessed dimensionality by computing d_{shared} [102, 104], which is defined as the number of dimensions required to explain 95% of the variance in the matrix of interest (Fig. 14*c*, lower right). We report d_{shared} and %sv separately for within (local) and across-area (global) components, and also separately for left and right hemisphere PFC populations.

4.4 pCCA-FA successfully recovers ground truth in various settings

To validate our model, we compared the ability of pCCA-FA to characterize across-hemisphere d_{shared} and %sv in simulations in which the ground truth was known. We randomly generated ground-truth pCCA-FA model parameters and simulated data for 30 neurons in each area of two brain areas from the pCCA-FA generative model (Fig. 14*b*; see Methods). The global and local components were designed to have a fixed number of dimensions and percent shared variance across simulation runs ("ground truth" d_{shared} and %sv). We then fit pCCA-FA to the simulated data by using 10-fold cross-validation to jointly select the across-area and within-area dimensionalities. We asked how well pCCA-FA was able to recover the ground truth dimensionality (d_{shared}) and %sv.

We found that pCCA-FA identified both the ground truth global (across-hemisphere) d_{shared} and %sv reliably with only 300 trials/samples (Fig. 15*a* left and *b* left). Importantly, we found that pCCA-FA required fewer trials than pCCA to recover d_{shared} (300 compared to 600 for pCCA; Fig. 15*a* left). Additionally, pCCA always underestimated the global %sv, even with a large number of trials (Fig. 15*b* left; see Methods and Supp. Fig. 9). The pCCA-FA method was also able to correctly identify within-hemisphere d_{shared} and %sv (Fig 15*a* right and *b* right). Note that since pCCA does not model within-hemisphere variability, we could not assess the ability of this model to identify within-hemisphere d_{shared} and %sv.

In the previous analyses, we fixed the global and local d_{shared} and %sv and asked how many trials were needed to recover ground truth. We next fixed the number of trials, and asked whether pCCA-FA could recover the ground truth under various settings of d_{shared} and %sv. We found that pCCA-FA was able to identify the ground truth d_{shared} and %sv at a variety of ground truth settings (Fig. 14*c*), both when global was larger than local and when local was larger than global. Taken together, these results demonstrate that pCCA-FA is able to identify and properly partition global (across-hemisphere) and local (within-hemisphere) shared variability, even in very data-limited settings and across a variety of ground truth settings.

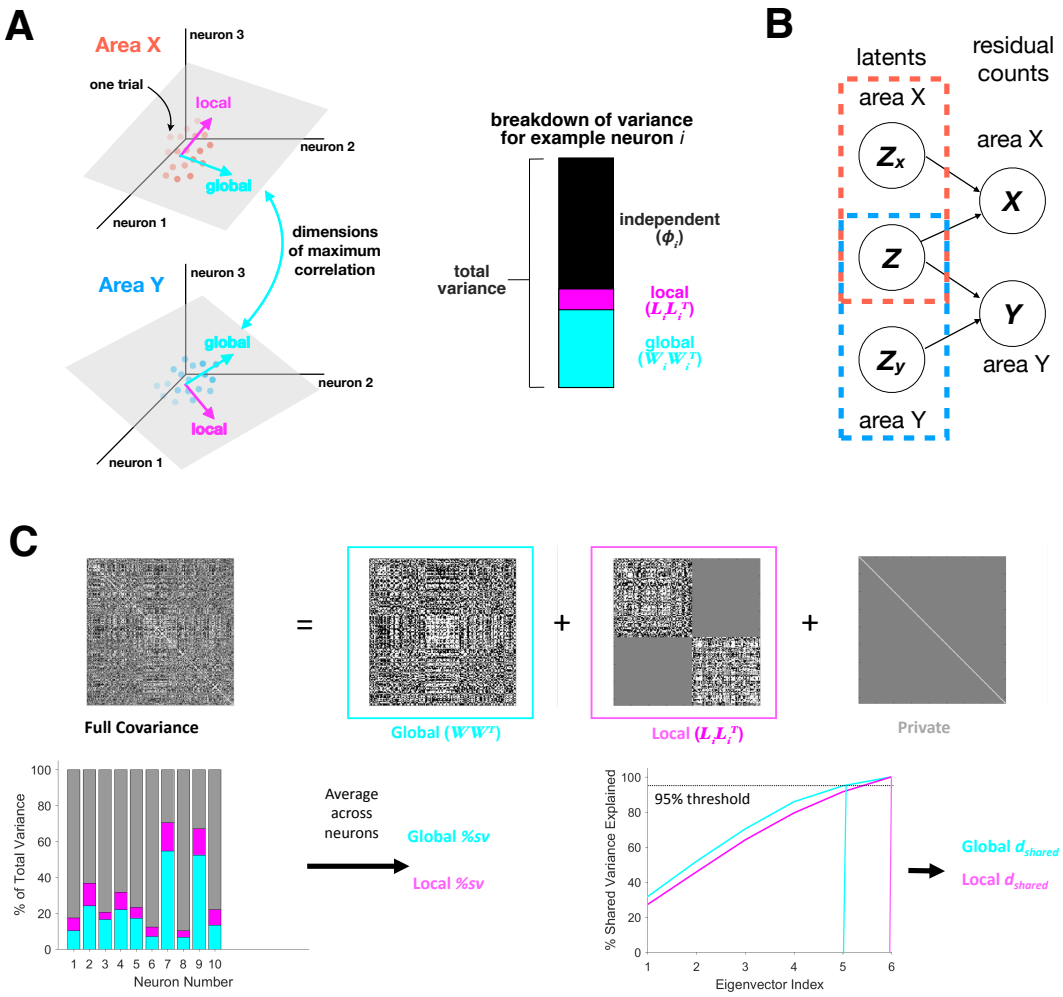


Figure 14: The pCCA-FA model partitions global and local shared variability. **a**. Left: visual representation of how pCCA-FA finds low-dimensional local (within-area) and global (across-area, or across-hemisphere) subspaces. In this illustration, transparency indicates correspondence between samples in area X and area Y. Global dimensions are those that are most correlated across areas (i.e., projections onto "across" arrow are highly correlated between area X and Y). Local dimensions explain dimensions of large covariance in neurons within the same area, which are not correlated across areas. Right: visual representation of how pCCA-FA partitions a neuron's variance: shared global variability, shared local variability, and independent private variability components. **b**. pCCA-FA graphical model. Global latent variables (z) contribute to variability in both brain areas. Local latent variables (z_x, z_y) only contribute to variability in their respective brain area. The distributions that define this graphical model are available in Methods. **c**. Top: visual representation of how pCCA-FA partitions a covariance matrix. The full-rank empirical covariance matrix is decomposed as the sum of a low-rank global covariance, a low-rank local covariance, and a diagonal private covariance. Bottom: illustration of how important metrics of fitted pCCA-FA models are computed. We evaluate the strength of shared variability using %sv and the dimensionality using d_{shared}, for both global and local subspaces.

4.5 Extracting fast-timescale trial-to-trial variability

Previous work has shown that neural activity can covary quickly from moment-to-moment and trial-to-trial [1, 99], but also more slowly over the course of many trials or the entire session [41, 42, 46]. Indeed, one might think of neuronal covariability as containing fast trial-to-trial component riding on top of a slow multi-trial component. Covariance and correlation matrices

computed directly on raw spike counts reflect both fast and slow co-fluctuations [128].

We separate these two timescales of covariation and study them separately with the assumption that they reflect distinct cognitive processes. Slow timescale co-fluctuations have been associated with arousal, impulsivity, and engagement [41, 42]. However, caution must be taken when studying correlations in slow processes due to autocorrelation and limited data. Failure to due so can result in spurious, large-magnitude correlations (Supp. Fig. 10). To simplify our analyses in this study, we removed the slow component from the raw spike counts and focus most analyses on the fast component (though see Supp. Fig. 13). We first identified the slow component using a moving average of 25 trials on each neuron's spike counts. We computed the fast component as the residuals–by subtracting the slow component from the raw spike counts of each neuron (Supp. Fig. 11). Previous work has shown that these faster-timescale trial-to-trial co-fluctuations are thought to limit the fidelity of sensory encoding [37–40].

4.6 Across-hemisphere shared variability is substantial, and often larger than within-area shared variability

We asked to what extent fast trial-to-trial variability is shared across hemispheres vs. within areas in PFC population activity. We consider spike counts computed in a one second window at the end of the delay period of each trial, and mean center the counts within each target condition. We then extract the fast processes for each neuron as described above. We fit pCCA-FA to these fast neural processes using 10-fold cross-validation to jointly choose dimensionalities for the three subspaces (across-hemisphere, within left PFC, and within right PFC) and then compared d_{shared} and %sv for within-area versus across-hemisphere subspaces. We found that our pCCA-FA model provided better fits to our neural data than alternative approaches and models that we considered (Supp. Fig. 9).

We found that across-hemisphere (global) d_{shared} and %sv were often of a similar magnitude or significantly greater than within-area (local) d_{shared} and %sv (Fig 16*a*-*b*). Pooled across subjects, sessions, and left and right hemisphere PFC, both d_{shared} and %sv were significantly larger for across-hemisphere shared variability than within-area shared variability. This stands in contrast to the mean pairwise correlation results in Supp. Fig. 8*a*, in which we found substantially less mean r_{sc} for across-hemisphere pairs than within-area pairs. Moreover, we found that the most correlated dimensions across hemispheres also explained the most shared variance (Supp. Fig. 12), which did not have to be the case as CCA can pick up on dimensions that have high correlation but low variance.

We also applied pCCA-FA to the slow component of neural activity removed earlier and compared the amount of the slow activity assigned to the global component to that of a control chance level. We found that the slow activity had higher canonical correlations in the across-area component of shared variability than expected by chance, indicating that a significant amount of the slow activity likely represents global processes (Supp. Fig. 13). Overall, these results show that a large proportion of trial-to-trial variability is shared across hemispheres of cortex.

4.7 Across-hemisphere latent variables predict pupil size

We next wanted to assess the behavioral relevance of the global across-hemisphere and local within-area components. One possibility was that the across-hemisphere component is related to latent variables that modulate activity in many areas. Such variables may be related to a variety of cognitive processes including arousal, impulsivity, engagement, satiation, and others [41, 42, 130]. One variable that has previously been used to indirectly measure these processes is pupil size. Previous work has linked large-scale cognitive processes with neural activity at slow timescales [41, 42] using pupil diameter as an indirect measure of these processes. Given that

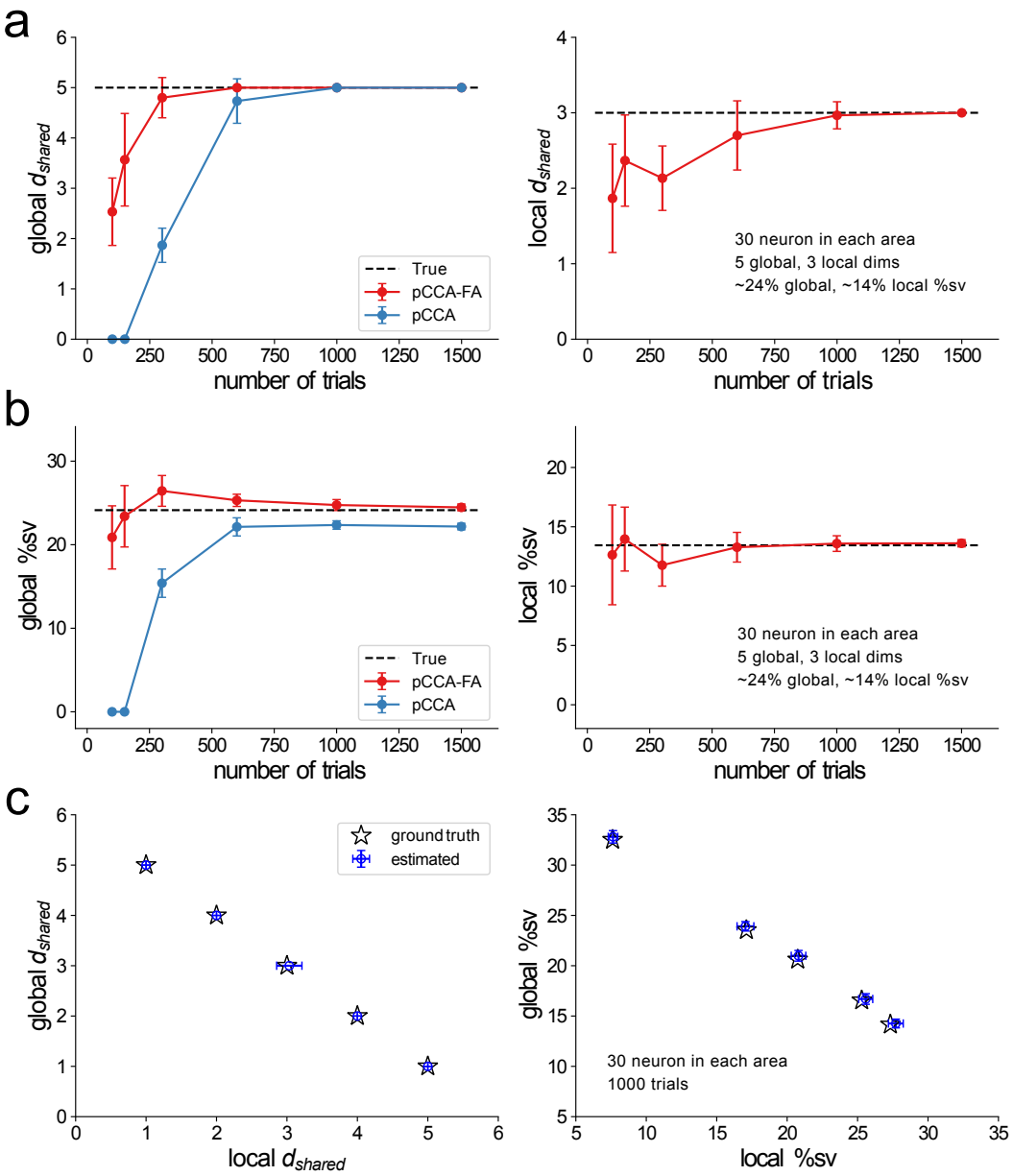


Figure 15: pCCA-FA recovers ground truth %sv and dimensionality. a. Recovery of ground truth dimensionality (d_{shared}). We asked how well pCCA-FA and pCCA could recover ground truth. We swept the number of trials (horizontal axis) to test the models under different data limitations. We crossvalidated pCCA to select global dimensionality, and crossvalidated pCCA-FA to jointly select global and local dimensionalities. Left: global dimensionality. pCCA-FA recovers ground truth global d_{shared} with relatively few (300) trials, and is more efficient that pCCA which requires 600 trials to recover ground truth. Right: local dimensionality. pCCA-FA is able to recover ground truth local d_{shared}; more trials are needed here as compared to recovering global d_{shared} because the local %sv ($\approx 14\%$) is smaller than the global %sv ($\approx 24\%$) in this simulation. pCCA does not have a concept of local dimensionality, and therefore has no data in this figure. In this and subsequent figures, error bars indicate 1 standard deviation, computed across 30 separate simulations. (continued on next page...)

this previous work focused on slow processes in a single brain area, it was unclear to what extent faster-timescale trial-to-trial co-fluctuations in neural activity reflects these same processes. Further, it has been assumed that the aspects of neural variability related to pupil are multi-area

Figure 15 ***(previous page)*****:** (continued from previous page...)

b. Recovery of ground truth %sv. Same simulations and fitting procedures as in **a**, but now showing recovery of global %sv in the left panel, and local %sv in the right panel. Left: both pCCA-FA and pCCA underestimate global %sv with very limited data because dimensionality is underestimated (panel **a**). However, there is a small regime (300 trials and 600 trials) where pCCA-FA overestimates global %sv. This is because, in general, eigenvalues of covariance matrices tend to be overestimated in high-dimensional regimes where the number of trials (i.e., samples) is small relative to number of neurons (i.e., features) [129]. However, estimates of global %sv improve with more trials. On the other hand, pCCA asymptotes and underestimates global %sv even with increasing data (see Methods and Supp. Fig. 4 for details on the shortcomings of pCCA). Right: pCCA-FA recovers ground truth local %sv with relatively few trials. Again, pCCA does not have a way to separate shared local variability and private neuron variability and therefore has no data in this figure. **c**. In **a** and **b**, we asked how many trial pCCA-FA needed to recover ground truth for a given setting of global and local shared variability. Here, given a reason number of trials (1000 trial, 30 neurons per area), we evaluate whether pCCA-FA can recover ground truth for various settings of global and local shared variability (i.e., %sv and dimensionality). Left: pCCA-FA can recover ground truth global and local dimensionality across various settings (blue circles are estimates, black starts are ground truth). Right: pCCA-FA can recover ground truth global and local %sv across various settings. Stars in left and right panels correspond to the same ground truth parameter settings, and error bars in estimates indicate 1 standard deviation computed across 30 simulations.

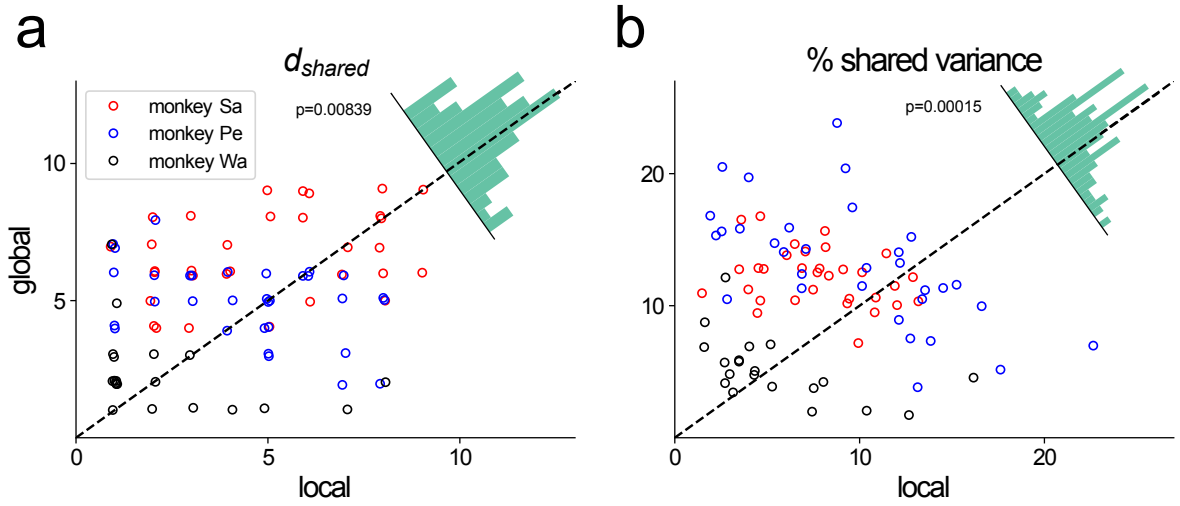


Figure 16: Global shared variability is substantial, and often larger than local shared variability. a. d_{shared} for pCCA-FA fits to "faster-timescale" neural activity. Results from each hemisphere and session is aggregated per monkey. Global (across-hemisphere) d_{shared} is larger than local (within-area) d_{shared} (pooled $p = 0.008386$; Sa $p = 0.000403$, Pe $p = 0.317343$, Wa $p = 0.941009$; paired sample t-test). **b**. Percent shared variance (%sv) for pCCA-FA fits to neural activity. Global %sv is larger than local %sv (pooled $p = 0.000148$; Sa $p = 0.000001$; Pe $p = 0.032285$; Wa $p = 0.801285$; paired sample t-test). In both **a** and **b** , histograms show the difference between local and global metrics.

processes. However, given that previous studies did not record from multiple brain regions, it was unknown whether this assumption was valid. Here we leveraged our two-hemisphere recording paradigm coupled with our pCCA-FA model to address these gaps (Fig. 17*a*).

We computed the global across-hemisphere and local within-area latent variables (see Methods) and used them to predict pupil size using linear regression. Qualitatively, we found that prediction of pupil diameter was robust for across-hemisphere latents but absent for within-area

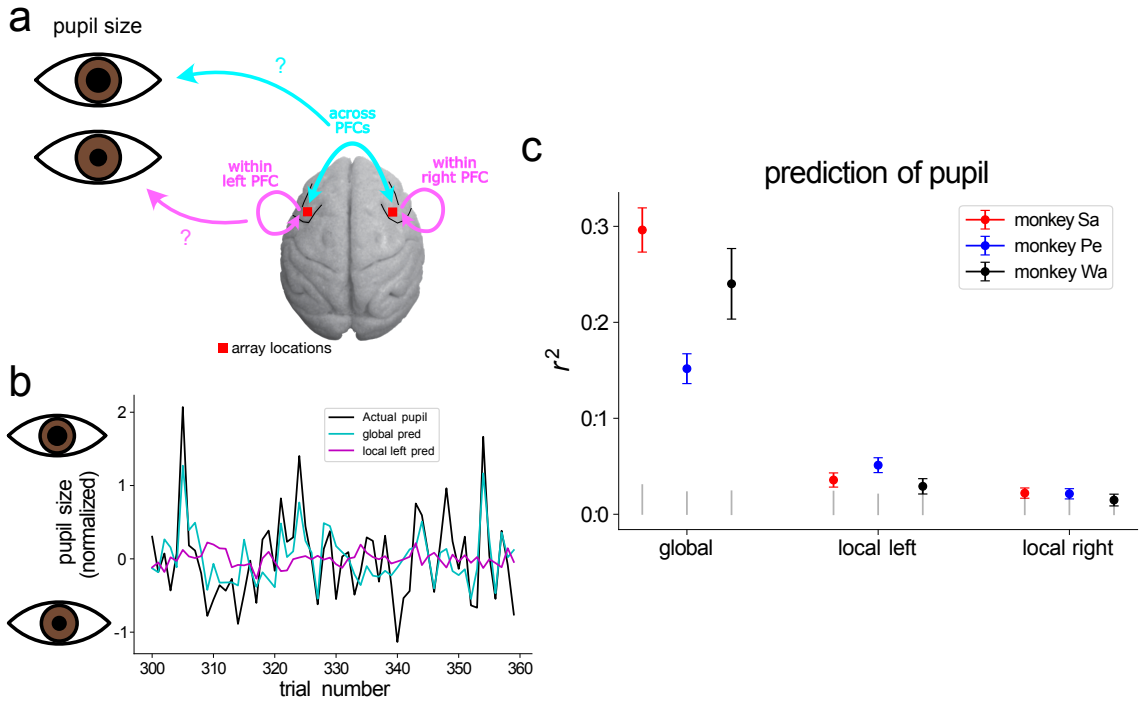


Figure 17: Global latent variables are predictive of pupil size, which is thought to reflect global cognitive phenomena such as arousal and wakefulness. a. Are the global (across-hemisphere) latents or the local (within-hemisphere) latents extracted from neural recordings in PFC predictive of pupil size? **b**. Example of 60 trials and "faster-timescale" fluctuations in pupil size (black). Prediction of pupil size using "faster-timescale" global latents (cyan) and local latents (magenta). For this session, global latents predict pupil size ($r^2 = 0.365$) better than local latents predict pupil size ($r^2 = 0.044$). **c**. r^2 aggregated across sessions for each subject, values are significantly positive for global latents and close to zero for local (within-area) latents. Gray bars show 95% of the null distribution, which is computed by taking the latents on session i and predicting pupil on session j, where $i \neq j$. Global latents are significantly more predictive of pupil than local latents for all subjects (Sa left $p < 10^{-6}$; Sa right $p < 10^{-6}$; Pe left $p = 0.000064$; Pe right $p = 0.000001$; Wa left $p = 0.000541$; Wa right $p = 0.000156$; paired sample t-test). To account for the fact that there can be a different number of latents for global, local left, and local right on any given session (since we use crossvalidation to select dimensionality on each session separately), we reran the same analysis, but only used a single latent to predict pupil. For global, we used the latent with highest correlation; for local, we used the latent that explained the most shared variance. We found that global latents had higher r^2 than local latents (Sa left $p < 10^{-6}$; Sa right $p < 10^{-6}$; Pe left $p = 0.00010$; Pe right $p = 0.000010$; Wa left $p = 0.000274$; Wa right $p = 0.000096$; paired sample t-test), consistent with our result in panel **c**. Thus, the result in **c** cannot be explained by the fact that global shared variability was higher dimensional than local shared variability (Fig. 16*a*).

latents (Fig. 17*b*). We quantified goodness of fit by measuring the coefficient of determination (r^2) for predictions. We found that the global across-hemisphere latent variables demonstrated significantly larger ability to predict pupil than did the local within-area latent variables (Fig. 17*c*; r^2 is significantly higher for global than either local left or local right). Interestingly, we found that this predicted pupil signal was related to but not synonymous with the pupillary evoked response on each trial (Supp. Fig. 14). Overall, these results are consistent with shared trial-by-trial encoding of a global cognitive process across areas and hemispheres of cortex.

4.8 Discussion

In this work, we utilized simultaneous dual hemisphere recordings to study interactions across hemispheres of cortex in prefrontal cortex. Using pairwise analyses, we found that correlations tended to appear to be larger within hemisphere pairs compared to across hemisphere pairs. However, using a new dimensionality reduction approach that we developed, called pCCA-FA, we identified across hemisphere components that were larger in magnitude (%sv) and dimensionality (d_{shared}) than variability shared among neurons within the same hemisphere. We found that across-hemisphere latent variables were predictive of pupil size, while within-hemisphere latent variables were not. Taken together, our results suggest that a large portion of shared neuronal variability in PFC can be explained by across-hemisphere interactions, which are predictive of signatures of global cognitive phenomena.

Neural variability shared across hemispheres of cortex may arise from a variety of mechanisms. In our paper, we highlighted the fact that across hemisphere shared variability predicted pupil diameter. This is consistent with global cognitive modulatory signals like arousal or fatigue contributing to the observed variability. Another source of across hemisphere shared variability could be shared information about the external world. For example, work in rodents has shown that small movements contribute to a large portion of variability in the visual cortex [113]. Further work will be needed to explore to what extent these movement related signals appear in non-motor regions (like PFC) of non-human primates. A third possible source of across hemisphere shared variability is direct communication between the hemispheres of PFC. Previous work has suggested ways that the two hemispheres may work together [63, 115, 123] using these connections. Further work is needed to understand the extent to which these and other sources contribute to across hemisphere shared variability.

Pupil diameter has been widely studied as an indirect measure of arousal signals in the brain. The majority of these studies have involved MRI or EEG signals that allow for a relatively coarse measurement of neural activity using a wide window [121]. Recent work using implanted electrodes have identified neural activity that predict pupil diameter in a wide range of brain areas, including rodent area V1 [113] and macaque area V4 [41]. One question that arises from this literature is whether the signals that predict pupil diameter in any given brain region are correlated with analogous signals in other brain areas. Here, we identified signals that predict pupil diameter and are shared across hemispheres. We found almost no within-hemisphere interactions that predicted pupil diameter beyond what was shared across hemispheres. Our work suggests that brain regions and neurotransmitters that modulate pupil and cortical activity likely do so in a non-specific manner, with many cortical brain regions likely receiving the modulatory signals.

Although the focus of our study was on variability shared across hemispheres, we also identified a substantial amount of variability that was shared among neurons of the same hemisphere but not neurons across hemispheres. The origin of this shared within hemisphere variability is unclear. There are a number of possible sources that likely contribute to this variability. One possibility is that there are feedforward or feedback input signals that modulate brain areas in the two hemispheres separately. For example, Rabinowitz et al. (2015) [46] used recordings of area V4 during a spatial attention task and identified two latent variables that accounted for attention-related modulation of neural activity. Their analysis found that each latent variable described the attention modulation in one hemisphere of V4, and were uncorrelated with one another. Similar signals (e.g., spatial attention) that could selectively modulate the activity of many PFC neurons in one hemisphere may account for some of the within hemisphere variability. Another possible source of shared within-hemisphere variability is constraints on patterns of neural activity imposed by the cortical circuitry in each hemisphere. Previous studies have shown that clustering structure in neural network models can lead to shared trial-to-trial fluctuations

within groups of neurons in a recorded population [102, 103, 131].

4.9 Methods

Surgical preparation

We implanted three adult rhesus macaque monkeys each with two 100-electrode "Utah" arrays (Blackrock Microsystems, Salt Lake City, UT). Electrode arrays were placed in the prefrontal cortex anterior to the arcuate sulcus and dorsal to the medial sulcus in both hemispheres. In a prior procedure, titanium headposts were fixed onto the skull of each subject using titanium screws. This was done to limit head movement during experiments. Surgeries in each subject were performed in sterile conditions under general anesthesia using isoflurane. All experimental procedures were approved by the Institutional Animal Care and Use Committee of the University of Pittsburgh.

Electrophysiological methods

Signals from the implanted electrodes were band-pass filtered (0.3 - 7500 Hz) and then digitized at 30,000 Hz before being stored offline for analysis. Waveforms were defined as a 52-sample (1.73 ms) window of the filtered voltage signal triggered by the signal crossing a predefined threshold. The threshold was defined as a multiple of the root-mean-square of a short snippet of the raw signal collected at the beginning of the session.

Behavioral Task

Subjects were trained to perform a standard memory-guided saccade task [132]. At the beginning of each trial, a 0.5 degree blue circle appeared at the center of a gray screen. The subject initiated fixation within an invisible 2.3 degree diameter window centered on the blue circle and then 200 ms later a white circle appeared in the subjects periphery 12 or 16 degrees from fixation at one of 4, 8, or 16 locations depending on the session. The white circle remained on the screen for either 100, 200, or 400 ms depending on the session after which the white circle was removed from the screen. The subjects then continued to fixate the blue fixation circle until it disappeared from the screen (after 1.5 to 3 seconds) indicating for the subject to saccade toward the location where the white target flash occurred. The subject had 400 ms to initiate fixation, defined by the eye position leaving a 0.9 degree window centered on the blue fixation circle. After saccade initiation, the subject had 200 ms to reach the target window, defined by a 2.1 degree radius window centered on the target location. The subject then needed to maintain fixation within the target window for 150 ms after which the subject was provided with a liquid reward for a saccade to the correct location. For a subset of sessions, a dim white target was flashed after saccade initiation to assist the subject in target acquisition. Trials were pseudo-randomized in mini-blocks during which the subject was required to correctly complete all target directions before beginning a new mini-block. While some of the above parameters varied slightly from session to session or subject to subject, all parameters remained constant within a session.

Preprocessing of neural data

To remove non-neural artifacts from among the saved waveforms, we used a neural network to classify waveforms as neural or not neural. Details of the method were described previously in [105]. Briefly, a neural network was trained on human sorted waveforms to distinguish between waveforms putatively of neural origin and waveforms not of neural origin.

We further removed channels that were likely to contain artifacts. To do this, we first binned neural activity by counting threshold crossing that occurred between target onset and fixation

offset. We then removed channels with mean spike count lower than 2 spikes/second and Fano factor greater than 10.

We also removed channels affected by artifactual cross-talk due to electrical coupling. For each pair of channels, we flagged spikes as coincident if they occurred within 100 us of each other. If either neuron in the pair had 20% of its spikes flagged as coincident, we flagged that pair as having artifactual crosstalk. We then removed the fewest number of channels as possible to eliminate crosstalk on the array.

After this process the number of remaining units in Monkey P was 79.3 ± 8.3 for right and 78.3 ± 7.8 for left hemisphere, in Monkey W was 24.9 ± 4.1 for right and 85.1 ± 8.0 for left hemisphere, and in Monkey S was 62.6 ± 9.8 for right and 75.3 ± 13.2 for left hemisphere. For all analyses in this paper, neural activity was preprocessed as described above and then binned using a 1 second window at the end of the delay period to compute spike counts.

Removing target information

For analyses of trial-to-trial variability (e.g., r_{sc} and population analyses using pCCA-FA), we removed target information and analyzed residual spike counts. For fitting the pCCA-FA model, we simply subtracted the condition mean from spike counts within each condition. When computing r_{sc}, we first z-scored spike counts (mean-subtracted and divided by standard deviation) within each condition.

Separation of slow and fast components

As we investigated interactions between left and right hemispheres, it soon became apparent that both hemispheres contained a component that varied slowly over the course of the session [41]. This was problematic for our analyses because slow processes like the ones we identified result in non-independent samples, which violates a key assumption in correlation analysis (e.g., regression, Pearson correlation, pCCA, or pCCA-FA; Supp. Fig. 10). It was therefore unclear whether these slow processes actually represented global across-hemisphere signals or whether their assignment to the global subspace was due to potentially spurious correlations induced by slow-timescale fluctuations [133]. Therefore, we removed slow components from all neural and pupillometry data, and focused most analyses in this work on faster-timescale trial-to-trial variability (Supp. Fig. 11; though see Supp. Fig. 13 for an analysis of slow components). We did this by computing the slow components using a centered boxcar filter of length 25 trials, computed after removing target information (as described above). We then subtracted this component from the raw spike counts or pupil size data to remove slow-timescale correlations that could have induced spurious correlations. We performed this pre-processing procedure independently for each neuron, and for pupil size data. All analyses in this study were performed on the residual faster-timescale component.

Measuring tuning

To study the delay period tuning of neurons to target location, we measured the average spike count of a neuron to each of the 4, 8, or 16 possible targets during the final one second window in the delay period. We then fit cosine tuning curves to these mean responses [134].

$$f_\theta = b + (f_{\max} - b)\cos(\theta - PD)$$

Where b is baseline, max is the maximum of the tuning curve, and PD is the preferred direction. We defined modulation depth to be the amount of modulation relative to the baseline: $(f_{\max} - b)/b$. To assess significance, we computed a null distribution of modulation depths using a

permutation test (shuffling the target angle labels), and labeled a neuron as significantly tuned if the actual modulation depth was larger than 95% of modulation depths in the null distribution (i.e. $p<0.05$).

Measuring signal correlation

We defined signal correlation between two neurons as the Pearson correlation between the two neuron's average responses to each condition during the delay period. To assess significance, we generated a null distribution of correlation values by using a permutation test (again shuffling the target angle labels), and labeled a pair as having significant positive signal correlation if the actual signal correlation was larger than 99% of the null distribution, and significant negative signal correlation if the actual signal correlation was smaller than 99% of the null distribution.

Probabilistic Canonical Correlation Analysis - Factor Analysis (pCCA-FA)

We develop a model called pCCA-FA to partition neuronal population structure into a global (across-area or across-hemisphere) component, a local (within-area) component, and a component independent to each neuron. The model is a novel combination of two existing dimensionality reduction and latent variable methods, namely probabilistic canonical correlation analysis, or pCCA [135] (which finds dimensions that maximize correlation between two brain areas) and factor analysis, or FA [136] (which maximizes covariance between neurons in a given brain area).

The pCCA-FA model (Fig. 14) explains spike counts in area x ($\mathbf{x}$) and area y ($\mathbf{y}$) according to global latent variables $\mathbf{z}$ and local latent variables $\mathbf{z_x}, \mathbf{z_y}$. To fully define the probabilistic graphical model (Fig. 14*b*), the priors over the latents and the conditional spike count observation distributions are:

$$\begin{aligned} \mathbf{z} \sim N(0, I_d) \quad \mathbf{z_x} \sim N(0, I_{d_x}) \quad \mathbf{z_y} \sim N(0, I_{d_y}) \\ \mathbf{x}|\mathbf{z}, \mathbf{z_x} \sim N(\mu_\mathbf{x} + W_x\mathbf{z} + L_x\mathbf{z_x}, \Phi_x) \\ \mathbf{y}|\mathbf{z}, \mathbf{z_y} \sim N(\mu_\mathbf{y} + W_y\mathbf{z} + L_y\mathbf{z_y}, \Phi_y) \end{aligned} \tag{18}$$

where $\mathbf{z} \in \mathbb{R}^{d\times 1}$ are the d latents shared across-areas, $\mathbf{z_x} \in \mathbb{R}^{d_x\times 1}$ are the d_x latents shared between neurons in area x, and $\mathbf{z_y} \in \mathbb{R}^{d_y\times 1}$ are the d_y latents shared between neurons in area y. If we assume that we record n_x neurons from area x, then $W_x \in \mathbb{R}^{n_x\times d}$ is the loading matrix for the global subspace in area x, and $L_x \in \mathbb{R}^{n_x\times d_x}$ is the loading matrix for the local subspace and area x. $\Phi_x \in \mathbb{R}^{n_x\times n_x}$ is a diagonal matrix containing the independent variances of each neuron, and $\mu_x \in \mathbb{R}^{n_x\times 1}$ is a vector of average responses of each neuron in area x. The parameters for area y are defined analogously.

Following the definitions in Eqn. 18, the marginal distributions for $\mathbf{x}$ and $\mathbf{y}$ are:

$$\begin{aligned} \mathbf{x} \sim N(\mu_\mathbf{x}, W_xW_x^T + L_xL_x^T + \Phi_x) \\ \mathbf{y} \sim N(\mu_\mathbf{y}, W_yW_y^T + L_yL_y^T + \Phi_y) \end{aligned} \tag{19}$$

By inspecting the marginal distributions, we observe that pCCA-FA decomposes the covariance of each area as the sum of a low-rank global component ($W_xW_x^T$), a low-rank within-area component ($L_xL_x^T$), and a diagonal independent neuron component (Φ_x).

Fitting pCCA-FA and computing latent variables

For fitting pCCA-FA to data, and computing latent variables, it is helpful to think of pCCA-FA as a generalized and structured factor analysis model. First, we define a joint vector of neural

activity in both areas and a joint vector of global and local latent variables:

$$\tilde{\mathbf{X}} := \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \qquad \tilde{\mathbf{Z}} := \begin{bmatrix} \mathbf{z} \\ \mathbf{z_x} \\ \mathbf{z_y} \end{bmatrix}$$

Now, the joint prior distribution of latent variables, and the joint conditional observation distribution of neural activity from Eqn. 18 can be written as:

$$\begin{aligned} \tilde{\mathbf{Z}} &\sim N(0, I_{d+d_x+d_y}) \\ \tilde{\mathbf{X}}|\tilde{\mathbf{Z}} &\sim N(\mu_{\tilde{\mathbf{X}}} + \tilde{L}\tilde{\mathbf{Z}}, \tilde{\Phi}) \end{aligned} \tag{20}$$

where d, d_x, d_y represent the global dimensionality, local dimensionality of area x, and local dimensionality of area y respectively. The model parameters in Eqn. 20 are:

$$\mu_{\tilde{\mathbf{X}}} = \begin{bmatrix} \mu_{\mathbf{x}} \\ \mu_{\mathbf{y}} \end{bmatrix} \qquad \tilde{L} = \begin{bmatrix} W_x & L_x & 0 \\ W_y & 0 & L_y \end{bmatrix} \qquad \tilde{\Phi} = \begin{bmatrix} \Phi_x & 0 \\ 0 & \Phi_y \end{bmatrix}$$

where all parameters are exactly the same as those in Eqn. 18. The model definitions in Eqn. 1 and Eqn. 2 are equivalent. The definition in Eqn. 2 makes it easy to see that we can think of pCCA-FA as a generalized and structured factor analysis model [50, 102, 104]. The structure in the loading matrix $\tilde{L}$ (i.e., the zeros) ensure that the local latent variables ($\mathbf{z_x}$ and $\mathbf{z_y}$) only contribute to variability in their respective areas, while global latent variables ($\mathbf{z}$) contribute to variability in both areas. Based on the definition in Eqn. 20, the marginal is:

$$\tilde{\mathbf{X}} \sim N(\mu_{\tilde{\mathbf{X}}}, \tilde{L}\tilde{L}^T + \tilde{\Phi}) \tag{21}$$

Based on the definitions in Eqns. 20 and 21, we can fit pCCA-FA model parameters to data using the EM algorithm [137]. This EM algorithm is the same as that for factor analysis, with the added step of maintaining the structure in $\tilde{L}$ after the M-step parameter updates. When fitting to data, we jointly chose the dimensionalities for global (d), local area x (d_x), and local area y (d_y) subspaces using 10-fold cross-validation.

To obtain the global and local latent variables in pCCA-FA (e.g., for use in predicting pupil size; Fig. 17), we compute the posterior mean of latent variables:

$$E\left[\tilde{\mathbf{Z}}|\tilde{\mathbf{X}}\right] = \tilde{L}^T\left(\tilde{L}\tilde{L}^T + \tilde{\Phi}\right)^{-1}\left(\tilde{\mathbf{X}} - \mu_{\tilde{\mathbf{X}}}\right)$$

where the first d entries of $E\left[\tilde{\mathbf{Z}}|\tilde{\mathbf{X}}\right]$ are $E\left[\mathbf{z}|\mathbf{x},\mathbf{y}\right]$ (global latents), the next d_x entries are $E\left[\mathbf{z_x}|\mathbf{x}\right]$ (local latents for area x), and the final d_y entries are $E\left[\mathbf{z_y}|\mathbf{y}\right]$ (local latents for area y).

Connection to canonical correlations, CCA, and pCCA

In words, the objective of Canonical Correlation Analysis (CCA) is to find a dimension in area x, and a dimension in area y, such that when neural activity is projected onto these dimensions, the Pearson correlation is maximized. Further dimensions can be found by also maximizing correlation, subject to the constraint that new dimensions are uncorrelated with previous dimension that are found. The correlations along these dimensions are known as canonical correlations (ρ).

CCA also has a probabilistic interpretation [pCCA; 135], defined by the graphical model:

$$\begin{aligned} \mathbf{z} &\sim N(0, I_d) \\ \mathbf{x}|\mathbf{z} &\sim N(\mu_{\mathbf{x}} + W_x\mathbf{z} + L_x\mathbf{z_x}, \Psi_x) \\ \mathbf{y}|\mathbf{z} &\sim N(\mu_{\mathbf{y}} + W_y\mathbf{z}, \Psi_y) \end{aligned} \tag{22}$$

where d is the number of CCA (global) dimensions, W_x, W_y are loading matrices as in Eqn. 18, but now Ψ_x, Ψ_y, the within-area or "noise" covariance matrices, are full rank. One advantage of pCCA-FA over pCCA is that it generalizes the within-area noise covariance matrix by modeling it as low rank plus diagonal ($L_x L_x^T + \Phi_x$). The means that pCCA-FA has fewer parameters than pCCA, and is thus more robust and performant in limited data regimes (i.e., able to recover ground truth with fewer samples; Fig. 15).

It is not straightforward from the graphical model for pCCA (Eqn. 22) to see how pCCA maximizes correlation. However, it can be shown that pCCA recovers the same dimensions and subspaces (W_x, W_y) as CCA [135]. Intuitively, this means that the canonical correlations should be the same as well. Indeed, it can be shown mathematically, that the canonical correlations are equal to the Pearson correlation between the latent posterior means, defined as:

$$E[\mathbf{z}|\mathbf{x}] = W_x^T(W_x W_x^T + \Psi_x)^{-1}(\mathbf{x} - \mu_\mathbf{x})$$
$$E[\mathbf{z}|\mathbf{y}] = W_y^T(W_y W_y^T + \Psi_y)^{-1}(\mathbf{y} - \mu_\mathbf{y})$$

Analogously, the canonical correlations in the pCCA-FA model can be computed as the Pearson correlation between latent posterior means:

$$E[\mathbf{z}|\mathbf{x}] = W_x^T(W_x W_x^T + L_x L_x^T + \Phi_x)^{-1}(\mathbf{x} - \mu_\mathbf{x})$$
$$E[\mathbf{z}|\mathbf{y}] = W_y^T(W_y W_y^T + L_y L_y^T + \Phi_y)^{-1}(\mathbf{y} - \mu_\mathbf{y})$$

Measuring percent shared variance (%sv) and dimensionality (d_{shared})

We defined two metrics to characterize the global (across-hemisphere) and local (within-hemisphere) subspaces: percent shared variance (%sv) and dimensionality (d_{shared}). We used %sv to measure the amount of shared variance attributed to either global or local subspaces. We used dimensionality, measure using d_{shared} [102], to measure the complexity of these interactions. We computed these metrics similar to how they are for FA [102–104], but modified for the pCCA-FA model developed in this study.

To assess the amount of each neurons variance that could be explained by global latents, we computed the global percent shared variance, defined as:

$$\text{Global \%sv for neuron } k = \frac{W_{xk} W_{xk}^T}{W_{xk} W_k^T + L_{xk} L_{xk}^T + \Psi_k} \tag{23}$$

where W_{xk} is the k^{th} row of the global loading matrix for area x, L_{xk} is the k^{th} row of the local loading matrix for area x, and Ψ_{xk} is the independent variance for the k^{th} neuron in area x.

We similarly defined the local percent shared variance as:

$$\text{Local \%sv for neuron } k = \frac{L_{xk} L_{xk}^T}{W_{xk} W_k^T + L_{xk} L_{xk}^T + \Psi_k} \tag{24}$$

We defined global d_{shared} as the minimum number of modes needed to explain 95% of the global shared covariance matrix $W_x W_x^T$. To do this, we first identified the eigenvalues of $W_x W_x^T$ and sorted them from largest to smallest. Note that these eigenvalues indicate the amount of variance in $W_x W_x^T$ explained by the corresponding eigenvector. We then defined d_{shared} as the minimum number of eigenvalues needed such that the sum of the eigenvalues explains at least 95% of the sum of all of the eigenvalues. We defined local d_{shared} for each hemisphere using the procedure described above except that $W_x W_x^T$ was replaced with $L_x L_x^T$.

Here, we have described computation of %sv and d_{shared} for area x; we computed the metrics analogously for area y.

Pupil prediction

Pupil prediction for global and local latents was performed using linear regression and assessed by computing the proportion of variance explained by the predictions. First, mean pupil diameter for each trial was computed in the same time bin as was used for computing spike counts (i.e., a 1-second bin at the end of the delay period). We normalized pupil size for each session by using the mean and standard deviation of pupil sizes across the session. Additionally, we removed slow-timescale fluctuations in pupil size using the same method used for neural activity and focused analyses on faster-timescale trial-to-trial variability (Supp. Fig. 11).

We then computed the posterior means: for the global latents $E[z|x, y]$ and for the local latents $E[z_x|x]$ and $E[z_y|y]$. Here, z represents global latents, z_x and z_y represent left and right hemisphere local latents respectively, and x and y represent left and right hemisphere spike counts respectively.

We then fit a linear regression model between global latents and pupil size (Fig. 17*b* global), and reported the proportion of variance in pupil size explained by the model (i.e., r^2; Fig. 17*c* global). We repeated the same procedure for local left and right hemisphere latents, and reported local left and right r^2 (Fig. 17*b*-*c* local). To compute null distributions for pupil prediction, we used the latents on a given session i and repeated the procedure above, except using the pupil size on another session j (where $i \neq j$; trials were truncated in the session with more trials to ensure equal trial numbers). This resulted in null distributions with 240 samples for subjects Sa and Pe with 16 sessions, and a null distribution with 90 samples for subject Wa with 10 session. We report the 95% confidence intervals of this null distribution (Fig. 17*c* gray bars). We also test whether global or local latents are more predictive of pupil size using a paired sample t-test (Fig. 17*c*).

5 [Data augmentation] How to augment your ViTs? Consistency loss and StyleAug

Chapters 2, 3, and 4 focus on understanding neuronal variability in a natural neural system, i.e., the primate brain. However, variability also plays a crucial role in modern deep learning and artificial neural systems. For example, stochasticity plays an important role in regularization during training (e.g., dropout regularization, stochastic depth) and "internal" model variability plays an important role in generative modeling (e.g., variational autoencoders, generative adversarial networks). "External" model variability, in the form of the amount and diversity of data that is used to train a model, is also an important factor in the success of modern deep learning. As compute power increases and models become larger, there is an increasing need for larger and more diverse datasets. One way to improve the size, quality, and variability of training data is to use data augmentation–a term that encompasses a variety of techniques to generate new training samples from a given training set or distribution. While data augmentation is widely used, not much is known about how data augmentation strategies interact with the architecture of the deep learning model that is being trained. In this chapter, I explore the interaction between commonly-used and state-of-the-art data augmentations and model architectures for the task of image classification. I also introduce a new data augmentation loosely inspired by human visual perception, called StyleAug, that improves performance of the vision transformer (ViT), an architecture that has recently been shown to work very well for computer vision applications.

5.1 Introduction

For nearly a decade, convolution neural networks (CNNs) have been the de-facto deep learning architecture for a variety of computer vision tasks from image classification to object detection to segmentation [138–141]. A major reason for their success is due to the inductive biases imposed by the convolution operation, namely sparse interactions, weight sharing, and translational equivariance [142]. These inductive biases allow for efficient training of feature representations that are useful for vision tasks. Despite their widespread adoption, CNNs have room for improvement–they can be prone to adversarial attacks and perform poorly when there are distribution shifts (e.g., when images have been corrupted [143]). Other work has shown that CNNs rely on textures to categorize objects, while humans rely on object shape [144]. This can be problematic for using CNNs as a model of the human visual system [145].

Taking inspiration from the success of the Transformer architecture in language modeling [146], Vision Transformers (ViTs) are an alternative architecture that utilize the key mechanism of multi-head self-attention (as opposed to the key mechanism of convolution in CNNs). ViTs have recently shown promise for image classification, even outperforming state-of-the-art CNNs [147, 148]. Follow-up work has shown that ViTs also have other advantages relative to CNNs, including: 1) increased adversarial robustness [149], 2) increased robustness to corruptions [150], 3) ability to provide pixel-level segmentation using attention maps [150, 151], and 4) smaller texture bias and greater shape bias [150, 152], making them a good candidate model for human vision.

Although ViTs have attained competitive performance on vision tasks, they are known to be more difficult to train than CNNs. In ViTs, only multi-layer perceptron (MLP) layers operate locally and are translationally equivariant, while the self-attention layers [146] operate globally [147]. As such, ViTs are thought to have a weaker inductive bias than CNNs, thus requiring more data, augmentations, and/or regularization than training a similarly-sized CNN [148, 153, 154]. However, the strategies for data augmentation for ViT training have largely been adapted from the techniques used for CNNs. While these augmentations have worked reasonably well, certain training and augmentation strategies may be more beneficial for ViTs than for CNNs.

In this work, we performed a systematic empirical evaluation of data augmentation strategies on CNNs and ViTs. Importantly, we found that using a consistency loss penalty term between different augmentations of the same image [143] was especially helpful when training ViTs. We then introduced a novel data augmentation, called StyleAug, inspired by shape bias in human visual perception [144]. StyleAug performs neural style transfer from a given image to another randomly chosen image in the dataset during training. When combined with a consistency loss, StyleAug improves validation accuracy, robustness to corruptions, shape bias, and transferability. For training ViTs, StyleAug outperforms previous state-of-the-art augmentations such as RandAugment [155] and AugMix [143] across several metrics.

5.2 Related work

ViT training. ViTs have a weaker inductive bias than CNNs. To achieve classification performance better than CNNs, Dosovitskiy *et al.* [147], trained ViTs on very large datasets, either ImageNet-21k or the proprietary JFT-300M. To train ViTs with limited data and compute resources, Steiner *et al.* [154] explore data, augmentations, and regularization. They suggest that, for a fixed dataset size, one should generally prefer data augmentations over extensive regularization. In another study, Touvron *et al.* [148] trained data-efficient vision transformers using a combination of various augmentations, regularization strategies, and a novel distillation strategy. For distillation, they create a special "distillation token" in the transformer architecture that uses a CNN as the teacher network. Their data-efficient image transformer (DeiT) achieves competitive performance without large datasets (i.e., with only ImageNet-1k).

Data augmentation. Proper data augmentation can increase the size and quality of datasets, which can help prevent overfitting and greatly improve generalization of deep learning models. Since ViTs have a weaker inductive bias, they can be prone to overfitting [153], and thus benefit greatly from many strong augmentations [148].

From another perspective, data augmentation can also help deep learning models learn invariances such as scale (i.e., with cropping) and color. Indeed, the increasingly popular self-supervised learning methods learn feature representations by becoming invariant to image transformations. The goal of self-supervised learning is to map different augmentations of the same image to similar locations in the feature embedding space [151, 156, 157]. Caron *et al.* [151] show that multi-scale cropping is an especially useful augmentation for training self-supervised ViTs. Hendrycks *et al.* [143] take inspiration from the self-supervised learning literature, and use a Jensen-Shannon consistency loss (between a training image, and two augmentations of the image) in addition to a classification loss when training CNNs.

Shape vs. texture bias Geirhos *et al.* [144] used psychophysics experiments to show that humans make image classification decisions based on object shape, rather than relying on image texture. Presented with the same images, CNNs made decisions based on image texture. Geirhos *et al.* [144] created a new dataset, called Stylized ImageNet, in which they performed style transfer with ImageNet images as content and images of art as style. Trained on this data, CNNs showed improved shape bias and lower texture bias, but at the expense of classification performance. Xu *et al.* [158] used a random convolution augmentation (to distort textures) combined with a consistency loss to improve CNN generalization to unseen domains such as ImageNet-sketch.

5.3 Augmentation strategies

A Image transformations

For training models, we tested several basic and several state-of-the-art augmentations for image classification. All images first went through Inception-style preprocessing: 1) a resized crop with

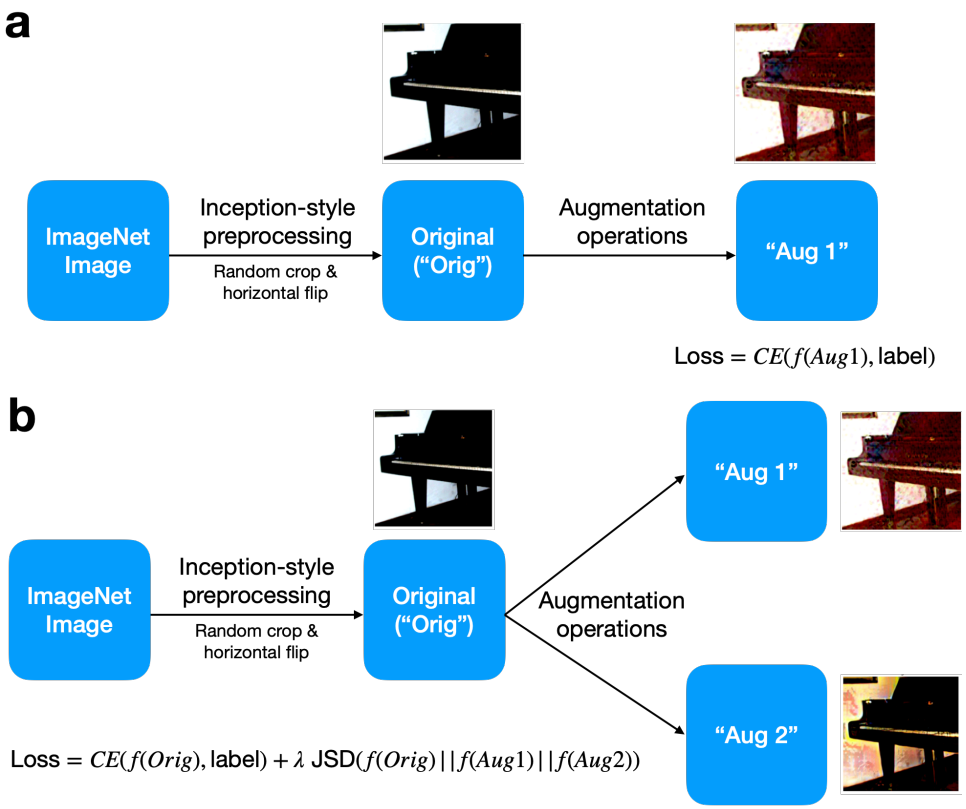


Figure 18: Augmentation setup. (a) Classic augmentation setup. Cross-entropy loss between the network prediction of the augmented image, $f(Aug1)$, and the true label. (b) Setup with a Jensen-Shannon (JSD) consistency loss. Cross-entropy loss between prediction of the original image $f(Orig)$, and an addition of a JSD loss between the three network predictions of each of the original image ($f(Orig)$) and two augmentations ($f(Aug1)$, $f(Aug2)$).

a randomly chosen scale in $[0.5, 1]$ and resized to 224×224, and 2) a random horizontal flip with $p = 0.5$ (Fig. 18*a*, ImageNet Image to "Orig"). We used a relatively large cropping scale in this step to allow for testing of multi-scale cropping augmentations later (see JSD loss below; [151]).

We then performed additional augmentation operation to this image (Fig. 18*a*, "Orig" to "Aug 1"). First, we tested basic augmentations such as random cropping, color jittering, and translation. Second, we tested RandAugment [155], a state-of-the-art augmentation for training CNNs on ImageNet, and AugMix [143], another state-of-the-art augmentation that improves CNN robustness to image corruptions. As in the Augmix paper, for both RandAugment and AugMix, we exclude transformations that overlap with ImageNet-C corruptions to allow for fair evaluations of model generalization and robustness. Third, we tested our new human perception-inspired augmentation StyleAug (described in detail below), and StyleAug with random cropping. Finally, we also tested another augmentation, called Neurofovea (Deza *et al.*, 2021 [159]), inspired by the human perceptual phenomena of foveation and metamerism [160].

In experiments, we considered the random cropping augmentation as a baseline as it is effectively the same as only training models with Inception-style preprocessing. Examples of all augmentations tested, along with further details such as any torchvision transforms used, are provided in the Appendix and Supp. Fig. 15.

B Jensen-Shannon divergence (JSD) consistency loss

For the typical training augmentation setup (Fig. 18*a*), we trained models using a cross-entropy classification loss (with label smoothing=0.1) between the model predictions (i.e., $f(\hat{y}|x_{aug1})$ posterior distribution over class labels given an image "Aug 1") and the true class label y. We tested the impact of using a consistency loss to train different model architectures. Following the AugMix paper [143], we used a Jensen-Shannon divergence (JSD) consistency loss between an image ("Orig") and two augmentations of the image ("Aug 1" and "Aug 2"). This JSD loss was applied in addition to a classification loss (Fig. 18*b*):

$$\mathcal{L}(f(\hat{y}|x_{orig}), y) + \lambda \text{JSD}(f(\hat{y}|x_{orig}) \mid\mid f(\hat{y}|x_{aug1}) \mid\mid f(\hat{y}|x_{aug2})) \tag{25}$$

We used $\lambda = 12$, the value used in AugMix [143]. The JSD loss is computed as follows:

$$\text{JSD}(p_{orig} \mid\mid p_{aug1} \mid\mid p_{aug2}) = \frac{1}{3}\left(KL(p_{orig} \mid\mid M) + KL(p_{aug1} \mid\mid M) + KL(p_{aug2} \mid\mid M)\right) \tag{26}$$

where KL is the KL divergence, and $M = (p_{orig} + p_{aug1} + p_{aug2})/3$. The JSD loss imposes a large penalty when the posterior distribution predictions for the three versions of the training image ("Orig", "Aug 1", and "Aug 2") are very different. Thus, the JSD consistency loss requires models to learn similar feature representations and output distributions across the different augmented versions of the same image. This explicitly trains models to become invariant to the augmentations used.

5.4 StyleAug

Figure 19: StyleAug: neural style transfer from a given image in the batch to another randomly chosen image in the dataset. "Orig" (left) shows the original image after Inception-style preprocessing; "Aug 1" (middle) and "Aug 2" (right) show two StyleAug augmentations of "Orig".

Geirhos *et al.* [144] showed that CNNs trained on ImageNet make classification decision mainly based on image textures (i.e., they have high texture bias). However, high shape bias and low texture bias is desirable because models with this property tend to show better generalization and increased robustness [144, 150, 158]. New datasets (i.e., Stylized ImageNet [144]) and augmentations (i.e., random convolutions [158]) have been developed to try to improve CNN shape bias, robustness, and generalization. However, training with these techniques are expensive and/or do not improve validation accuracy on the original ImageNet dataset. We sought to develop an augmentation that: 1) is fast and can be used in real-time during training, 2) improves shape bias, and 3) improves performance on ImageNet.

Algorithm 1: StyleAug training with Jensen-Shannon (JSD) consistency loss

Input : Model f, classification loss $\mathcal{L}$, training image x and its class label y, two images sampled randomly from the current mini-batch x_{rand1}, x_{rand2}

Function `StyleAug`(x,x_{style},$\alpha = 50$,$\beta = 50$)**:**

- $z = VGG_{enc}(x)$ `// VGG encoder from [161]`
- $z_{style} = VGG_{enc}(x_{style})$
- $z_{adain} = AdaIn(z, z_{style})$ `// adaptive instance normalization`
- $x_{adain} = VGG_{dec}(z_{adain})$ `// VGG decoder from [161]`
- $m \sim Beta(\alpha, \beta)$
- $x_{aug} = m \cdot x + (1 - m) \cdot x_{adain}$ `// mix with original representation`
- **return** x_{aug}

$x_{orig} = \text{InceptionStylePreprocess}(x)$ `// Random crop and horizontal flip`

$x_{style1} = \text{InceptionStylePreprocess}(x_{rand1})$

$x_{style2} = \text{InceptionStylePreprocess}(x_{rand2})$

$x_{aug1} = \texttt{StyleAug}(x_{orig}, x_{style1})$

$x_{aug2} = \texttt{StyleAug}(x_{orig}, x_{style2})$ `//` $x_{aug1} \neq x_{aug2}$

Loss Output: $\mathcal{L}(f(\hat{y}|x_{orig}), y) + \lambda \text{JSD}(f(\hat{y}|x_{orig}) \mid\mid f(\hat{y}|x_{aug1}) \mid\mid f(\hat{y}|x_{aug2}))$

We introduce a new data augmentation called StyleAug (Fig. 19). StyleAug performs style transfer between two images, using one as the content image and another as the style image. The resulting style-transferred images have the shapes present in the content image and the colors and textures present in the style image. StyleAug uses a style transfer method that performs real-time arbitrary style transfer (AdaIn; [161]), which computed fast enough to use for data augmentation. For the style transfer, we use a training image (whose label is preserved) as the content image, and another randomly chosen image in the batch as the style image. To ensure that the training label is preserved, the augmented image is a mix of the original image and the style-transferred image, where the mixing weight m was drawn from a $\beta(50, 50)$ distribution (i.e., most of the time m was close to 0.5, but there was some stochasticity in the amount of style distortion). In terms of computational resources, training models with StyleAug used approximately the same amount of time and resources as training models with RandAugment or AugMix.

StyleAug tended to preserve the shape content of an image but distorted its colors and textures (e.g., Fig. 19). By combining StyleAug with the JSD consistency loss, we explicitly trained networks to become invariant to the color, texture, and other distortions/transformations that were induced by StyleAug. Pseudocode for StyleAug training with the JSD loss is provided in Algorithm 1.

For both CNNs and ViTs, StyleAug greatly improved shape bias over other augmentations we tested (Fig. 22). Moreover, for ViTs StyleAug also provided the best ImageNet validation performance (Fig. 20), mean corruption accuracy on ImageNet-C (Fig. 21, and transfer learning performance to the Pet37 dataset (Fig. 23a).

5.5 Experiments

For fair comparison, we trained models of similar size: ResNet-50 for CNNs ($\sim$ 25 million parameters), and ViT-Small with 16x16 patch size for vision transformer ($\sim$ 22 million parameters)

[153]. All models were trained from random initialization for 100 epochs on ImageNet-1k. We used the AdamW optimizer with a peak learning rate of 0.001 with linear warmup for 10 epochs followed by a cosine learning rate decay schedule. For ResNet-50 training, we used a weight decay of 0.05, a batch size of 512 for typical training (no JSD) and 200 for training with the JSD consistency loss (since the JSD loss uses $\sim 3\times$ the number of images per batch). ViT-Small/16 required more GPU memory during training, requiring smaller batch sizes of 400 (no JSD) and 150 (JSD). We also used a larger weight decay of 0.3 for ViT-Small/16. We trained models using 2 GPUs (Nvidia Tesla V100) in parallel for training without the JSD loss, and 4 GPUs for training with the JSD loss. Basic augmentations (crop, color, translate) required approximately 30-45 minutes per epoch for training, while the other augmentations required approximately 60-75 minutes.

We trained models using the augmentations described in Sections 3 and 4, with and without a JSD consistency loss, and evaluated their performance on:

1. ImageNet validation accuracy
2. Robustness to corruptions / distribution shift (i.e., accuracy on ImageNet-C [162])
3. Shape bias vs. texture bias on cue-conflict images [144]
4. Transfer learning to: The Oxford-IIIT Pet dataset (pet37, which is a dataset of natural images [163]) and to resisc45 (a dataset of satellite images [164]).

A ImageNet-1k validation accuracy

To evaluate models' performance on the ImageNet-1k, we preprocessed validation images by resizing to 256 pixels, and then taking a 224×224 center crop. Here, we report the accuracy of models on the ImageNet-1k validation set. We note that ViT-S validation accuracy was lower than ResNet-50 performance. This is expected when training on the relatively small ImageNet-1k; ViT models that outperformed ResNets were trained on larger datasets (ImageNet-21k or JFT-300M [147]) or used knowledge distillation [148]. Here, we were more interested in comparing

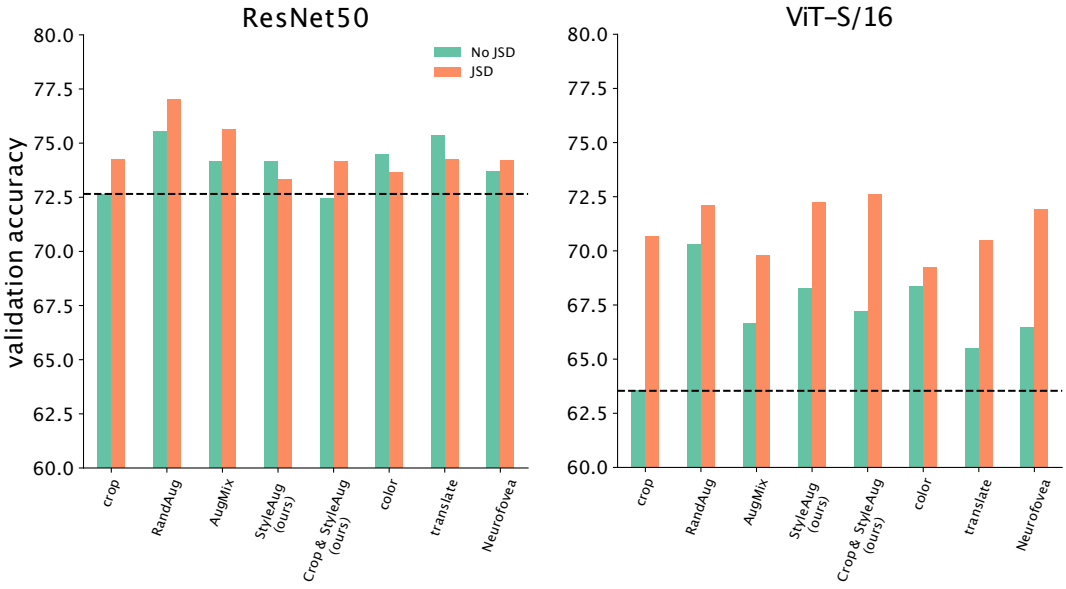


Figure 20: Validation accuracy of different augmentations on ImageNet-1k.

the relative performance improvements of using one augmentation strategy over another, and how that differed between CNNs and ViTs.

Most importantly, we found that using a JSD consistency loss provided a large boost in accuracy across all augmentations when training ViT-S (Fig. 20 right panel, orange bars all above green bars). For ResNet-50, using a JSD loss improved accuracy for some augmentations but resulted in lower accuracy for others (Fig. 20 left panel).

Second, we found that the augmentations that worked best for ResNet-50 were different from those that worked best for ViT-S. For ViT-S, our proposed augmentation, StyleAug and StyleAug + crop, have the best accuracy, followed closely by RandAugment and Neurofovea. For ResNet-50, the state-of-the-art RandAugment and Augmix do best, while accuracy for StyleAug is lower.

Finally, we note that our cropping + JSD loss augmentation (see Supp. Fig. 15a) is very similar to multi-scale cropping used in DINO [151]. Thus, our finding that using a JSD consistency loss with random cropping supports the finding in the DINO paper that multi-scale cropping in a self-supervised setting is a very beneficial augmentations for ViTs.

B Robustness to corruptions

We tested the models trained on ImageNet-1k on their robustness to distribution shift (i.e., image corruptions). To do so, we evaluated each models' performance on ImageNet-C [162], which contains 19 different corruptions across 5 different severity levels each. We report the mean corruption accuracy as the model's average accuracy across the 95 datasets present in ImageNet-C.

For ViT models, we found that training with StyleAug and a JSD consistency loss attained the highest corruption accuracy, and in fact also outperformed all ResNet-50 models (Fig. 21). RandAugment and Augmix with a JSD loss had the highest corruption accuracy among ResNet-50 models. Secondly, we again found that using the JSD consistency loss during training boosted the corruption accuracy of ViT models by close to 5% in many cases (Fig. 21 right panel, orange bars above green bars). However, using a JSD loss again provided mixed results in corruption accuracy for ResNet-50 models (Fig. 21 left panel).

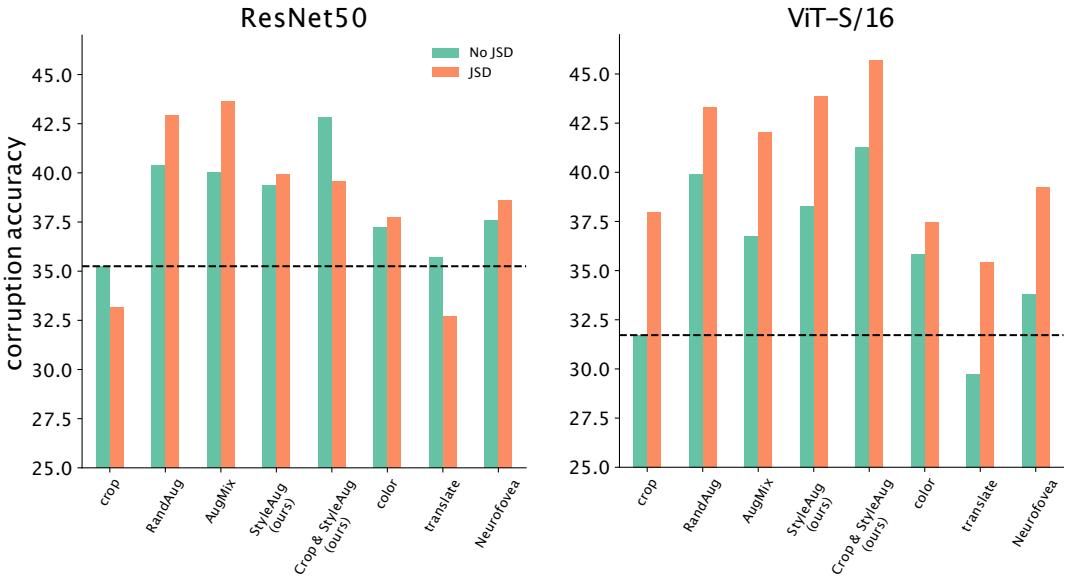


Figure 21: Mean corruption accuracy of different augmentation strategies on ImageNet-C.

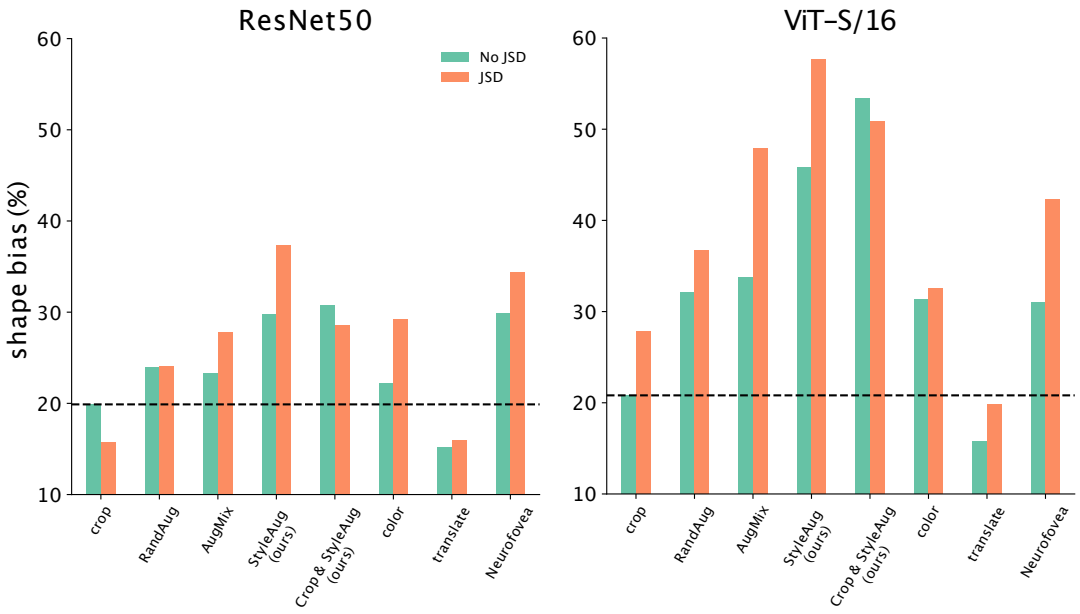


Figure 22: Shape bias of different augmentation strategies.

We also note that while ImageNet-1k validation accuracy was lower for ViT-S than ResNet-50, here we found that, for the same models as in Fig. 20, many ViT-S models had higher corruption accuracy than ResNet-50 models. This supports findings in Naseer *et al.* [150] that suggest that vision transformers tend to be more robust than CNNs.

C Shape bias

We tested each models' shape bias relative to its texture bias. To do so, we evaluate the models trained on ImageNet-1k on the cue-conflict images from Geirhos *et al.* [144]. Cue-conflict images were generated by performing iterative style transfer between two images. Thus, they have an object shape label (based on the content image) and an texture label (base on the style image; see Supp. Fig. 16 for an example). The shape bias is defined as the number of correctly classified shape labels relative to the total number of correctly classified images (either shape or texture) [144]:

$$\text{shape bias} = \frac{\#\text{ correct shape labels}}{\#\text{ correct shape labels} + \#\text{ correct texture labels}}$$

We found that training ViT-S with the JSD consistency loss greatly improved shape bias (Fig. 22 right panel, orange bars above green bars), and that StyleAug provided the highest shape bias. Training with the JSD loss also tended to increase shape bias across augmentations for ResNet-50, and we also found that our proposed StyleAug provided the highest shape bias for ResNet-50 (Fig. 22 left panel). However, we note that ResNet-50 models had much lower shape bias than ViT-S models.

D Transfer learning

We tested the transferability of models trained in on ImageNet-1k. To do so, we froze the backbone weights, replaced the classification heads, and only finetuned the weights of the new classification heads on Pet37 or Resisc45. For Pet37, we used SGD with momentum, with a batch size of 512, and learning rate of 0.01 for 10 epochs, followed by a learning rate of 0.003 for

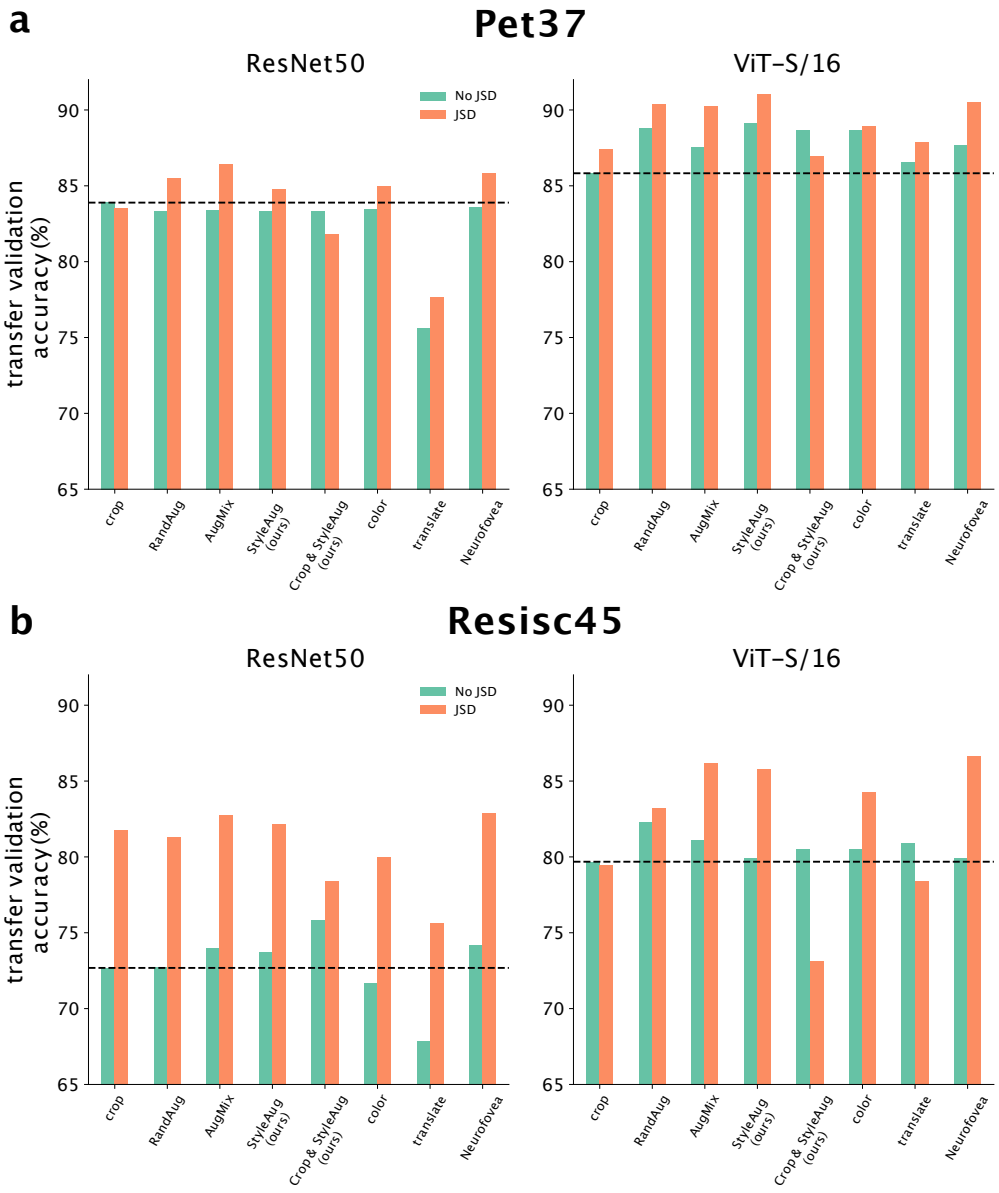


Figure 23: Transfer learning of ImageNet trained models. (a) Validation accuracy on Pet37 after transfer learning. (b) Validation accuracy on Resisc45 after transfer learning.

10 epochs. We evaluated performance on the test split used in [163]. For Resisc45, we used SGD with momentum, with a batch size of 512, and learning rate of 0.01 for 10 epochs, 0.003 for 5 epochs, and 0.001 for 5 epochs. Since there is not a standard training / test split, we performed a single random 80/20 split which was kept constant across training and evaluation of different models. We did not use any augmentation during transfer learning; we only used augmentations while training on ImageNet-1k.

For both Pet37 and Resisc45, we found that JSD consistency loss improved transfer learning for both ResNet-50 and ViT-S models (Fig. 23*a*-*b*, orange bars typically above green bars). For both datasets, we found that ViT-S models transferred better than ResNet-50 models (Fig. 23*a*-*b*, bars in right panels higher than those in left panels). For Pet37, StyleAug worked best again for ViT-S while AugMix worked best for ResNet-50 (Fig. 23*a*). For the satellite images of Resisic45, Neurofovea, Augmix, and StyleAug worked well for both ViT-S and ResNet-50 (Fig.

23*b*).

5.6 Conclusion

In this work, we systematically evaluated how different commonly-used augmentation strategies perform on different model architectures. We showed that the data augmentations that work best for ViTs are different than those that work best for CNNs. Importantly, we found that using a Jensen-Shannon consistency loss in addition to a classification loss when training ViTs provided considerable performance improvement in almost all cases. Importantly, although ViT performance lagged CNN performance on ImageNet-1k, they were generally more robust to corruptions, had higher shape bias, and were more transferable. We hope that future work will scrutinize other existing data augmentations and training strategies that have worked well for CNNs, and consider whether they should be used for other model architectures like ViTs.

We also introduced StyleAug: real-time neural style transfer from a training image to another randomly chosen image in the dataset. For ViTs, StyleAug outperforms other state-of-the-art augmentations in accuracy, robustness, transfer learning, and shape bias. We hope that future research will continue to develop augmentations and training strategies that work well for vision transformers, even if they might not benefit the previously dominant CNN architecture.

6 Conclusion

This dissertation presents work that furthers our understanding of variability in both artificial and natural neural systems (i.e., the brain). For artificial neural networks and deep learning, this work highlights that one should consider different techniques to increase training data variability (i.e., data augmentation) for different model architectures (Chapter 5). For systems neuroscience, this work advances the understanding of the structure of shared neural variability (Chapter 2), its distinct sources (Chapter 4), and to what degree it can be controlled (Chapter 3).

6.1 Summary of contributions

Structure of shared neural variability [Chapter 2]

Pairwise correlations (Pearson correlation between spike counts) [7] and population metrics computed from dimensionality reduction methods [50] both aim to characterize shared trial-to-trial neuronal variability. Although they are both computed from the same spike count covariance matrix, the relationship between the two is not known. We established the relationship between pairwise and population metrics both analytically (i.e., through mathematical proofs) and empirically using simulations. Our results demonstrated that changes in the mean pairwise correlation could correspond to one (or several) of a number of changes in population metrics: 1) the strength of shared variability (%sv), 2) the patterns of shared variability (loading similarity), and 3) dimensionality (d_{shared}). We showed that the standard deviation of pairwise correlations, which is rarely reported, provides complementary information to the mean pairwise correlation about population covariance structure. In recordings of macaque area V4, we found that the previously-reported decrease in mean pairwise correlation with attention corresponds to multiple distinct changes in population metrics. Overall, our framework builds the intuition to navigate between pairwise correlations and dimensionality reduction, allowing for a more interpretable and richer description of the structure of shared neuronal variability.

Control of shared neural variability [Chapter 3]

Neural activity drifts slowly over time, and the direction of these drifts are often shared among the neurons in a population. These slow drifts in population activity have been linked to slow changes in cognitive phenomena such as arousal, impulsivity, and engagement [41, 42, 113]. In this study, we asked to what degree animals could volitionally modulate these slow co-fluctuations and stabilize neuronal activity over the course of several hours. We trained two rhesus macaques to control a novel brain computer interface (BCI) paradigm that provided visual feedback about their prefrontal cortex population activity. The size of an on-screen annulus was linked to the distance of the animal's neuronal activity from a "target" neuronal state that was defined at the beginning of the session. By using the BCI, animals: 1) were successfully able to reduce the distance of their internal neuronal state to the target state, and 2) control shared variability to suppress slow neuronal drifts. Future work will investigate whether this suppression in slow neuronal drifts also corresponded to suppression of slow changes in pupil size and internal cognitive states (e.g., arousal, impulsivity, or engagement).

Sources of shared neural variability [Chapter 4]

Shared trial-to-trial variability in one area of cortex may be shared with another area (e.g., an input or output area), or be due to brain-wide signals that impact many areas (e.g., arousal, impulsivity) [41, 42, 113]. In this work, we utilized simultaneous bilateral prefrontal cortex (PFC) recordings to study shared variability across hemispheres of cortex vs within a single

brain area. We developed a new probabilistic graphical model, called pCCA-FA, to identify and separate global (across-hemisphere) and local (within-area) sources of shared variability. In our PFC data, we identified across-hemisphere components that were larger in magnitude (%sv) and dimensionality (d_{shared}) than variability shared among neurons within the same hemisphere. We found that across-hemisphere latent variables were predictive of pupil size, while within-hemisphere latent variables were not. Taken together, our results suggest that a large portion of shared neuronal variability in PFC can be explained by across-hemisphere interactions, and these across-hemisphere interactions are predictive of signatures of global cognitive phenomena.

Different augmentations for different neural network architectures [Chapter 5]

Variability is an important aspect of modern artificial neural networks and deep learning. Internal model variability is used in generative models [43, 165] and in regularization during model training [44, 45]. External variability in terms of training dataset size and diversity is also necessary to train models that are robust and generalizeable. In order to improve training data variability and quality, data augmentation is used to generate new training samples from a given dataset or distribution. In this work, we systematically evaluated how different augmentation strategies perform on different model architectures for image classification. The data augmentations that worked best for vision transformers (ViTs; [147]) were different than those that worked best for convolutional neural networks (CNNs). We found that using a Jensen-Shannon consistency loss in addition to a classification loss when training ViTs provided considerable performance improvement in almost all cases. We also introduced a new data augmentation, called StyleAug: real-time neural style transfer from a training image to another randomly chosen image in the mini-batch. For ViTs, StyleAug outperformed other state-of-the-art augmentations in accuracy, robustness, transfer learning, and shape bias. We hope that future research will continue to develop augmentations and training strategies that work well for vision transformers and other neural network architectures, even if they might not benefit the previously dominant CNN architecture.

6.2 Discussion and future directions

Shared neuronal variability and information coding

An improved understanding of the characteristics of shared trial-to-trial neuronal variability is critical to elucidating how the brain encodes and processes information. Previous literature has noted that correlated variability can impact the amount of information encoded in a neuronal population. Some of these studies measured trial-to-trial variability using spike count correlations (r_{sc}; [34, 35]), while others used a high-dimensional approach [37, 38]. In chapter 2, we provided a framework that related r_{sc} and dimensionality reduction, allowing one to bridge between the literatures of r_{sc} and dimensionality reduction (and population metrics). By considering the three population metrics–percent shared variance, loading similarity, and dimensionality–used in Chapter 2, along with the way in which mean population responses vary across conditions, we can more incisively characterize how trial-to-trial variability impacts information coding than by using r_{sc} mean alone. We can use the three population metrics to measure how patterns of shared variability are related to (e.g., align with or are orthogonal to) patterns of stimulus encoding and downstream readouts [14, 38, 41].

Some cognitive phenomena such as attention and learning have been shown to change the properties of shared neuronal variability in the brain [13, 14, 46, 104, 166]. These changes could impact the amount of information that can be encoded. For example, paying attention to a location in space may enable a neuronal population to improve encoding fidelity for stimuli in that location. However, it is not know whether the nature of these changes are global to the

entire brain, or local to the relevant neuronal population performing the encoding. We hope that future work will utilize simultaneous multi-area recordings [39] and statistical techniques like pCCA-FA in Chapter 4 to separate different sources of shared variability and investigate their potentially distinct impacts on information coding. Moreover, while information-limiting correlations have begun to be measured empirically [39, 85], little has been shown empirically about their behavioral impact. We hope that future work will utilize difficult behavioral tasks (such as fine discrimination) or brain computer interfaces (Chapter 3; [167]) to directly investigate how information-limiting correlations impact behavioral performance.

Data augmentation for neuroscience

Data augmentation strategies (e.g., those in Chapter 5) are commonly used in deep learning in a variety of domains and tasks, including computer vision [168], natural language processing [169, 170], and self-supervised learning [151, 156, 157] among others. However, data augmentation has not often been used for neuroscience applications. This is presumably because of the much smaller signal to noise ratio in neural data (e.g., due to the variable nature of neural responses discussed in Chapter 2-4). However, careful development and application of data augmentation techniques to neural data is potentially a fertile ground for future research. For brain computer interfaces, augmentation might be used to train predictive models with less data and greater robustness to instabilities or phenomena such as slow drift discussed in Chapter 3 [41, 42, 171, 172]. For inference, augmentation can be used in addition to or as an alternative to regularization to train models such as factor analysis (Chapter 2; [99, 102, 104]) or pCCA-FA (Chapter 4), enabling model fitting with limited data while also preventing overfitting to noise. Because of the constraints and expenses involved in electrophysiology experiments, developing data augmentation strategies for neural data will prove invaluable for future systems neuroscience research.

Intersection of deep learning and neuroscience/perception

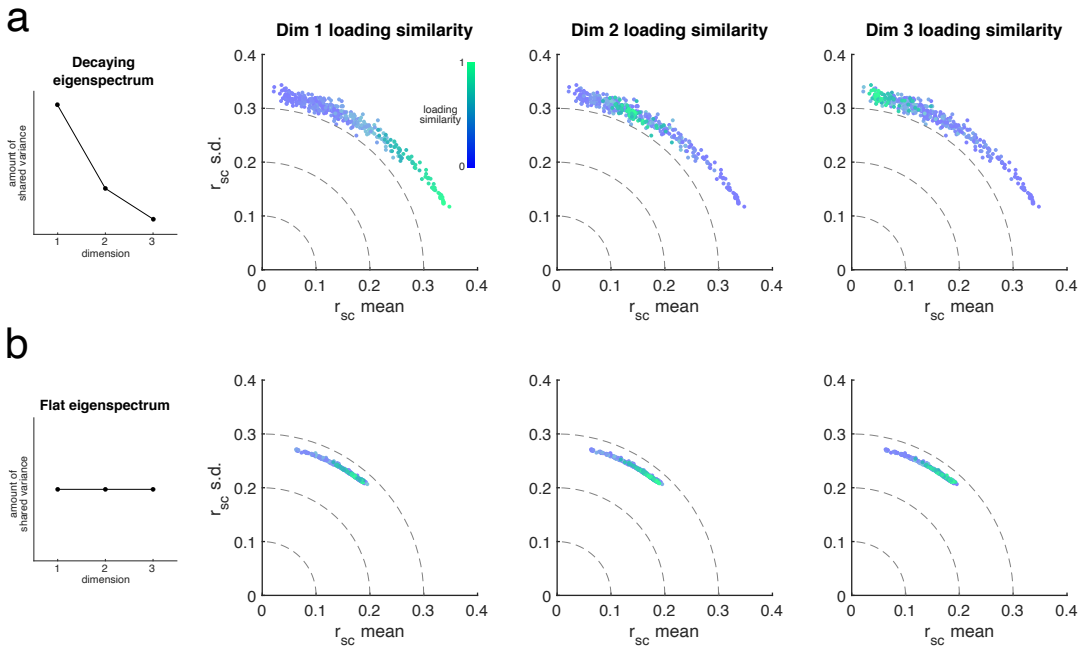
Artificial neural networks and deep learning have been hugely successful and are often used throughout a number of modern applications. Yet, deep learning models are somewhat brittle in that they are highly specialized to a specific task, are prone to adversarial attacks [173], and can make errors with even natural and small distribution shifts [162]. On the other hand, as humans we may not be as good as deep learning models at a highly specialized task, but we are much more robust and can generalize much more easily. Indeed, recent deep learning research has started to focus less on training the best model for a single task, and more on training general models that are good at many tasks (i.e., self-supervised learning; [151, 156, 157]). This was apparent with our comparison of the recent vision transformer (ViTs [147]) and the older convolution neural network (CNN) architectures in Chapter 5. Although our ViTs did not outperform CNNs at the image classification task, they were more robust to distribution shifts, had higher shape bias, and generalized better to other datasets (i.e., better transfer learning).

Can we take further inspiration from neuroscience and perception to train better deep learning models? Recent deep learning research has started to do so–the ViT [147] is a new computer vision architecture loosely inspired by the cognitive phenomena of attention. Additionally, the increasingly popular domain of self-supervised learning [151, 156, 157, 169, 170] is inspired by the fact that we as humans are constantly experiencing the external world, learning, and updating our beliefs and internal models without direct supervision. Self-supervised learning is an unsupervised technique that aims to learn general structure and feature representations in training data that will be useful for downstream tasks. Models pretrained with self-supervised learning, and then fine-tuned to a task achieve state of the art performance in both computer vision

[151, 156, 157] and natural language processing [169, 170]. The secret sauce to the success of self-supervised learning is massive amounts of data and extensive data augmentation. StyleAug (Chapter 5) is one such data augmentation technique that was inspired by shape bias in human vision, and showed excellent performance for the ViT architecture. Future research might focus on developing augmentation or training techniques that take inspiration from shared trial-to-trial variability (Chapters 2-4; [174]) or neural phenomena like slow drift (Chapter 3; [41, 42, 174]). The intersection of neuroscience and deep learning research has an exciting future–both fields can gain inspiration from the other and benefit from the cross-pollination of ideas.

7 Appendix

A Appendix for Chapter 2



Supplementary Figure 1: Relationship between pairwise metrics, loading similarity of each latent dimension, and the relative strengths of each dimension. Related to Figure 6.

In Fig. 5*e* and Math Note A, we considered the relationship between loading similarity and pairwise metrics when population activity was one dimensional. Here, we asked about the informativeness of loading similarity when population activity varies along multiple dimensions, and the impact of the relative strengths of each dimension (i.e., the shape of the eigenspectrum of Σ_{shared}, which specifies the amount of shared variance explained by each dimension).

We considered two cases. First, we considered an eigenspectrum that decays quickly, as has been widely reported in population recordings [49, 57, 66, 67, 71, 73, 76]. In this case, we found that the loading similarity of the strongest dimension (i.e., dimension with largest eigenvalue) was most informative about pairwise metrics, while the loading similarities of the other dimensions were less informative. Second, we considered a flat eigenspectrum. In this case, the loading similarities of each dimension were equally informative.

a. Loading similarity for a decaying eigenspectrum of the shared covariance matrix (Σ_{shared} in Supplementary Fig. 5*a*). We reproduced the simulation in Fig. 5 for a latent dimensionality of 3 and %sv=50%. For each 3-d model, we evaluated the r_{sc} mean and s.d., and then plotted the same point in 3 separate panels colored by loading similarity of each of the 3 different dimensions. The loading similarity of strongest dimension ('Dim 1') is very informative–high loading similarity implies high r_{sc} mean and low r_{sc} s.d. (green dots), whereas low loading similarity implies low r_{sc} mean and high r_{sc} s.d. (blue dots). This is the same relationship as shown in Fig. 5*e* for the case of one dimension. The loading similarities of 'Dim 2' and 'Dim 3' are less informative–in both cases, low loading similarity points (blue dots) are scattered throughout the arc. The only case when the loading similarity of 'Dim 2' or 'Dim 3' is informative is when either of them have a high loading similarity (green dots). This is informative because it implies that 'Dim 1' must have low loading similarity ('Dim 1' is blue for dots where 'Dim 3' is green; see Math Note E), implying low r_{sc} mean and high r_{sc} s.d. (continued on next page...)

Supplementary Figure 1 ***(previous page)*****:** (continued from previous page...)

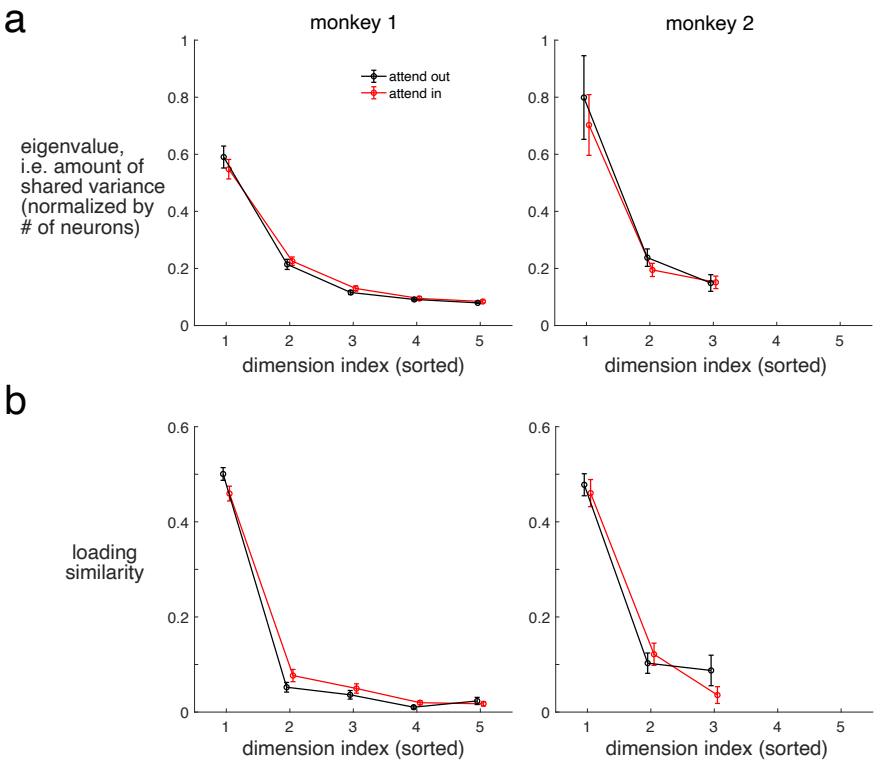
b. Same as panel **a** but for flat eigenspectrum across the three dimensions. In this case, r_{sc} mean will tend to be small and r_{sc} s.d. will tend to be large because: 1) all three dimensions contribute equally, and 2) it is not possible for all three dimensions to have high loading similarity, while multiple dimensions can have low loading similarity (Math Note E). However, knowing whether any of the three dimensions have high loading similarity can provide more specific information about r_{sc} mean and s.d. within this limited range (green dots tend to have high r_{sc} mean and lower r_{sc} s.d. in each panel).

Because most studies of population neuronal recordings have shown quickly decaying eigenspectra as in panel **a** [49, 57, 66, 67, 71, 73, 76], we recommend considering the loading similarity of the strongest dimension for concision and simplicity (as we do in Fig. 8*c*; and see eigenspectra in Supplementary Fig. 2). However, if it happens that the data have an eigenspectrum that decays slowly or has multiple dimensions that are very strong, then one may benefit by considering the loading similarities of additional dimensions as well.

This analysis also highlights how the shape of the eigenspectrum influences pairwise metrics. First, an exponentially-decaying eigenspectrum tended to have a higher r_{sc} mean and s.d. compared to its corresponding flat eigenspectrum (dots in panel **a** are farther from origin here than in panel *a*). This occurs because, for an exponentially-decaying eigenspectrum, an added dimension explains relatively little shared variance. Thus, the added dimension tends to result in only a small decrease in r_{sc} mean and s.d. On the other hand, adding a dimension to the flat eigenspectrum affects r_{sc} mean and s.d. as much as any other dimension, leading to larger changes (i.e., decreases) in r_{sc} mean and s.d. than in the case of an exponentially-decaying eigenspectrum.

Second, we observed a greater radial and angular spread for exponentially-decaying eigenspectra (panel **a**) compared to flat eigenspectra (panel **b**). This occurs because, when the eigenspectra are not flat, there is greater diversity in how the co-fluctuation patterns of different dimensions can contribute to r_{sc}. In other words, permuting the eigenvectors of three dimensions with equal eigenvalues (i.e., both dimensions explain the same amount of shared variance) results in the same model and same covariance matrix—yielding the same values for r_{sc} mean and s.d. However, permuting the eigenvectors of three dimensions with different eigenvalues will likely result in a different covariance matrix and different values of r_{sc} mean and s.d. Thus, for non-flat eigenspectra, the greater diversity by which co-fluctuation patterns can contribute to the shared covariance matrix leads to greater spread in the r_{sc} mean vs s.d. plots. The mathematical details regarding this observation are provided in Math Note D.

An implication of this analysis is that it is important to report the eigenspectrum shape whenever one reports dimensionality. Thus, considering both dimensionality and the eigenspectrum curve, instead of dimensionality alone, will lead to a more complete picture of the structure of population activity. Inspecting the eigenspectrum will also help determine whether assessing loading similarity in the strongest dimension is sufficient (panel **a**), or whether one needs to consider the loading similarities of other dimensions as well (panel **b**).



Supplementary Figure 2: Eigenvalues and loading similarity by dimension for V4 population activity. Related to Figure 8.

Although we observed only a modest change in dimensionality with attention (Fig. 8*c*), our simulations showed that the relative strength of each dimension (i.e., shape of the shared eigenspectrum) could alter the "effective dimensionality" of population activity and have large effects on pairwise metrics (Fig. 6*a*). Here, we asked whether the relative strengths of each dimension changed with attention. We also considered the loading similarities across different dimensions.

a. We found that the shape of the eigenspectra was qualitatively similar for 'attend in' and 'attend out' conditions (red and black curves have similar shape). In both conditions, the eigenvalues of the shared covariance matrix decayed (dot for each subsequent dimension was below dot for the previous dimension), indicating that a small number of dimensions were needed to explain the population-wide covariability.

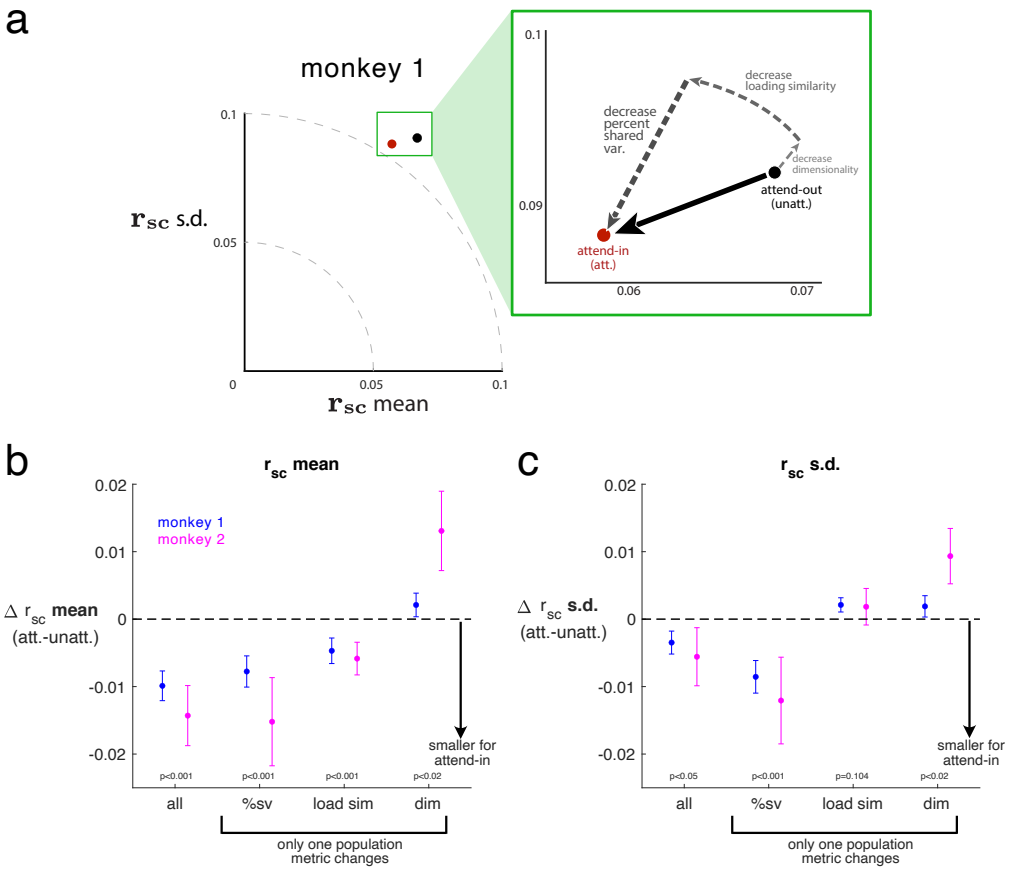
When comparing eigenspectra (i.e., the amount of shared variance explained by each dimension), one also needs to consider the firing rates under each condition. Mean firing rates tend to be higher for attend-in than attend-out trials. Higher firing rates typically correspond to higher spike count variance due to the Poisson-like firing of neurons. All else being equal, the higher mean firing rates imply higher levels of both shared variance and independent variance [69]. Thus, a direct comparison of the eigenspectra should be done with caution. Nonetheless, we plotted attend-in and attend-out together to relate our results to previous reports [49, 76]. Consistent with these studies, we found that attention decreased the strength of the strongest dimension (red below black dot for dimension index 1), though the magnitude of the decrease we observed was more consistent with [76] than [49]. Had we been able to equalize the mean firing rate across the two conditions, we likely would have observed an even greater difference between attend-in and attend-out. We note that the caveat described here for comparison of eigenspectra (i.e., the amount of shared variance) does not apply to comparisons of %sv (Fig. 8*c*) because %sv is normalized by the overall spike count variance.

(...continued on next page)

Supplementary Figure 2 ***(previous page)*****:** (...continued from previous page)

The eigenspectra were computed in the following way. We decomposed the V4 spike count covariance matrix into shared and independent components using factor analysis (see Methods). We then computed the eigendecomposition of the shared covariance matrix (Supplementary Fig. 5, $\Sigma_{\text{shared}} = U\Lambda U$). We found that eigenvalues (diagonal of Λ) tended to increase linearly with the number of neurons recorded; therefore, in order to combine across sessions, we normalized the eigenvalues by dividing by the number of neurons recorded in each session. After normalizing, we computed the eigenspectrum averaged across sessions and stimulus orientations. Because the dimensionality identified by cross-validation differed across sessions, there were a different number of sessions that contributed to each average. We did not plot mean eigenvalues when there were fewer than 5 sessions to average (i.e., dimensions ≥ 6 for monkey 1; dimensions ≥ 4 for monkey 2). Error bars indicate standard error. Data points have been jittered horizontally for visual clarity.

b. Loading similarity for 'attend-in' (red) and 'attend-out' (black) by dimension. Pooled across monkeys, the loading similarity for the first (i.e., strongest) dimension was larger for 'attend-out' than 'attend-in' (same result as Fig. 8*c*). We also observed some differences in loading similarity across the other dimensions in both monkeys. These differences could be important for specific scientific questions (see Fig. 9, for example). However, as we show in Fig. 6, Supplementary Fig. 1, and Math Note C, the first dimension plays the largest role in determining the r_{sc} distribution because it explains the greatest amount of shared variance.



Supplementary Figure 3: Quantifying the extent to which each population metric contributes to changes in pairwise metrics. Related to Figure 8.

In Fig. 8*c*, we observed changes in several population metrics with attention in V4 population responses. However, it was unclear to what degree the change we observed in each population metric contributed to the overall changes in pairwise metrics (Fig. 8*b*). In order to quantify this, here we used a population metric matching procedure to assess how much each individual change in a population metric contributed to the changes in a pairwise metric. We found that for these V4 data, %sv contributes the most, followed by loading similarity, and finally dimensionality. We illustrate these results in Fig. 8*d* (also reproduced here as panel **a** for convenience).

a. Reproduction of Fig. 8*d* to aid the interpretation of panels **b** and **c** here. For pairwise metrics, we observed decreases in both r_{sc} mean and s.d. with attention. For population metrics, we observed decreases in %sv, loading similarity, and dimensionality with attention.

b. Contribution of population metrics to changes in r_{sc} mean. For each recording session, we assessed how allowing all population metrics to vary ("all") or only a single population metric to vary between "attend-out" (unatt.) and "attend-in" (att.) influenced r_{sc} mean. The procedure for assessing this is detailed at the end of the caption ("Details of population metric matching procedure"). When only %sv or only loading similarity were allowed to vary, r_{sc} mean decreased with attention; when only dimensionality was allowed to vary, r_{sc} mean increased. When all population metrics were allowed to vary, r_{sc} mean decreased, consistent computations directly from data (Fig. 8*b*). Results for both monkeys were consistent; means and standard errors across sessions are shown.

(continued on next page...)

Supplementary Figure 3 ***(previous page)*****:** (continued from previous page)

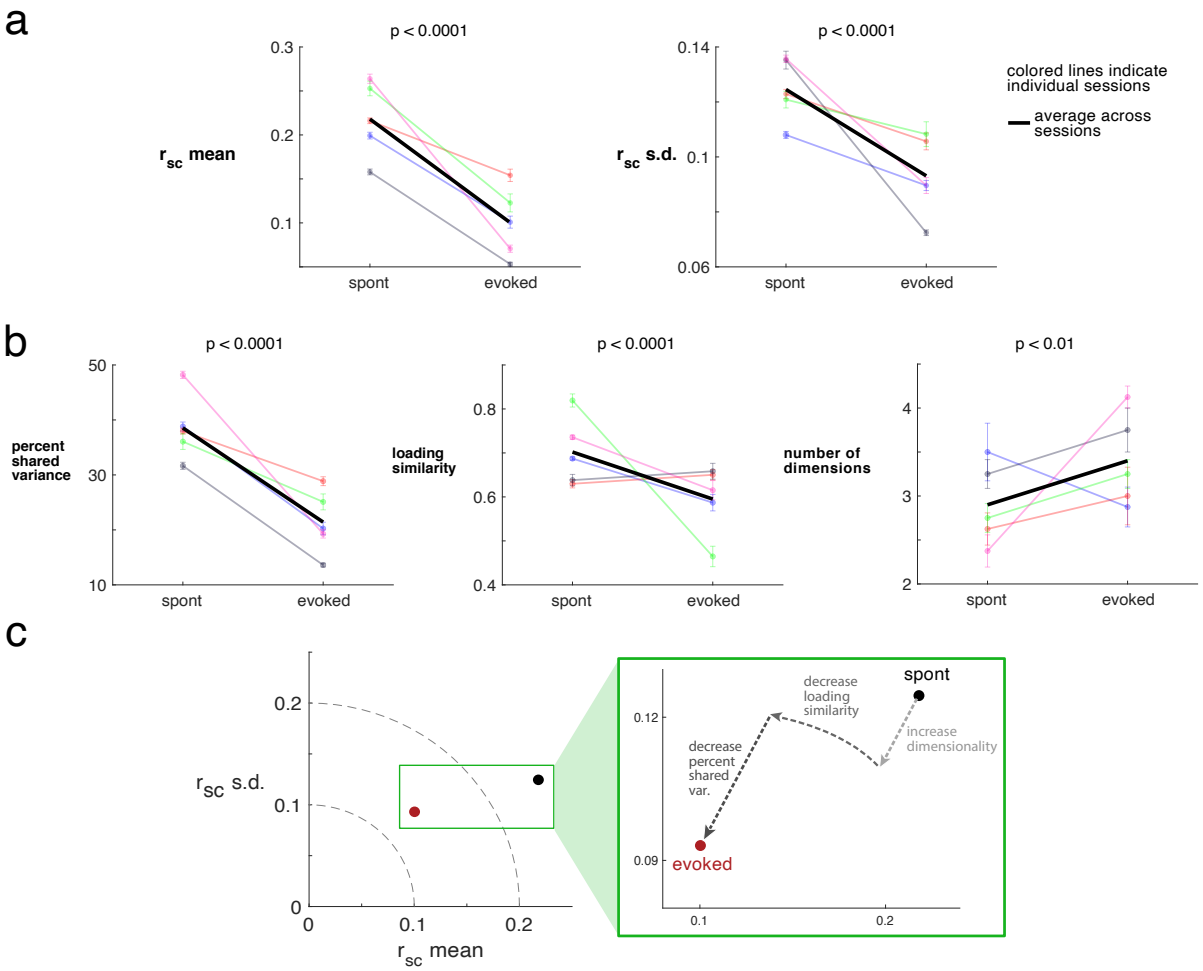
c. Contribution of population metrics to changes in r_{sc} s.d. Same format as **b**. When only %sv was allowed to vary, r_{sc} s.d. decreased with attention. When only loading similarity was allowed to vary, r_{sc} s.d. slightly increased (not significant). Also, when only dimensionality could vary, r_{sc} s.d. increased. When all population metrics were allowed to vary, we found that r_{sc} s.d. decreased with attention, consistent with our computations from data (Fig. 8*b*).

These results provide a systematic quantification of the illustration that relates pairwise and population metrics in V4 (panel **a**). Based on direction and magnitude of contributions, we conclude that for overall changes in pairwise metrics in these data: 1) %sv is most important, 2) followed by loading similarity, 3) followed by dimensionality. More generally, the population metric matching procedure (described below) provides a framework for assessing how changes in population metrics contribute to changes in pairwise metrics in recorded neuronal population activity.

Details of population metric matching procedure. Given two factor analysis (FA) models (e.g., fitted to two different experimental conditions), we first assess the overall change in pairwise metrics by computing r_{sc} mean and s.d. directly from the two fitted models (see Methods). In this case, all three population metrics are allowed to change between the two conditions, and contribute to the overall observed change in pairwise metrics (labeled "all" in the plots above).

Next, we use population metric matching to assess the contribution of each individual population metric change to the overall change in pairwise metrics. To do so, we choose one of the two fitted FA models (e.g., the model fitted to "attend-out") and systematically change the model such that one of its population metrics matches that of the other FA model (e.g., "attend-in"), while the other two population metrics remain the same. We then assessed the change in pairwise metrics between the base FA model (i.e., "attend-out") and the "matched" FA model (i.e., modified "attend-out" model). This allowed us to assess the change in pairwise metrics that would have resulted from a change in a single population metric.

For systematically modifying %sv, we scaled the eigenspectrum (see Methods) of the base FA model in order to match the %sv of the other FA model. For systematically modifying loading similarity, we replaced the co-fluctuation patterns (U in Supplementary Fig. 5*a*) in the base FA model (e.g., "attend-out") with the co-fluctuation patterns from the other FA model (e.g., "attend-in"). In cases where the dimensionality of the two models was different, we swapped the top k co-fluctuation patterns, where k is equal to the smaller dimensionality in the two models. For systematically modifying dimensionality, we removed dimensions from the base FA model if it had higher dimensionality than the other FA model, or added dimensions (after orthogonalization) from the other model to the base FA model if it had lower dimensionality. Because adding or removing dimensions changes the %sv, we then scaled the eigenspectrum to match the original %sv of the base FA model. These procedures allowed us, using two FA models fit to real data, to systematically vary one of the population metrics while keeping the other two the same and assess the contribution to a change in pairwise metrics.



Supplementary Figure 4: Relationship between pairwise and population metrics in V1 population responses. Related to Figure 8.

In Fig. 8, we assessed the relationships between pairwise and population metrics in V4 population recordings where a decrease in r_{sc} mean with spatial attention had been widely reported [1, 2, 10, 11, 13]. To demonstrate the applicability of the identified relationships to other brain areas, we applied the same analysis to population recordings in primary visual cortex (V1). Previous studies have shown that the r_{sc} mean is lower after stimulus onset (i.e., evoked activity) than before stimulus onset (i.e., spontaneous activity) in V1 [17, 69]. Here, we analyzed population activity recorded using Utah arrays in V1 (88 to 159 units per session, 112.2 on average) in three macaque monkeys [previously reported in 64, 78, http://dx.doi.org/10.6080/K0B27SHN]. Two monkeys had two recording sessions each, while the third monkey had a single recording session, for a total of 5 recording sessions. Animals were presented with 1.28s of oriented gratings (1 of 8 possible orientations) interleaved with 1.5s of a blank screen.

(continued on next page...)

Supplementary Figure 4 ***(previous page)*****:** (...continued from previous page)

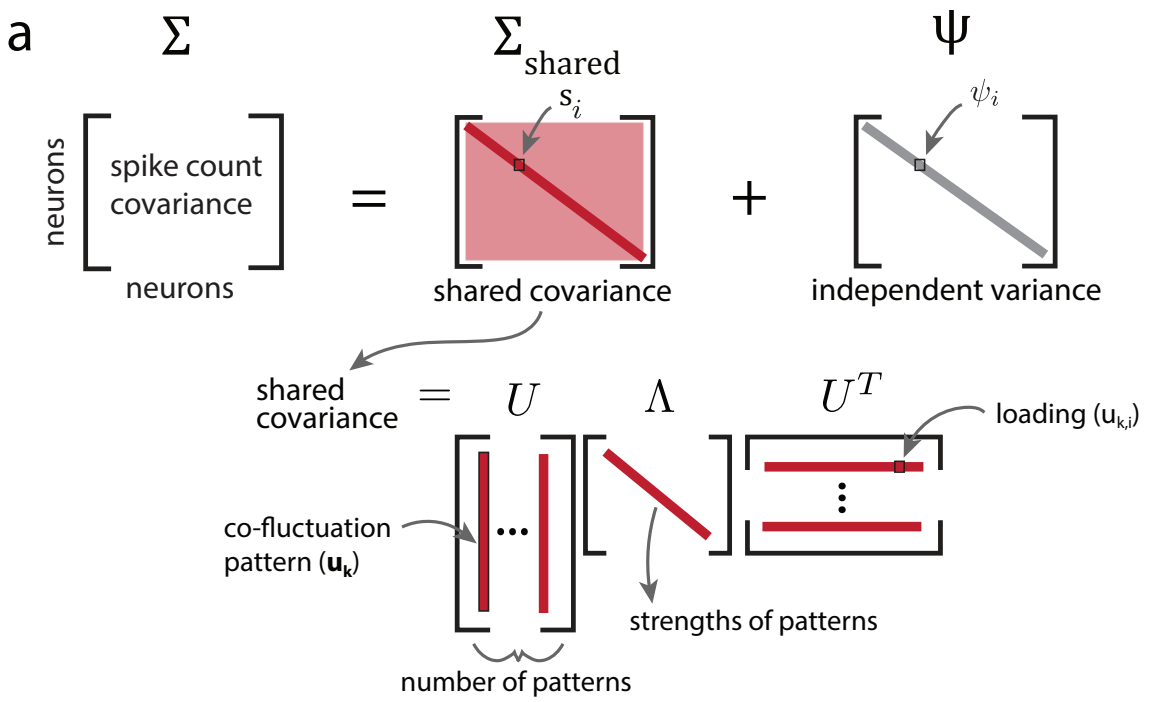
In V1, evoked activity had smaller r_{sc} mean and s.d. than spontaneous activity. We also found that evoked activity had smaller %sv and loading similarity, but larger dimensionality than spontaneous activity. Most changes in pairwise and population metrics between V1 "spontaneous" and "evoked" activity matched the direction of the changes in V4 "attend-out" and "attend-in", *except* for the change in dimensionality (cf. panel **c** and Fig. 8*d*). Taken together, our analyses of V1 and V4 population activity demonstrate that similar changes in pairwise metrics need not correspond to precisely the same changes in population metrics. In this case, measuring population metrics provided insight about the dimensionality of the population-wide variability that would not have been gleaned from changes in pairwise metrics alone.

a. The r_{sc} mean was smaller in evoked activity than in spontaneous activity (left panel; $p < 0.0001$) [17, 69]. We also found that r_{sc} s.d. was smaller in evoked activity than in spontaneous activity (right panel; $p < 0.0001$), which has not been previously reported.

b. Next, we assessed how population metrics changed between evoked and spontaneous V1 activity. Consistent with [69], we found that %sv was smaller for evoked activity than spontaneous activity (left panel; $p < 0.0001$). We also found that loading similarity for the dominant dimension was smaller for evoked activity than spontaneous activity (middle panel; $p < 0.0001$). Finally, we found that dimensionality was higher for evoked activity than spontaneous activity (right panel; $p < 0.01$). This result differed from a previous study in which dimensionality was lower for evoked activity than spontaneous in neural recordings from rat gustatory cortex and in a clustered network model [67]. This could be explained by a difference in sensory modality or the way in which dimensionality was measured.

c. Using the framework we developed to understand the relationships between pairwise and population metrics, the decrease in both r_{sc} mean and s.d. with evoked V1 activity corresponds to: 1) a decrease in %sv, 2) a decrease in loading similarity, and 3) an increase in dimensionality. The direction of the changes in pairwise metrics between spontaneous and evoked activity are the same as those we observed between "attend-out" and "attend-in" in V4 (Fig. 8*b*), as are the changes in %sv and loading similarity population metrics (Fig. 8*c*). However, the increase in dimensionality from V1 spontaneous to evoked is in the opposite of what we observed from "attend-out" to "attend-in" in V4 (Fig. 8*c*, right panel).

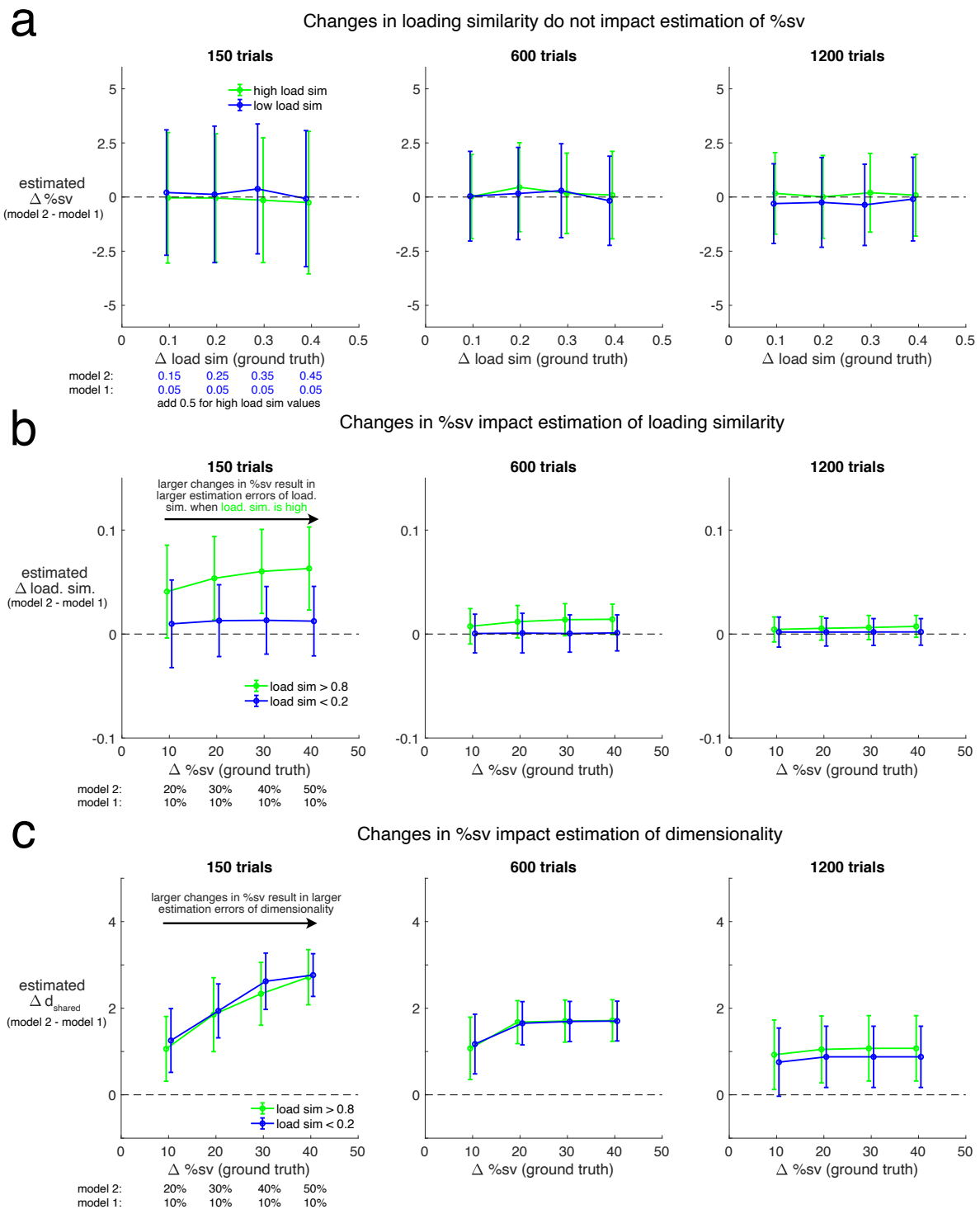
Methods. For evoked activity, we computed spike counts for each trial in the time period 160-260 ms after stimulus onset. For spontaneous activity, we computed spike counts during the blank screen in the 100 ms immediately prior to stimulus onset. We chose to use 100 ms bin sizes to match those used in [64]. We define spike counts during these two time periods during a trial as a "spont-evoked pair". Each recording session consisted of 400 repeats of a spont-evoked pair for each of the 8 oriented stimuli. For each session, we assessed changes in metrics for each orientation and computed the mean and standard error of the metric across the 8 orientations (transparent colored data points connected by lines). We also plot the average across the 5 sessions (thick black line). To compare metrics for spontaneous and evoked activity, we computed p-values across all 40 datasets (5 sessions, 8 orientations per session) using a paired t-test.



b

population metric	mathematical definition	intuition
loading similarity	$\text{loading sim.} = 1 - \frac{\text{var}(\mathbf{u_k})}{1\ /\ \text{n}}$	how similar a pattern is to one in which all neurons contribute equally
percent shared variance	$\%\text{sv} = \frac{1}{\text{n}} \sum_{i=1}^{\text{n}} \frac{\text{s}_i}{\text{s}_i + \psi_i} \cdot 100\%$	% of neuron's fluctuations explained by fluctuations of other neurons
dimensionality	$\text{dim.} = \text{rank}(U)$	number of co-fluctuation patterns

Supplementary Figure 5: Decomposition of the spike count covariance matrix and defining population metrics. a. We use factor analysis to decompose the spike count covariance matrix Σ into the sum of a low-rank shared covariance matrix Σ_{shared} and a diagonal independent variance matrix Ψ. The ith diagonal entry of Σ_{shared} (s_i) corresponds to the spike count variance that neuron i shares with other neurons in the population (i.e., shared variance), while the ith diagonal entry of Ψ_i corresponds to spike count variance of neuron i that cannot be explained by the other neurons (i.e., independent to neuron i). We can further decompose Σ_{shared} via an eigendecomposition to extract the co-fluctuation patterns (i.e., the eigenvectors) and the strength of each latent co-fluctuation (i.e., the eigenvalues). **b**. The population metrics used in this study are loading similarity, percent shared variance (%sv), and dimensionality.



Supplementary Figure 6: Characterizing how changes in one population metric can impact the estimates of another population metric. Related to Figure 8.

(continued on next page...)

Supplementary Figure 6 ***(previous page)*****:** (...continued from previous page)

In the main text, we related population metrics to pairwise metrics by systematically changing a single population metric and measuring the resulting changes in r_{sc} mean and r_{sc} s.d. (Figs. 5. However, in real neuronal data, multiple population metrics could change together between experimental conditions (e.g., see Fig. 8 and Supplementary Fig. 4). When we measure that multiple population metrics changed, it could be the case that a change in one population metric impacted the estimates of the other population metrics (e.g., we could have measured a change in multiple population metrics when only one metric truly changed). This can affect the precision by which we can distinguish population metric changes in real neuronal data.

Here, we assessed this by systematically changing one population metric while keeping the other two population metrics constant. We then simulated data and fit factor analysis (FA) to the data to obtain the population metrics. We examined in turn each of the three population metrics under conditions when they did not actually change (but one of the other metrics did). If there were no dependencies between estimates of population metrics, then all the vertical values in panels **a-c** would be 0. We found that this was the case for estimates of %sv were under conditions in which the true loading similarity changed (**a**). However, estimates of loading similarity and dimensionality were affected by true changes in %sv. Increasing the number of simulated trials reduced the estimation error caused by true changes in %sv (**b**, **c**). These findings allow us to better interpret changes in population metrics estimated from neuronal activity.

a. Estimation error in %sv due to changes in loading similarity. "Model 1" and "model 2" had the same dimensionality (1) and %sv (20%). The only difference between the two models was their loading similarity. We varied how different the loading similarity was between the two models (horizontal axis), while assessing how different was the estimated %sv across the two models (vertical axis). We found that %sv estimates remained unaffected in the presence of true changes in loading similarity (all changes in %sv are near 0). This was true for both low loading similarities where "model 1" had loading similarity of 0.05 (blue) and high loading similarities where "model 1" had loading similarity of 0.55 (green). As we simulated more trials, estimates of %sv became more precise (error bars decrease in size going from left to right panels). Error bars show means and standard deviations across simulations.

b. Estimation error in loading similarity due to changes in %sv. "Model 1" and "model 2" had the same dimensionality (1) and loading similarity. The only difference between the two models was their %sv. We varied how different the %sv was between the two models (horizontal axis), while assessing how different was the estimated loading similarity between the two models (vertical axis). We found little changes in estimates of loading similarity when the true loading similarity was low (blue points). However, we found larger changes in estimates of loading similarity when the true loading similarity was high (green points). The the size of the change increased with larger true changes in %sv. To understand this, recall that there are relatively few ways to have high loading similarity (e.g., all loadings must be the same to have loading similarity of 1), while there are many ways to have low loading similarity (Math Note E). Thus, high loading similarities are more likely to be underestimated than low loading similarities. This underestimate tends to be larger when %sv is low than when %sv is high. However, increasing the trial counts reduced the estimation error of loading similarity (vertical values closer to 0 going from left to right panels).

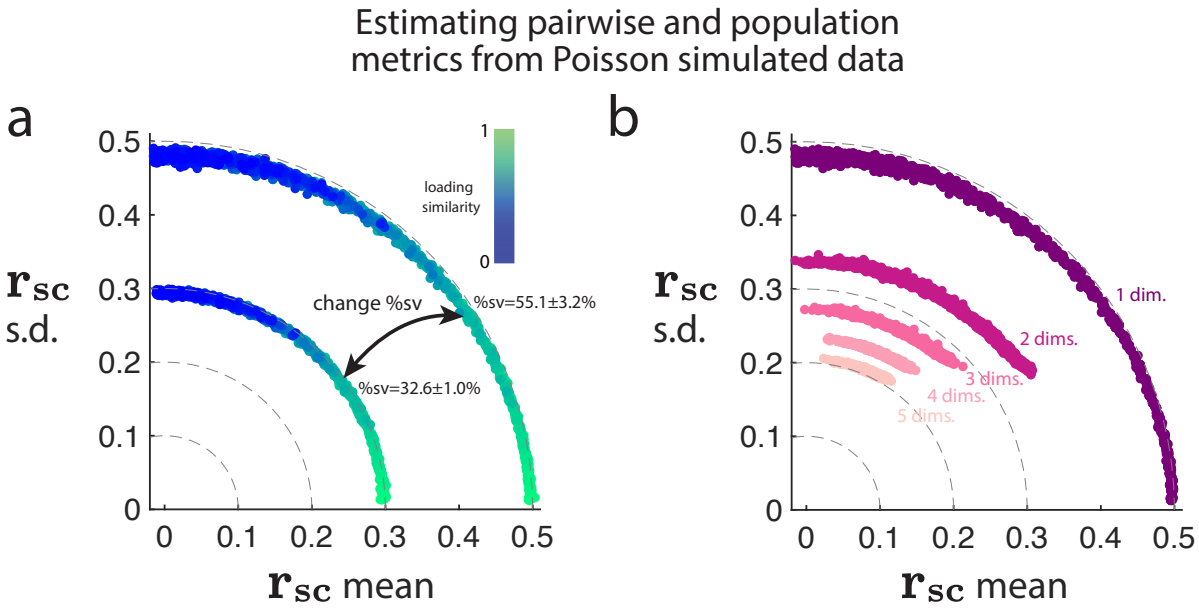
(...continued on next page)

Supplementary Figure 6 ***(previous page)*****:** (...continued from previous page)

c. Estimation error in dimensionality due to changes in %sv. "Model 1" and "model 2" had the same dimensionality (5; with eigenspectrum defined as $\lambda_k = e^{-0.75k}$) and loading similarity. The only difference between the two models was their %sv. We varied how different the %sv was between the two models (horizontal axis), while assessing how different was the estimated dimensionality (d_shared; see Methods) between the two models (vertical axis). We found changes in the estimates of dimensionality between "model 1" and "model 2", and the size of the change increased with larger true changes in %sv. This can be understood by the fact that dimensions with small eigenvalues can be difficult to recover when fitting FA to data, particularly when %sv is low. However, increasing the trial counts reduced the estimation error of dimensionality (vertical values closer to 0 going from left to right panels).

These results have important implications for interpreting estimated changes in population metrics in real neuronal data. First, because estimation error depends on trial count, one should equalize the number of trials across conditions in order to make fair comparisons across conditions using population metrics. Second, when changes in %sv are large and trial counts are small, one may need to be careful in interpreting estimated changes in loading similarity and dimensionality. For trial count, the key quantity to consider is the ratio of observed trials to the number of recorded neurons. In the simulations above, we used 30 neurons–the left column represented a ratio of 5x trials to neurons, the middle column represented 10x, and the right column represented 20x.

In our V4 data (Fig. 8), most sessions had 10 times (or more) the number of trials as the number of neurons (ratio of trials to neurons: 9.90 ± 0.66 for monkey 1, 27.60 ± 2.68 for monkey 2). We observed a difference in %sv of ≈3% between "attend-out" and "attend-in". Based on the results in panels **a** and **b**, the differences we measured in %sv and loading similarity in our V4 data are unlikely to be due to estimation error. Based on panel **c**, the small difference we measured in dimensionality in our V4 data could potentially be explained by a change in %sv, if the only true change between conditions was in %sv (and not loading similarity or any other aspect of the population activity).



Supplementary Figure 7: Relationships between pairwise and population metrics hold for metrics estimated from Poisson simulated data. Related to Figure 5

In our simulations and analytical derivations, we created covariance matrices with specified population metrics from which we computed pairwise metrics (Fig. 5). However, when assessing population metrics in neuronal recordings, one needs to fit a factor analysis (FA) model to data. Here, we simulated Poisson data to assess whether the relationships between pairwise and population metrics were impacted by: 1) needing to estimate metrics from data, and 2) the mismatch between the linear-Gaussian assumption of FA and the Poisson-like statistics of neuronal activity. We found that the relationships between pairwise and population metrics were very similar to those shown in Fig. 5.

a. Estimating loading similarity and %sv. We simulated data from a model with a single co-fluctuation pattern and Poisson observations (see details at the end of the caption). In the ground truth models, we varied loading similarity smoothly between 0 and 1 and chose %sv equal to 30% or 50%. We estimated r_{sc} mean and s.d. from the simulated data. To estimate population metrics, we fit the FA parameters to the same simulated data. We then plotted estimates of pairwise metrics and colored or labeled points according to the *estimated* population metrics (as opposed to the ground truth population metrics used to generate the data). We found that as estimated loading similarity increased, r_{sc} mean increased and r_{sc} s.d. decreased (blue to green). We also found that as estimated %sv increased, r_{sc} mean and s.d. both increased (inner arc with %sv=32.6±1.0% to outer arc with %sv=55.1±3.2%). These results are consistent with Fig. 5*e-f*.

b. Same as **a**, but fixing %sv=50% and varying the dimensionality of the ground truth model, with a flat eigenspectrum (corresponding to Fig. 5*g*). We colored points according to *estimated* dimensionality as opposed to the ground truth dimensionality. We found that as the estimated dimensionality increased, r_{sc} mean and s.d. both tended to decrease (purple outer arc with dim=1 to salmon inner arc with dim=5), consistent with Fig. 5**g**.

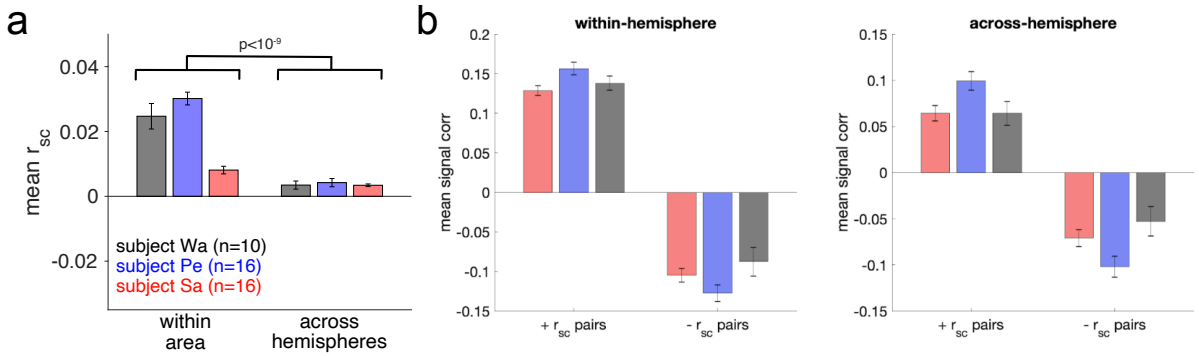
(continued on next page...)

Supplementary Figure 7 *(previous page)*: (...continued from previous page)

The results here, based on estimating factor analysis parameters from Poisson simulated data, are qualitatively the same as those in the main text (Fig. 5*e*-*g*) and analytical derivations (Appendices). This indicates that the relationships between pairwise and population metrics are robust to: 1) having to estimate these metrics from data and 2) the Poisson-like variability of neuronal activity.

Simulating from a Poisson observation model. According to FA, the observations x (i.e., spike counts) have a linear-Gaussian relationship with latent variables z (which represent shared activity among neurons): $\mathbf{z} \sim N(0, I)$ and $\mathbf{x}|\mathbf{z} \sim N(L\mathbf{z} + \mu, \Psi)$). We fit the FA parameters to data simulated from a Poisson observation model. We generated Poisson spike counts for 30 neurons as follows. For neuron i, we sample from from $x_i|\mathbf{z} \sim Poisson(ReLu(L_{i,:}\mathbf{z} + \mu_i))$, where $ReLu$ indicates a rectified linear unit, and $L \in R^{30\times d}$ is the loading matrix with $L_{i,:}$ as the i^{th} row. We set $\mu_i = 10$ for each neuron, a typical average firing rate (10 Hz) for neurons across many areas of macaque cortex (assuming a 1 second time bin). We consider the asymptotic case by simulating many trials for each model (corresponding to a single dot in panels *a* and *b*; see Methods for how model parameters are randomly chosen). We consider estimation from limited data in Supplementary Fig. 6. We drew 6000 samples (i.e., 6000 trials) of $\mathbf{x}$ from the Poisson observation model. Thus, this procedure generated a data matrix $X \in R^{30\times 6000}$ of simulated spike counts, which we then used to estimate pairwise and population metrics.

B Appendix for Chapter 4

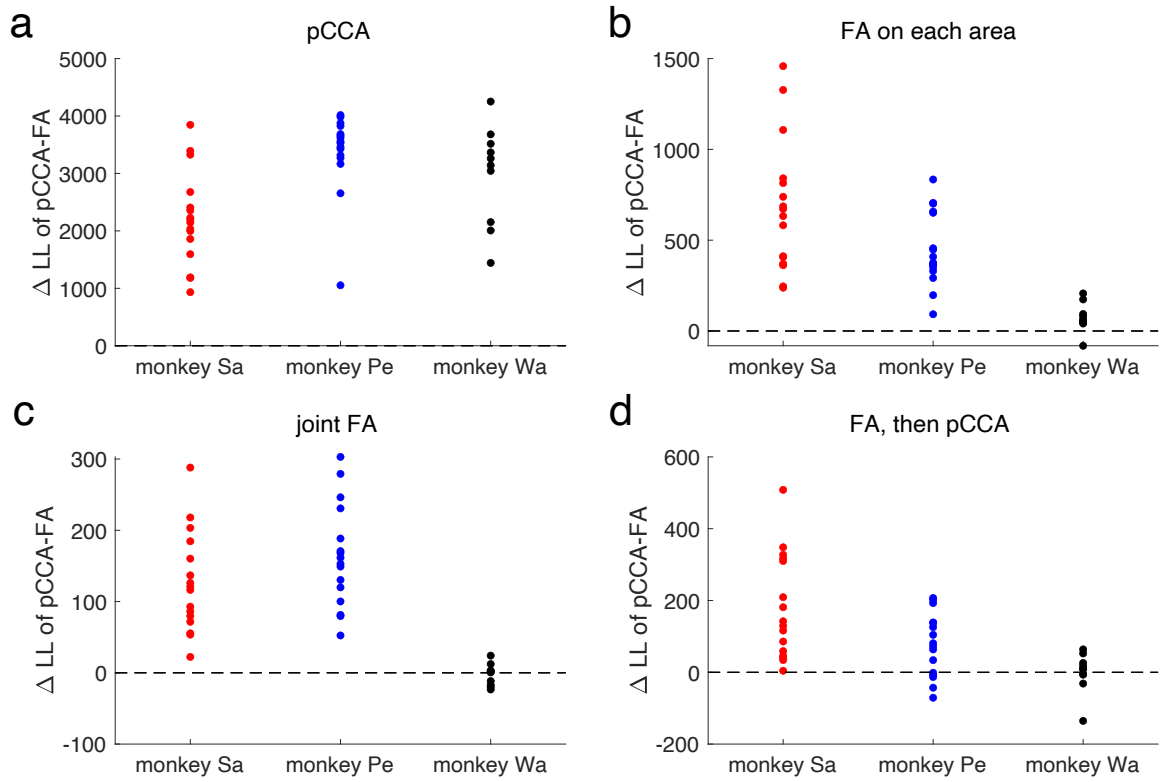


Supplementary Figure 8: Relationship between signal and noise correlation (r_{sc}) in within-area and across-hemisphere pairs. Related to Fig. 1.

Here, we consider the r_{sc} mean, a commonly used metric to summarize the r_{sc} distribution and characterize pairwise neuronal correlations. We also asked whether r_{sc} (i.e., noise correlation) was related to signal correlation between pairs of neurons 1) in the same brain area and 2) across hemispheres.

a. Mean r_{sc} aggregated across sessions for each subject in our data. The r_{sc} mean is significantly larger for within-hemisphere pairs than across-hemisphere pairs. This might lead one to conclude that there is shared variability between neurons in the same area, but not across hemispheres. However, r_{sc} mean is a coarse metric in that it averages across many pairs of neurons and could therefore be veiling population covariability structure [104]. Error bars indicate standard error across sessions.

b. Relationship between signal and noise correlation. First, we determined significantly positive r_{sc} pairs and significantly negative r_{sc} pairs for both within-area and across-hemisphere pairs of neurons. We then computed the signal correlation for the two groups (positive r_{sc} pairs and negative r_{sc} pairs). Aggregated across sessions, we found that positive r_{sc} pairs had positive and significantly higher signal correlations than negative r_{sc} pairs, which had negative signal correlations. This was true for both within-area pairs and across-hemisphere pairs in all three subjects ($p<0.001$). Thus, by using metrics that are finer-grained than r_{sc} mean, we start to see that that are interesting shared fluctuations among pairs of neurons in different hemispheres.



Supplementary Figure 9: pCCA-FA provides a better fit to data than alternative models. Related to Figs. 15 and 16.

We proposed pCCA-FA as a model that allows the partitioning of global and local shared variability. To do so, pCCA-FA assumes both low-rank global (across-area) interactions and low-rank local (within-area) interactions. In Fig. 15*a*-*b*, we found that pCCA-FA is more data-efficient than pCCA in recovering ground truth. Thus, it is reasonable to expect that pCCA-FA might provide better fits to our neural data than alternative models. Here, we detail the assumptions of alternative models under consideration and compare their performance with pCCA-FA. Overall, we found that pCCA-FA provided better fits to data (in terms of higher log likelihood) than alternative models in most cases. In addition to fitting better to neural data, pCCA-FA also allows for a clean distinction between global and local interactions, providing better scientific interpretability (e.g., computing metrics and inferring latent variables) than the alternative models that we considered.

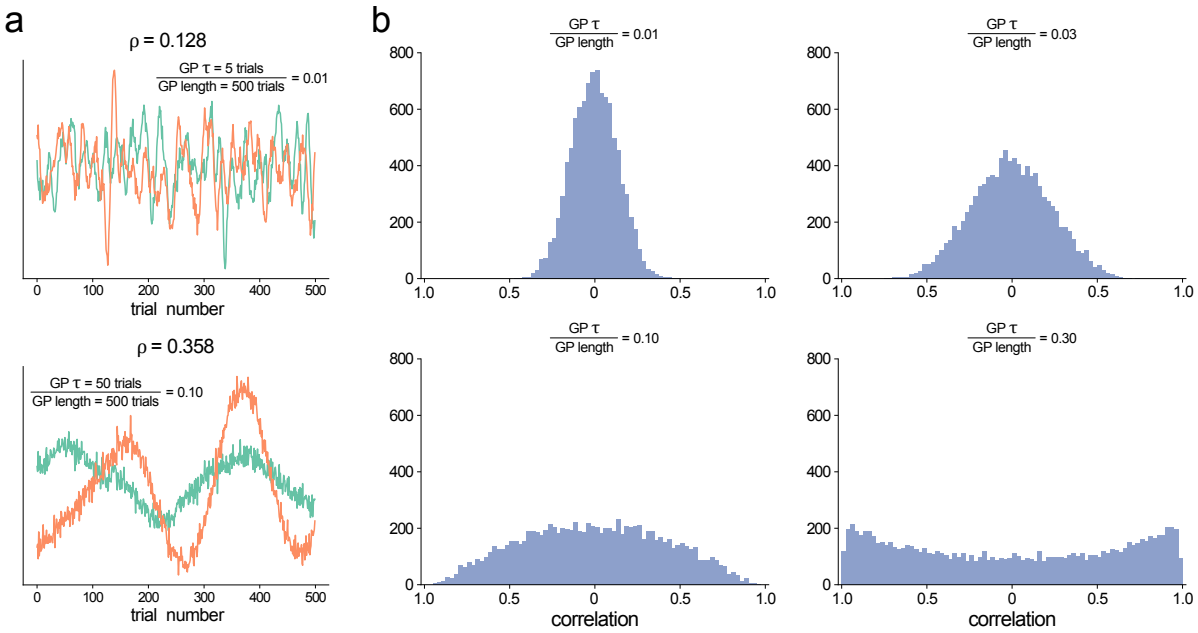
a. Probabilistic canonical correlation analysis (pCCA; see Methods) assumes low-rank across-area interactions and full-rank within-area interactions. We found that pCCA-FA provided better fits to our data than pCCA in all cases (Fig. 15*a*-*b*; Satchel $p<0.0001$, Pepe $p<0.0001$, Wakko $p<0.0001$; paired sample t-test). This is due to limited trial counts and the subsequent large estimation errors in estimating the full-rank within-area interactions.

b. We tested simply fitting factor analysis (FA) to each area. This model assumes no global (across-area) interactions, and low-rank within-area interactions. Note that this model is equivalent to pCCA-FA, when the global dimensionality is fixed to 0. We found that pCCA-FA fit better in almost all cases, except for a single session in subject Wa (Satchel $p<0.0001$, Pepe $p<0.0001$, Wakko $p<0.02$; paired sample t-test).

c. We tested fitting FA jointly to all recorded neurons, without consideration that they are recorded from different brain hemispheres. In other words, there is no separation of global and local interactions into different subspaces as in pCCA-FA, but rather they are both capture as part of a single subspace defined by FA. We found that pCCA-FA provided better fits to data for subject Sa (red; $p<0.0001$) and Pe (blue; $p<0.0001$). Both models fit similarly for subject Wa (black; $p>0.05$), who had many neurons in one area (60 per session) but few recorded neurons in the other area (20 per session). Subjects Pe and Sa had many neurons in both areas (60 to 80). This meant that modeling the cross-covariance (i.e., the interactions between areas) was less important for subject Wa than it was for subjects Sa and Pe.

d. We tested sequentially fitting FA to each area, and then fitting pCCA to the latents inferred from FA on each area. This model first captures all low-rank variability in an area (which could be either local or global), and then finds global interactions that exist within those low-rank subspaces. This procedure reduces dimensionality first (FA) before fitting pCCA, which can help prevent the overfitting and poor heldout performance seen in pCCA (panel **a**). Still, we found that pCCA-FA fit better for all sessions in subject Sa (red; $p<0.0001$) and most sessions in subject Pe (blue; $p<0.004$). For similar reasons as in **c**, pCCA-FA and sequential FA+pCCA provided similar fits to data for subject Wa (black; $p>0.05$).

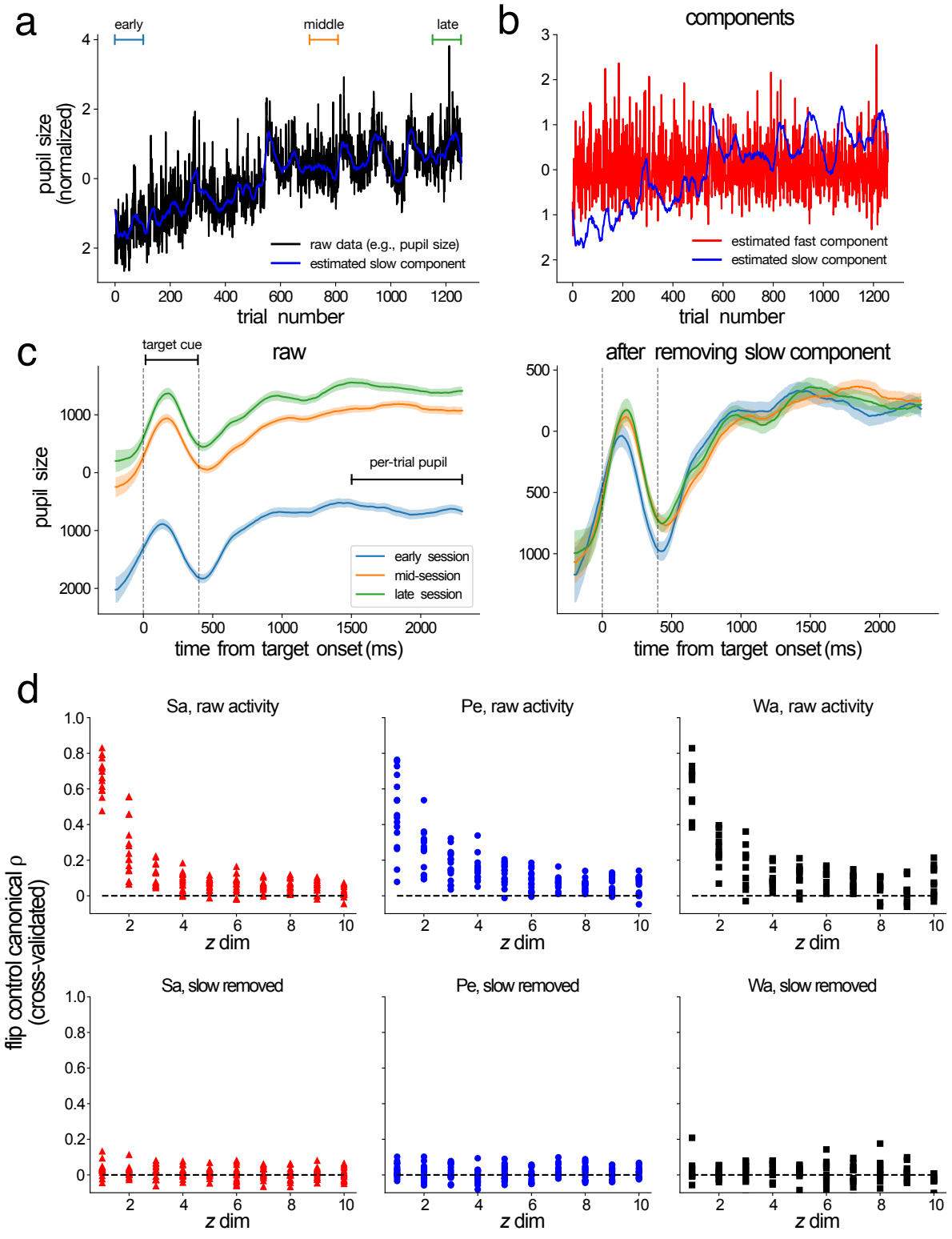
Method. For each model, we perform 10-fold cross-validation to choose dimensionalities, and then assess the held-out log likelihood of our neural data under the model (i.e., the marginal data likelihoods of observed neural activity). We assess this for each subject (different colors), and each session (each dot). We compare the difference in log likelihood (ΔLL) of pCCA-FA relative to the alternative model under consideration. Thus, any value above 0 means pCCA-FA is a better fit to the data than the alternative model.



Supplementary Figure 10: Slow-timescale (autocorrelated) fluctuations in data could lead to potentially spurious correlations. Related to Figs. 16 and 17.

a. Two independently drawn Gaussian Processes (GPs) can have non-zero correlations. Since they are independently drawn samples, they would ideally have a correlation close to 0. However, given a fixed sample size (500 trials here), two independent GPs with slower timescale fluctuations (bottom; τ = 50 trials) tend to have larger correlations (rho = 0.358) than two independent GPs with faster timescale fluctuations (top; τ = 5 trials).

b. If we repeat the procedure in **a** many times, we can obtain a null distribution of correlations we would expect to see between two independent GPs, for a given ratio between the GP τ (timescale) and the GP length. The null distributions have have a mean of 0. However, as the ratio of GP τ to GP length increases, the width of the null distributions increases. Thus, any given draw of two *completely independent* GPs can have a large (either positive or negative) correlation value. Thus, whenever fluctuations in our data are slow relative to the number of samples (i.e., trials), we might expect large correlations just by chance–these are often called spurious or nonsense correlations [133]. This is problematic for correlation analysis, and especially when using CCA, who's objective is to maximize correlation. Thus, we preprocess data to remove slow fluctuations in our data (Supp. Figs. 11 and 13).



Supplementary Figure 11: Estimating slow and fast components of neural activity and pupil. Related to Figs. 16 and 17.

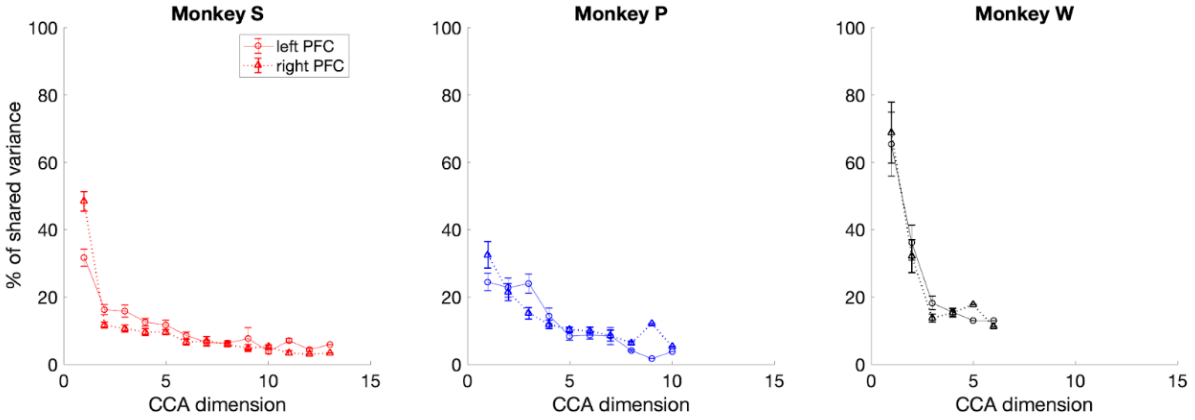
a. The slow component (blue) is estimated using a moving average of 25 trials on the raw neural activity or, as shown here, pupil size (black). This procedure is done separately for each neuron

spike count data, and for pupil size.

b. The fast component is the residual, i.e., the raw data minus the slow component (red).

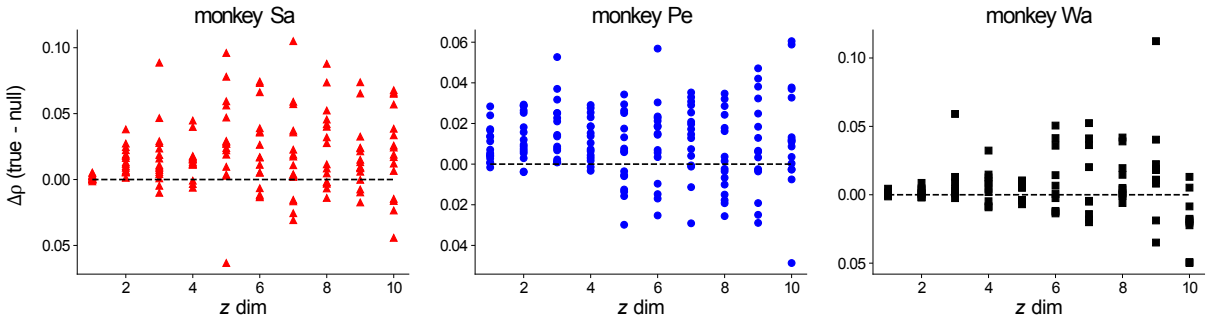
c. The slow component captures slow baseline fluctuations in pupil size or neural activity.

d. Limited data and potentially spurious correlations that arise with slow components (Supp. Fig. 10; [133]). Here, we run control analysis to demonstrate the importance of removing slow components before fitting pCCA-FA (or performing any correlation analysis for that matter). To generate a null distribution, we flip the trial order of neural activity in one hemisphere of the brain. Our assumption is that this flipping breaks any trial-to-trial correspondence and should thus result in correlations close to 0. Any recovered correlations would be spurious (induced by slow timescale fluctuations; Supp. Fig. 10). If we perform the flip control on raw neural activity, and fit pCCA-FA we find that we recover large canonical correlations (top row), indicating that the global dimensions could be picking up on spurious correlations. However, if we remove the slow components from neural activity (i.e., use the fast components), and then perform the flip control and fit pCCA-FA, we recover canonical correlations close to 0. This suggests that fitting pCCA-FA to the fast components would not be recoving spurious correlations. Therefore, most analyses in the remainder of the work focuses on the estimated fast component (though see Supp. Fig. 13 for an analysis of slow components).



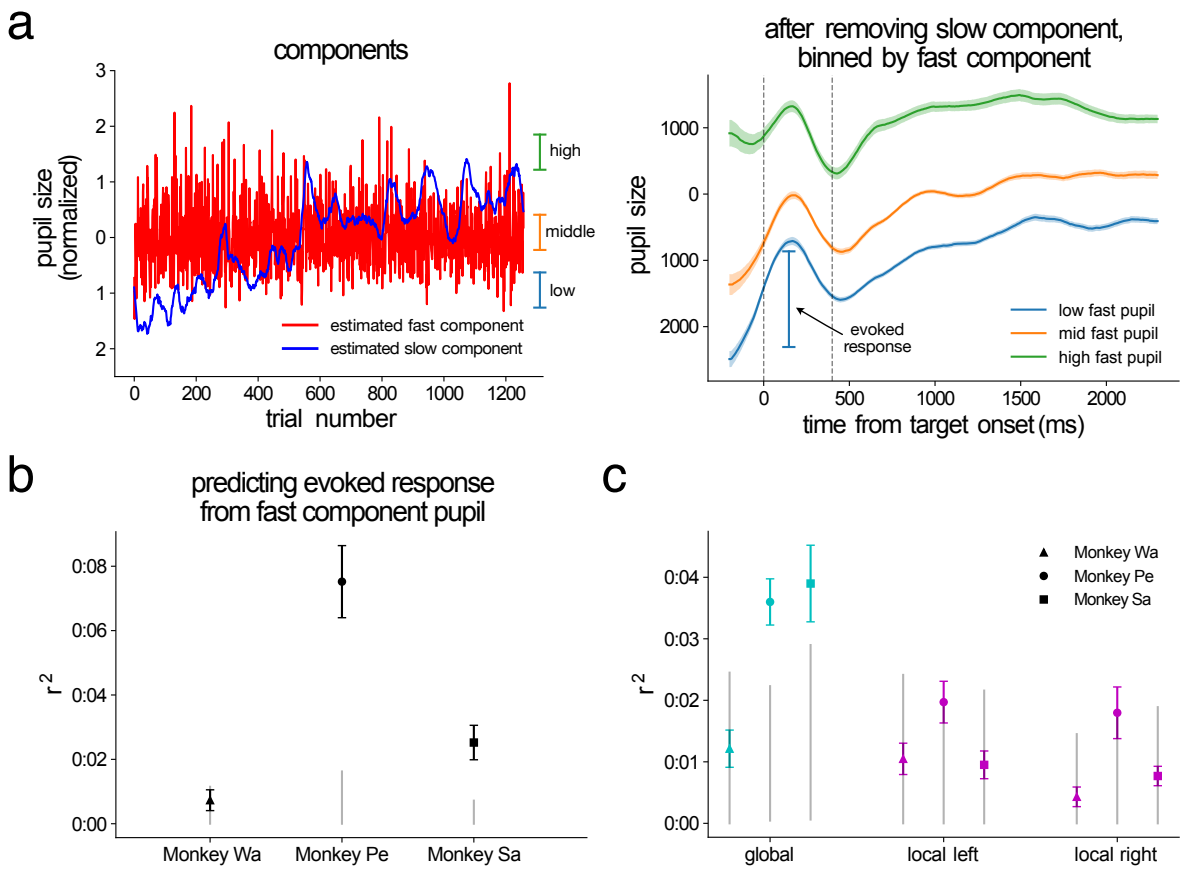
Supplementary Figure 12: The most correlated dimensions in the global subspace also explain the most variance. Related to Fig. 16.

The global latent variables in pCCA-FA are defined by the dimensions in area x and area y that are most correlated with one another (as in pCCA and CCA). In contrast to PCA or factor analysis which find dimensions that maximize variance or covariance respectively, there is no requirement for how much variability each dimension explains in CCA or pCCA-FA. We found that the top global dimensions (i.e., the most correlated dimensions across areas) of pCCA-FA explained the most variance. We computed the % of global shared variance in area x and area i as: $\frac{tr(W_{x,i}W_{x,i}^T)}{tr(W_xW_x^T)}$, where $tr(\cdot)$ is the trace, W_x is the loading matrix for area x onto the global subspace, and $W_{x,i}$ is the i^{th} column of W_x.



Supplementary Figure 13: Slow-timescale global interactions exist in neural activity. Related to Fig. 16.

Most of the work in this study focuses on faster-timescale trial-to-trial variability because of difficulties in interpretability that can arise due to spurious correlations in autocorrelated time series (Supp. Figs. 10 and 11*d*). However, here we asked whether the slow-timescale autocrrelated fluctuations are significantly above chance. We fit pCCA-FA to the estimated slow components of neural activity (Supp. Fig. 11*a*) and asked whether the estimated canonical correlations (true ρ) were above those estimated when fitting pCCA-FA to the flip control of the same data (null ρ). Indeed, we found that the true ρ were above the null ρ ($\Delta\rho$ true-null > 0) across subjects for most global dimensions (Sa $< 10^{-6}$, Pe $< 10^{-6}$, Wa $p = 0.017300$; paired sample t-test). This suggests that slow-timescale global (across-hemisphere) interactions do indeed exist in PFC neural activity above the chance level.



Supplementary Figure 14: Pupillary evoked response can be predicted from fast pupil components and global latents. Related to Fig. 17.

a. Left: example fast and slow components of pupil size, computed as described in Methods and Supp. Fig. 11. Removing the slow components removes baseline fluctuations in the pupillary evoked response across the session (Supp. Fig. 11). Right: After removing the slow component, binning trials by fast pupil value starts to reveal that there is an interaction between the fast pupil value on each trial and the evoked response amplitude (change in pupil size from 100 ms pre-target presentation to 100 ms post-target presentation).

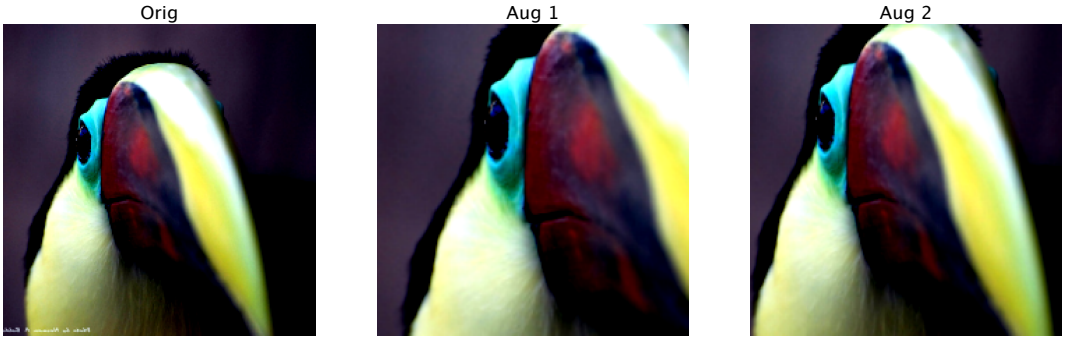
b. Per-trial fast pupil values predict the per-trial evoked response. Error bars indicate standard error computed across sessions. Null distributions are computed from predicting the evoked response on session i from fast pupil on session j, where $i \neq j$. Light gray bars indicate 95% confidence intervals of the null distributions.

c. Prediction of per-trial evoked response the latent variables computed from pCCA-FA. Global latent activity indeed does predict the evoked response for subjects Sa and Pe, while local latent activity does not for any subject. Error bars and null distributions are computed analogously as in panel *b*.

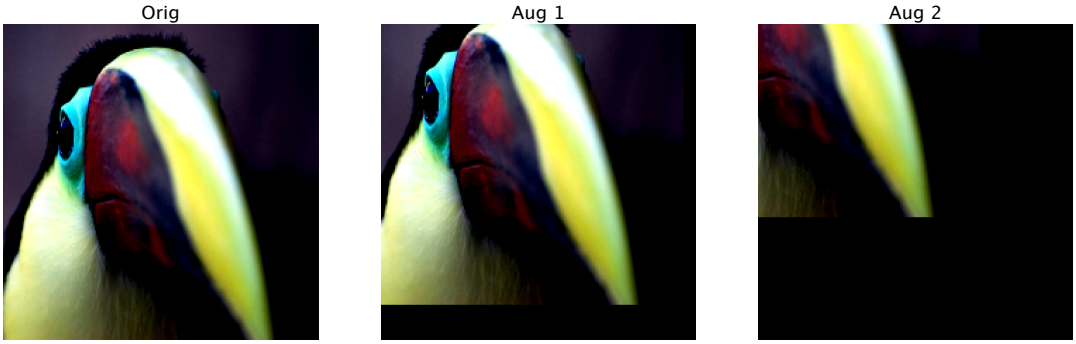
C Appendix for Chapter 5

Augmentation examples

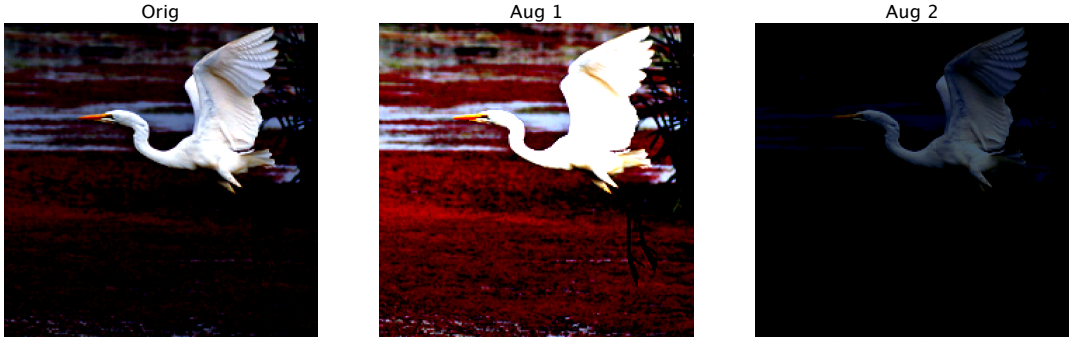
Here, we provide examples images of all augmentations that we tested (Supp. Fig. 15). In the examples, "Orig" represents the image after Inception-style preprocessing (random crop with a large scale (0.5,1.0) and horizontal flip). "Aug1" represents the augmentation we used for training when no JSD loss was used. When a JSD loss was used, we trained with the 3 versions of the displayed image ("Orig", "Aug1", "Aug2"; see Fig. 18, Algorithm 1, and Equation 26). Where applicable, the PyTorch torchvision transform used for augmentation is described in the caption.



(a) Cropping augmentation. RandomResizedCrop(scale=(0.25,1.0)).



(b) Translate augmentation. RandomAffine(rotate=0,translate=(0.5,0.5),scale=None,shear=None).



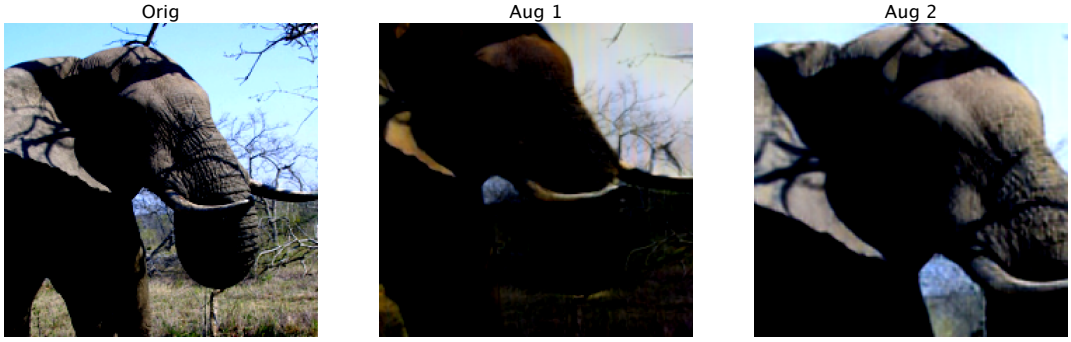
(c) Color augmentation. ColorJitter(brightness=0.4,contrast=0.4,saturation=0.2,hue=0.1).

(d) AugMix [143]. We used the PyTorch image models (timm; [175]) implementation of AugMix. As in the original AugMix work, the set of transformations used were mutually exclusive with the transfomations present in the Imagenet-C dataset [143].

(e) RandAugment [155]. We used the PyTorch image models (timm; [175]) implementation of RandAugment. We again removed any transformations that overlapped with corruptions in the ImageNet-C dataset [143].

(f) Neurofovea. We adapted the Neurofovea transformation described by Deza *et al.* [176]. For the foveation step, we simply mixed the original image and style-transferred noise image using a weighted mask. The original image received larger weights for points closer to the foveation point with exponentially decaying weights for farther pixels, with a minimum weight of 0.25. The style-transferred noise image received weights that were 1 minus the weights for the original image. Deza *et al.* [176] used a more perceptually accurate transformation. However, the computational cost of that approach made it infeasible as an augmentation strategy.

(g) StyleAug (ours). The augmentation is detailed in section 4 of the main text.

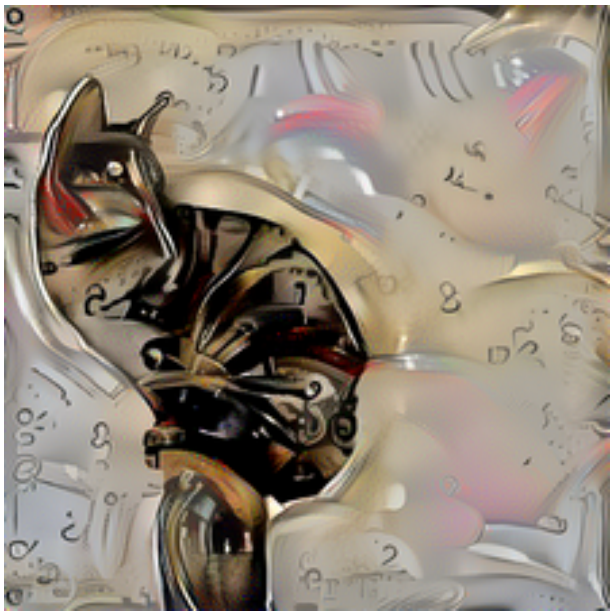


(h) StyleAug and crop (ours). As above in panel (g), with the additional step of RandomResizedCrop(scale=(0.25,1.0)).

Supplementary Figure 15: Example augmentations. "Orig" shows an example image after Inception-style preprocessing. "Aug 1" and "Aug 2" show augmentations applied to "Orig".

Cue-conflict image example

Example image of the cue conflict experiment dataset generated from [144] using style transfer. Each image has two "correct" labels, one relating to dominant shape of the image, and one relating to the dominant texture of an image. In the example image in Supp. Fig. 16, the shape of the object is a cat and the texture of the image is clocks.



Supplementary Figure 16: Example of a shape vs. texture cue-conflict image [144]. The shape of the object is a cat and the texture of the image is clocks. The displayed image was generated by Geirhos *et al.* [144].

References

[1] Cohen, M.R. and Maunsell, J.H. (2009). Attention improves performance primarily by reducing interneuronal correlations. Nature Neuroscience *12*, 1594–1600.

[2] Mitchell, J.F., Sundberg, K.A., and Reynolds, J.H. (2009). Spatial attention decorrelates intrinsic activity fluctuations in macaque area V4. Neuron *63*, 879–888.

[3] Snyder, A.C., Morais, M.J., and Smith, M.A. (2016). Dynamics of excitatory and inhibitory networks are differentially altered by selective attention. Journal of neurophysiology *116*, 1807–1820.

[4] Gu, Y., Liu, S., Fetsch, C.R., Yang, Y., Fok, S., Sunkara, A., DeAngelis, G.C., and Angelaki, D.E. (2011). Perceptual learning reduces interneuronal correlations in macaque visual cortex. Neuron *71*, 750–761.

[5] Erisken, S., Vaiceliunaite, A., Jurjut, O., Fiorini, M., Katzner, S., and Busse, L. (2014). Effects of locomotion extend throughout the mouse early visual system. Current Biology *24*, 2899–2907.

[6] Miura, K., Mainen, Z.F., and Uchida, N. (2012). Odor representations in olfactory cortex: distributed rate coding and decorrelated population activity. Neuron *74*, 1087–1098.

[7] Cohen, M.R. and Kohn, A. (2011). Measuring and interpreting neuronal correlations. Nature Neuroscience *14*, 811–819.

[8] Herrero, J.L., Gieselmann, M.A., Sanayei, M., and Thiele, A. (2013). Attention-induced variance and noise correlation reduction in macaque V1 is mediated by NMDA receptors. Neuron *78*, 729–739.

[9] Ruff, D.A. and Cohen, M.R. (2014). Global cognitive factors modulate correlated response variability between V4 neurons. Journal of Neuroscience *34*, 16408–16416.

[10] Gregoriou, G.G., Rossi, A.F., Ungerleider, L.G., and Desimone, R. (2014). Lesions of prefrontal cortex reduce attentional modulation of neuronal responses and synchrony in V4. Nature Neuroscience *17*, 1003–1011.

[11] Luo, T.Z. and Maunsell, J.H. (2015). Neuronal modulations in visual cortex are associated with only one of multiple components of attention. Neuron *86*, 1182–1188.

[12] Ruff, D.A. and Cohen, M.R. (2016). Attention increases spike count correlations between visual cortical areas. Journal of Neuroscience *36*, 7523–7534.

[13] Snyder, A.C., Yu, B.M., and Smith, M.A. (2018). Distinct population codes for attention in the absence and presence of visual stimulation. Nature Communications *9*, 4382.

[14] Ni, A.M., Ruff, D.A., Alberts, J.J., Symmonds, J., and Cohen, M.R. (2018). Learning and attention reveal a general relationship between population activity and behavior. Science *359*, 463–465.

[15] Maynard, E.M., Hatsopoulos, N.G., Ojakangas, C.L., Acuna, B.D., Sanes, J.N., Normann, R.A., and Donoghue, J.P. (1999). Neuronal interactions improve cortical population coding of movement direction. Journal of Neuroscience *19*, 8083–8093.

[16] Kohn, A. and Smith, M.A. (2005). Stimulus dependence of neuronal correlation in primary visual cortex of the macaque. Journal of Neuroscience *25*, 3661–3673.

[17] Smith, M.A. and Kohn, A. (2008). Spatial and temporal scales of neuronal correlation in primary visual cortex. The Journal of Neuroscience *28*, 12591–12603.

[18] Ponce-Alvarez, A., Thiele, A., Albright, T.D., Stoner, G.R., and Deco, G. (2013). Stimulus-dependent variability and noise correlations in cortical MT neurons. Proceedings of the National Academy of Sciences *110*, 13162–13167.

[19] Ruff, D.A. and Cohen, M.R. (2016). Stimulus dependence of correlated variability across cortical areas. Journal of Neuroscience *36*, 7546–7556.

[20] Nienborg, H., R. Cohen, M., and Cumming, B.G. (2012). Decision-related activity in sensory neurons: correlations among neurons and with behavior. Annual review of neuroscience *35*, 463–483.

[21] Bondy, A.G., Haefner, R.M., and Cumming, B.G. (2018). Feedback determines the structure of correlated variability in primary visual cortex. Nature Neuroscience *4*, 598–606.

[22] Ecker, A.S., Berens, P., Keliris, G.A., Bethge, M., Logothetis, N.K., and Tolias, A.S. (2010). Decorrelated neuronal firing in cortical microcircuits. science *327*, 584–587.

[23] Adibi, M., McDonald, J.S., Clifford, C.W., and Arabzadeh, E. (2013). Adaptation improves neural coding efficiency despite increasing correlations in variability. Journal of Neuroscience *33*, 2108–2120.

[24] Lee, D., Port, N.L., Kruse, W., and Georgopoulos, A.P. (1998). Variability and correlated noise in the discharge of neurons in motor and parietal areas of the primate cortex. Journal of Neuroscience *18*, 1161–1170.

[25] Smith, M.A. and Sommer, M.A. (2013). Spatial and temporal scales of neuronal correlation in visual area V4. Journal of Neuroscience *33*, 5422–5432.

[26] Ecker, A.S., Berens, P., Cotton, R.J., Subramaniyan, M., Denfield, G.H., Cadwell, C.R., Smirnakis, S.M., Bethge, M., and Tolias, A.S. (2014). State dependence of noise correlations in macaque primary visual cortex. Neuron *82*, 235–248.

[27] Solomon, S.S., Chen, S.C., Morley, J.W., and Solomon, S.G. (2015). Local and global correlations between neurons in the middle temporal area of primate visual cortex. Cerebral Cortex *25*, 3182–3196.

[28] Rosenbaum, R., Smith, M.A., Kohn, A., Rubin, J.E., and Doiron, B. (2017). The spatial structure of correlated neuronal variability. Nature Neuroscience *20*, 107.

[29] Khanna, S.B., Snyder, A.C., and Smith, M.A. (2019). Distinct sources of variability affect eye movement preparation. Journal of Neuroscience *39*, 4511–4526.

[30] Romo, R., Hernández, A., Zainos, A., and Salinas, E. (2003). Correlated neuronal discharges that increase coding efficiency during perceptual discrimination. Neuron *38*, 649–657.

[31] Huang, X. and Lisberger, S.G. (2009). Noise Correlations in Cortical Area MT and Their Potential Impact on Trial-by-Trial Variation in the Direction and Speed of Smooth-Pursuit Eye Movements. Journal of Neurophysiology *101*, 3012–3030.

[32] Bair, W., Zohary, E., and Newsome, W.T. (2001). Correlated firing in macaque visual area MT: time scales and relationship to behavior. Journal of Neuroscience *21*, 1676–1697.

[33] Qi, X.L. and Constantinidis, C. (2012). Correlated discharges in the primate prefrontal cortex before and after working memory training. The European Journal of Neuroscience *36*, 3538–3548.

[34] Zohary, E., Shadlen, M.N., and Newsome, W.T. (1994). Correlated neuronal discharge rate and its implications for psychophysical performance. Nature *370*, 140–143.

[35] Shadlen, M.N. and Newsome, W.T. (1998). The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. Journal of neuroscience *18*, 3870–3896.

[36] Abbott, L.F. and Dayan, P. (1999). The effect of correlated variability on the accuracy of a population code. Neural computation *11*, 91–101.

[37] Averbeck, B.B., Latham, P.E., and Pouget, A. (2006). Neural correlations, population coding and computation. Nature Reviews Neuroscience *7*, nrn1888.

[38] Moreno-Bote, R., Beck, J., Kanitscheider, I., Pitkow, X., Latham, P., and Pouget, A. (2014). Information-limiting correlations. Nature Neuroscience *17*, 1410–1417.

[39] Bartolo, R., Saunders, R.C., Mitz, A.R., and Averbeck, B.B. (2020). Information-Limiting Correlations in Large Neural Populations. Journal of Neuroscience *40*, 1668–1678.

[40] Rumyantsev, O.I., Lecoq, J.A., Hernandez, O., Zhang, Y., Savall, J., Chrapkiewicz, R., Li, J., Zeng, H., Ganguli, S., and Schnitzer, M.J. (2020). Fundamental bounds on the fidelity of sensory cortical coding. Nature *580*, 100–105.

[41] Cowley, B.R., Snyder, A.C., Acar, K., Williamson, R.C., Byron, M.Y., and Smith, M.A. (2020). Slow drift of neural activity as a signature of impulsivity in macaque visual and prefrontal cortex. bioRxiv .

[42] Hennig, J.A., Oby, E.R., Golub, M.D., Bahureksa, L.A., Sadtler, P.T., Quick, K.M., Ryu, S.I., Tyler-Kabara, E.C., Batista, A.P., Chase, S.M., et al. (2020). Learning is shaped by abrupt changes in neural engagement. bioRxiv , 2020.05.24.112714.

[43] Kingma, D.P. and Welling, M. (2014). Auto-Encoding Variational Bayes. arXiv:1312.6114 [cs, stat] ArXiv: 1312.6114.

[44] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research *15*, 1929–1958.

[45] Huang, G., Sun, Y., Liu, Z., Sedra, D., and Weinberger, K. (2016). Deep Networks with Stochastic Depth. arXiv:1603.09382 [cs] ArXiv: 1603.09382.

[46] Rabinowitz, N.C., Goris, R.L., Cohen, M., and Simoncelli, E.P. (2015). Attention stabilizes the shared gain of V4 populations. Elife *4*.

[47] Lin, I.C., Okun, M., Carandini, M., and Harris, K.D. (2015). The nature of shared cortical variability. Neuron *87*, 644–656.

[48] Ecker, A.S., Denfield, G.H., Bethge, M., and Tolias, A.S. (2016). On the structure of neuronal population activity under fluctuations in attentional state. Journal of Neuroscience *36*, 1775–1789.

[49] Huang, C., Ruff, D.A., Pyle, R., Rosenbaum, R., Cohen, M.R., and Doiron, B. (2019). Circuit models of low-dimensional shared variability in cortical networks. Neuron *101*, 337–348.

[50] Cunningham, J.P. and Yu, B.M. (2014). Dimensionality reduction for large-scale neural recordings. Nature Neuroscience *17*, 1500–1509.

[51] Harvey, C.D., Coen, P., and Tank, D.W. (2012). Choice-specific sequences in parietal cortex during a virtual-navigation decision task. Nature *484*, 62–68.

[52] Mante, V., Sussillo, D., Shenoy, K.V., and Newsome, W.T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature *503*, 78–84.

[53] Kiani, R., Cueva, C.J., Reppas, J.B., and Newsome, W.T. (2014). Dynamics of neural population responses in prefrontal cortex indicate changes of mind on single trials. Current Biology *24*, 1542–1547.

[54] Kaufman, M.T., Churchland, M.M., Ryu, S.I., and Shenoy, K.V. (2015). Vacillation, indecision and hesitation in moment-by-moment decoding of monkey motor cortex. Elife *4*, e04677.

[55] Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Foster, J.D., Nuyujukian, P., Ryu, S.I., and Shenoy, K.V. (2012). Neural population dynamics during reaching. Nature *487*, 51–6.

[56] Gallego, J.A., Perich, M.G., Miller, L.E., and Solla, S.A. (2017). Neural manifolds for the control of movement. Neuron *94*, 978–984.

[57] Sadtler, P.T., Quick, K.M., Golub, M.D., Chase, S.M., Ryu, S.I., Tyler-Kabara, E.C., Yu, B.M., and Batista, A.P. (2014). Neural constraints on learning. Nature *512*, 423.

[58] Vyas, S., Even-Chen, N., Stavisky, S.D., Ryu, S.I., Nuyujukian, P., and Shenoy, K.V. (2018). Neural Population Dynamics Underlying Motor Learning Transfer. Neuron *97*, 1177–1186.e3.

[59] Mazor, O. and Laurent, G. (2005). Transient dynamics versus fixed points in odor representations by locust antennal lobe projection neurons. Neuron *48*, 661–673.

[60] Pang, R., Lansdell, B.J., and Fairhall, A.L. (2016). Dimensionality reduction in neuroscience. Current Biology *26*, R656–R660.

[61] Cohen, M.R. and Maunsell, J.H. (2010). A neuronal population measure of attention predicts behavioral performance on individual trials. J Neurosci *30*, 15241–15253.

[62] Perich, M.G., Gallego, J.A., and Miller, L.E. (2018). A Neural Population Mechanism for Rapid Learning. Neuron *100*, 964–976.e7.

[63] Ames, K.C. and Churchland, M.M. (2019). Motor cortex signals for each arm are mixed across hemispheres and neurons yet partitioned within the population response. eLife *8*, e46159.

[64] Semedo, J.D., Zandvakili, A., Machens, C.K., Yu, B.M., and Kohn, A. (2019). Cortical areas interact through a communication subspace. Neuron .

[65] Veuthey, T.L., Derosier, K., Kondapavulur, S., and Ganguly, K. (2020). Single-trial cross-area neural population dynamics during long-term skill learning. Nature Communications *11*, 4057. Number: 1 Publisher: Nature Publishing Group.

[66] Williamson, R.C., Cowley, B.R., Litwin-Kumar, A., Doiron, B., Kohn, A., Smith, M.A., and Yu, B.M. (2016). Scaling properties of dimensionality reduction for neural populations and network models. PLoS computational biology *12*, e1005141.

[67] Mazzucato, L., Fontanini, A., and La Camera, G. (2016). Stimuli reduce the dimensionality of cortical activity. Frontiers in systems neuroscience *10*, 11.

[68] Recanatesi, S., Ocker, G.K., Buice, M.A., and Shea-Brown, E. (2019). Dimensionality in recurrent spiking networks: Global trends in activity and local origins in connectivity. PLOS Computational Biology *15*, e1006446.

[69] Churchland, M.M., Yu, B.M., Cunningham, J.P., Sugrue, L.P., Cohen, M.R., Corrado, G.S., Newsome, W.T., Clark, A.M., Hosseini, P., Scott, B.B., et al. (2010). Stimulus onset quenches neural variability: a widespread cortical phenomenon. Nature Neuroscience *13*, 369–378.

[70] Cowley, B.R., Smith, M.A., Kohn, A., and Yu, B.M. (2016). Stimulus-Driven Population Activity Patterns in Macaque Primary Visual Cortex. PLOS Computational Biology *12*, e1005185.

[71] Gallego, J.A., Perich, M.G., Naufel, S.N., Ethier, C., Solla, S.A., and Miller, L.E. (2018). Cortical population activity within a preserved neural manifold underlies multiple motor behaviors. Nature Communications *9*, 4233. Number: 1 Publisher: Nature Publishing Group.

[72] Yu, B.M., Cunningham, J.P., Santhanam, G., Ryu, S.I., Shenoy, K.V., and Sahani, M. (2009). Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. Journal of Neurophysiology *102*, 614–635.

[73] Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., and Harris, K.D. (2019). High-dimensional geometry of population responses in visual cortex. Nature .

[74] Stringer, C., Pachitariu, M., Steinmetz, N., Reddy, C.B., Carandini, M., and Harris, K.D. (2019). Spontaneous Behaviors Drive Multidimensional, Brain-wide Activity. Science .

[75] Musall, S., Kaufman, M.T., Juavinett, A.L., Gluf, S., and Churchland, A.K. (2019). Single-trial neural dynamics are dominated by richly varied movements. Nature neuroscience *22*, 1677–1686.

[76] Ruff, D.A., Xue, C., Kramer, L.E., Baqai, F., and Cohen, M.R. (2019). Low rank mechanisms underlying flexible visual representations. bioRxiv , 730978Publisher: Cold Spring Harbor Laboratory Section: New Results.

[77] Snyder, A.C., Yu, B.M., and Smith, M.A. (2018). Distinct population codes for attention in the absence and presence of visual stimulation. Nature Communications *9*, 1–14.

[78] Zandvakili, A. and Kohn, A. (2015). Coordinated neuronal activity enhances corticocortical communication. Neuron *87*, 827–839.

[79] Nevet, A., Morris, G., Saban, G., Arkadir, D., and Bergman, H. (2007). Lack of spike-count and spike-time correlations in the substantia nigra reticulata despite overlap of neural responses. Journal of neurophysiology *98*, 2232–2243.

[80] Liu, S., Gu, Y., DeAngelis, G.C., and Angelaki, D.E. (2013). Choice-related activity and correlated noise in subcortical vestibular neurons. Nature Neuroscience *16*, 89.

[81] Averbeck, B.B., Latham, P.E., and Pouget, A. (2006). Neural correlations, population coding and computation. Nature Reviews Neuroscience *7*, 358–366.

[82] Moreno-Bote, R., Beck, J., Kanitscheider, I., Pitkow, X., Latham, P., and Pouget, A. (2014). Information-limiting correlations. Nature Neuroscience *17*, 1410.

[83] Rigotti, M., Barak, O., Warden, M.R., Wang, X.J., Daw, N.D., Miller, E.K., and Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. Nature *497*, 585–590.

[84] Kohn, A., Coen-Cagli, R., Kanitscheider, I., and Pouget, A. (2016). Correlations and neuronal population information. Annual Review of Neuroscience *39*, 237–256.

[85] Rumyantsev, O.I., Lecoq, J.A., Hernandez, O., Zhang, Y., Savall, J., Chrapkiewicz, R., Li, J., Zeng, H., Ganguli, S., and Schnitzer, M.J. (2020). Fundamental bounds on the fidelity of sensory cortical coding. Nature *580*, 100–105.

[86] Bartolo, R., Saunders, R.C., Mitz, A.R., and Averbeck, B.B. (2020). Information-Limiting Correlations in Large Neural Populations. Journal of Neuroscience *40*, 1668–1678.

[87] Kobak, D., Brendel, W., Constantinidis, C., Feierstein, C.E., Kepecs, A., Mainen, Z.F., Qi, X.L., Romo, R., Uchida, N., and Machens, C.K. (2016). Demixed principal component analysis of neural population data. eLife *5*, e10989.

[88] Santhanam, G., Yu, B.M., Gilja, V., Ryu, S.I., Afshar, A., Sahani, M., and Shenoy, K.V. (2009). Factor-analysis methods for higher-performance neural prostheses. Journal of neurophysiology *102*, 1315–1330.

[89] Bittner, S.R., Williamson, R.C., Snyder, A.C., Litwin-Kumar, A., Doiron, B., Chase, S.M., Smith, M.A., and Yu, B.M. (2017). Population activity structure of excitatory and inhibitory neurons. PloS one *12*, e0181773.

[90] Okun, M., Steinmetz, N.A., Cossell, L., Iacaruso, M.F., Ko, H., Barthó, P., Moore, T., Hofer, S.B., Mrsic-Flogel, T.D., Carandini, M., et al. (2015). Diverse coupling of neurons to populations in sensory cortex. Nature *521*, 511–515.

[91] Insanally, M.N., Carcea, I., Field, R.E., Rodgers, C.C., DePasquale, B., Rajan, K., DeWeese, M.R., Albanna, B.F., and Froemke, R.C. (2019). Spike-timing-dependent ensemble encoding by non-classically responsive cortical neurons. eLife *8*.

[92] Harris, K.D. and Thiele, A. (2011). Cortical state and attention. Nature reviews neuroscience *12*, 509.

[93] Minces, V., Pinto, L., Dan, Y., and Chiba, A.A. (2017). Cholinergic shaping of neural correlations. Proceedings of the National Academy of Sciences of the United States of America *114*, 5725–5730.

[94] Ahrens, M.B., Li, J.M., Orger, M.B., Robson, D.N., Schier, A.F., Engert, F., and Portugues, R. (2012). Brain-wide neuronal dynamics during motor adaptation in zebrafish. Nature *485*, 471–477.

[95] Ahrens, M.B., Orger, M.B., Robson, D.N., Li, J.M., and Keller, P.J. (2013). Whole-brain functional imaging at cellular resolution using light-sheet microscopy. Nature methods *10*, 413.

[96] Jun, J.J., Steinmetz, N.A., Siegle, J.H., Denman, D.J., Bauza, M., Barbarits, B., Lee, A.K., Anastassiou, C.A., Andrei, A., Aydın, *c*C., et al. (2017). Fully integrated silicon probes for high-density recording of neural activity. Nature *551*, 232.

[97] Semedo, J., Zandvakili, A., Kohn, A., Machens, C.K., and Yu, B.M. (2014). Extracting Latent Structure From Multiple Interacting Neural Populations. In Advances in Neural Information Processing Systems. pp. 2942–2950.

[98] Kelly, R.C., Smith, M.A., Samonds, J.M., Kohn, A., Bonds, A., Movshon, J.A., and Lee, T.S. (2007). Comparison of recordings from microelectrode arrays and single electrodes in the visual cortex. Journal of Neuroscience *27*, 261–264.

[99] Santhanam, G., Yu, B.M., Gilja, V., Ryu, S.I., Afshar, A., Sahani, M., and Shenoy, K.V. (2009). Factor-Analysis Methods for Higher-Performance Neural Prostheses. Journal of Neurophysiology *102*, 1315–1330.

[100] Koralek, A.C., Jin, X., Long, J.D., Costa, R.M., and Carmena, J.M. (2012). CORTICOSTRIATAL PLASTICITY IS NECESSARY FOR LEARNING INTENTIONAL NEUROPROSTHETIC SKILLS. Nature *483*, 331–335.

[101] Neely, R.M., Koralek, A.C., Athalye, V.R., Costa, R.M., and Carmena, J.M. (2018). Volitional Modulation of Primary Visual Cortex Activity Requires the Basal Ganglia. Neuron *97*, 1356–1368.e4.

[102] Williamson, R.C., Cowley, B.R., Litwin-Kumar, A., Doiron, B., Kohn, A., Smith, M.A., and Yu, B.M. (2016). Scaling properties of dimensionality reduction for neural populations and network models. PLoS Computational Biology *12*, e1005141.

[103] Bittner, S.R., Williamson, R.C., Snyder, A.C., Litwin-Kumar, A., Doiron, B., Chase, S.M., Smith, M.A., and Yu, B.M. (2017). Population activity structure of excitatory and inhibitory neurons. PLOS ONE *12*, e0181773.

[104] Umakantha, A., Morina, R., Cowley, B.R., Snyder, A.C., Smith, M.A., and Yu, B.M. (2021). Bridging neuronal correlations and dimensionality reduction. Neuron *109*, 2740–2754.e12.

[105] Issar, D., Williamson, R.C., Khanna, S.B., and Smith, M.A. (2020). A neural network for online spike classification that improves decoding accuracy. Journal of Neurophysiology *123*, 1472–1485.

[106] Shadlen, M.N. and Newsome, W.T. (1998). The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience *18*, 3870–3896.

[107] Churchland, M.M., Afshar, A., and Shenoy, K.V. (2006). A Central Source of Movement Variability. Neuron *52*, 1085–1096.

[108] Druckmann, S. and Chklovskii, D. (2012). Neuronal Circuits Underlying Persistent Representations Despite Time Varying Activity. Current Biology *22*, 2095–2103.

[109] Churchland, A.K., Kiani, R., Chaudhuri, R., Wang, X.J., Pouget, A., and Shadlen, M.N. (2011). Variance as a Signature of Neural Computations during Decision Making. Neuron *69*, 818–831.

[110] Kiani, R., Cueva, C., Reppas, J., and Newsome, W. (2014). Dynamics of Neural Population Responses in Prefrontal Cortex Indicate Changes of Mind on Single Trials. Current Biology *24*, 1542–1547.

[111] Churchland, M.M., Yu, B.M., Cunningham, J.P., Sugrue, L.P., Cohen, M.R., Corrado, G.S., Newsome, W.T., Clark, A.M., Hosseini, P., Scott, B.B., et al. (2010). Stimulus onset quenches neural variability: a widespread cortical phenomenon. Nature Neuroscience *13*, 369–378.

[112] Wang, C.A., Baird, T., Huang, J., Coutinho, J.D., Brien, D.C., and Munoz, D.P. (2018). Arousal Effects on Pupil Size, Heart Rate, and Skin Conductance in an Emotional Face Task. Frontiers in Neurology *9*.

[113] Stringer, C., Pachitariu, M., Steinmetz, N., Reddy, C.B., Carandini, M., and Harris, K.D. (2019). Spontaneous behaviors drive multidimensional, brainwide activity. Science *364*.

[114] Bondy, A.G., Haefner, R.M., and Cumming, B.G. (2018). Feedback determines the structure of correlated variability in primary visual cortex. Nature Neuroscience *21*, 598–606.

[115] Cohen, J.Y., Crowder, E.A., Heitz, R.P., Subraveti, C.R., Thompson, K.G., Woodman, G.F., and Schall, J.D. (2010). Cooperation and Competition among Frontal Eye Field Neurons during Visual Target Selection. Journal of Neuroscience *30*, 3227–3238.

[116] Khanna, S.B., Snyder, A.C., and Smith, M.A. (2019). Distinct sources of variability affect eye movement preparation. Journal of Neuroscience .

[117] Smith, M.A. and Kohn, A. (2008). Spatial and Temporal Scales of Neuronal Correlation in Primary Visual Cortex. The Journal of Neuroscience *28*, 12591–12603.

[118] Ahrens, M.B., Orger, M.B., Robson, D.N., Li, J.M., and Keller, P.J. (2013). Whole-brain functional imaging at cellular resolution using light-sheet microscopy. Nature Methods *10*, 413–420.

[119] Mu, Y., Bennett, D.V., Rubinov, M., Narayan, S., Yang, C.T., Tanimoto, M., Mensh, B.D., Looger, L.L., and Ahrens, M.B. (2019). Glia Accumulate Evidence that Actions Are Futile and Suppress Unsuccessful Behavior. Cell *178*, 27–43.e19.

[120] Fries, P. (2015). Rhythms For Cognition: Communication Through Coherence. Neuron *88*, 220–235.

[121] van Kempen, J., Loughnane, G.M., Newman, D.P., Kelly, S.P., Thiele, A., O'Connell, R.G., and Bellgrove, M.A. (2019). Behavioural and neural signatures of perceptual decision-making are modulated by pupil-linked arousal. eLife *8*, e42541.

[122] Semedo, J.D., Zandvakili, A., Machens, C.K., Yu, B.M., and Kohn, A. (2019). Cortical Areas Interact through a Communication Subspace. Neuron *102*, 249–259.e4.

[123] Li, N., Daie, K., Svoboda, K., and Druckmann, S. (2016). Robust neuronal dynamics in premotor cortex during motor planning. Nature *532*, 459–464.

[124] Mayo, J.P., Cohen, M.R., and Maunsell, J.H.R. (2015). A Refined Neuronal Population Measure of Visual Attention. PLOS ONE *10*, e0136570.

[125] Vinci, G., Ventura, V., Smith, M.A., and Kass, R.E. (2018). Adjusted regularization in latent graphical models: Application to multiple-neuron spike count data. Annals of Applied Statistics *12*, 1068–1095. Publisher: Institute of Mathematical Statistics.

[126] Semedo, J.D., Gokcen, E., Machens, C.K., Kohn, A., and Yu, B.M. (2020). Statistical methods for dissecting interactions between brain areas. Current Opinion in Neurobiology *65*, 59–69.

[127] Bong, H., Liu, Z., Ren, Z., Smith, M., Ventura, V., and Robert, K.E. (2020). Latent Dynamic Factor Analysis of High-Dimensional Neural Recordings. Advances in neural information processing systems *33*.

[128] Smith, M.A. and Sommer, M.A. (2013). Spatial and Temporal Scales of Neuronal Correlation in Visual Area V4. The Journal of Neuroscience *33*, 5422–5432.

[129] Wainwright, M.J. (2019). High-Dimensional Statistics: A Non-Asymptotic Viewpoint (Cambridge University Press).

[130] Allen, W.E., Chen, M.Z., Pichamoorthy, N., Tien, R.H., Pachitariu, M., Luo, L., and Deisseroth, K. (2019). Thirst regulates motivated behavior through modulation of brainwide neural population dynamics. Science *364*.

[131] Mazzucato, L., Fontanini, A., and La Camera, G. (2015). Stimuli reduce the dimensionality of cortical activity. arXiv:1509.03621 [q-bio] ArXiv: 1509.03621.

[132] Hikosaka, O. and Wurtz, R.H. (1983). Visual and oculomotor functions of monkey substantia nigra pars reticulata. I. Relation of visual and auditory responses to saccades. Journal of Neurophysiology *49*, 1230–1253.

[133] Harris, K.D. (2021). Nonsense correlations in neuroscience. Technical report. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

[134] Georgopoulos, A.P., Kalaska, J.F., Caminiti, R., and Massey, J.T. (1982). On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. Journal of Neuroscience *2*, 1527–1537.

[135] Bach, F.R. and Jordan, M.I. (2006). A Probabilistic Interpretation of Canonical Correlation Analysis , 11.

[136] Everitt, B.S. (1984). Factor analysis. In An Introduction to Latent Variable Models, B.S. Everitt, ed., Monographs on Statistics and Applied Probability (Dordrecht: Springer Netherlands), pp. 13–31.

[137] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological) *39*, 1–38. Publisher: [Royal Statistical Society, Wiley].

[138] Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Advances in Neural Information Processing Systems, volume 25 (Curran Associates, Inc.).

[139] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. arXiv:1506.02640 [cs] ArXiv: 1506.02640.

[140] He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2018). Mask R-CNN. arXiv:1703.06870 [cs] ArXiv: 1703.06870.

[141] Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. (2020). Big Transfer (BiT): General Visual Representation Learning. arXiv:1912.11370 [cs] ArXiv: 1912.11370.

[142] Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning (MIT Press). `http://www.deeplearningbook.org`.

[143] Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. (2020). AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. arXiv:1912.02781 [cs, stat] ArXiv: 1912.02781.

[144] Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., and Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.

[145] Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences *111*, 8619–8624. Publisher: National Academy of Sciences Section: Biological Sciences.

[146] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. arXiv:1706.03762 [cs] ArXiv: 1706.03762.

[147] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs] ArXiv: 2010.11929.

[148] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. arXiv:2012.12877 [cs] ArXiv: 2012.12877.

[149] Shao, R., Shi, Z., Yi, J., Chen, P.Y., and Hsieh, C.J. (2021). On the Adversarial Robustness of Visual Transformers. arXiv:2103.15670 [cs] ArXiv: 2103.15670.

[150] Naseer, M., Ranasinghe, K., Khan, S., Hayat, M., Khan, F.S., and Yang, M.H. (2021). Intriguing Properties of Vision Transformers. arXiv:2105.10497 [cs] ArXiv: 2105.10497.

[151] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging Properties in Self-Supervised Vision Transformers. arXiv:2104.14294 [cs] ArXiv: 2104.14294.

[152] Tuli, S., Dasgupta, I., Grant, E., and Griffiths, T.L. (2021). Are Convolutional Neural Networks or Transformers more like human vision? arXiv:2105.07197 [cs] ArXiv: 2105.07197.

[153] Chen, X., Hsieh, C.J., and Gong, B. (2021). When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations. arXiv:2106.01548 [cs] ArXiv: 2106.01548.

[154] Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., and Beyer, L. (2021). How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. arXiv:2106.10270 [cs] ArXiv: 2106.10270.

[155] Cubuk, E.D., Zoph, B., Shlens, J., and Le, Q.V. (2019). RandAugment: Practical automated data augmentation with a reduced search space. arXiv:1909.13719 [cs] ArXiv: 1909.13719.

[156] He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum Contrast for Unsupervised Visual Representation Learning. arXiv:1911.05722 [cs] ArXiv: 1911.05722.

[157] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations. arXiv:2002.05709 [cs, stat] ArXiv: 2002.05709.

[158] Xu, Z., Liu, D., Yang, J., Raffel, C., and Niethammer, M. (2021). Robust and Generalizable Visual Representation Learning via Random Convolutions. arXiv:2007.13003 [cs] ArXiv: 2007.13003.

[159] Deza, A. and Konkle, T. (2021). Emergent Properties of Foveated Perceptual Systems. arXiv:2006.07991 [cs, eess, q-bio] ArXiv: 2006.07991.

[160] Freeman, J. and Simoncelli, E.P. (2011). Metamers of the ventral stream. Nature Neuroscience *14*, 1195–1201.

[161] Huang, X. and Belongie, S. (2017). Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. arXiv:1703.06868 [cs] ArXiv: 1703.06868.

[162] Hendrycks, D. and Dietterich, T. (2019). Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. arXiv:1903.12261 [cs, stat] ArXiv: 1903.12261.

[163] Parkhi, O.M., Vedaldi, A., Zisserman, A., and Jawahar, C.V. (2012). Cats and dogs. In 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3498–3505. ISSN: 1063-6919.

[164] Cheng, G., Han, J., and Lu, X. (2017). Remote Sensing Image Scene Classification: Benchmark and State of the Art. Proceedings of the IEEE *105*, 1865–1883. ArXiv: 1703.00121 version: 1.

[165] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Networks. arXiv:1406.2661 [cs, stat] ArXiv: 1406.2661.

[166] Cohen, M.R. and Maunsell, J.H. (2011). Using neuronal populations to study the mechanisms underlying spatial and feature attention. Neuron *70*, 1192–1204.

[167] Golub, M.D., Chase, S.M., Batista, A.P., and Yu, B.M. (2016). Brain–computer interfaces for dissecting cognitive processes underlying sensorimotor control. Current Opinion in Neurobiology *37*, 53–58.

[168] Shorten, C. and Khoshgoftaar, T.M. (2019). A survey on Image Data Augmentation for Deep Learning. Journal of Big Data *6*, 60.

[169] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language Models are Few-Shot Learners. arXiv:2005.14165 [cs] ArXiv: 2005.14165.

[170] Devlin, J., Chang, M.W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs] ArXiv: 1810.04805.

[171] Bishop, W., Chestek, C.C., Gilja, V., Nuyujukian, P., Foster, J.D., Ryu, S.I., Shenoy, K.V., and Yu, B.M. (2014). Self-recalibrating classifiers for intracortical brain–computer interfaces *11*, 026001. Publisher: IOP Publishing.

[172] Degenhart, A.D., Bishop, W.E., Oby, E.R., Tyler-Kabara, E.C., Chase, S.M., Batista, A.P., and Yu, B.M. (2020). Stabilization of a brain–computer interface via the alignment of low-dimensional spaces of neural activity. Nature Biomedical Engineering *4*, 672–685. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Brain–machine interface;Motor cortex Subject_term_id: brain-machine-interface;motor-cortex.

[173] Xu, H., Ma, Y., Liu, H., Deb, D., Liu, H., Tang, J., and Jain, A.K. (2019). Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. arXiv:1909.08072 [cs, stat] ArXiv: 1909.08072.

[174] Hennig, J.A., Oby, E.R., Losey, D.M., Batista, A.P., Yu, B.M., and Chase, S.M. (2021). How learning unfolds in the brain: toward an optimization view. Neuron , S0896627321006772.

[175] Wightman, R. (2019). PyTorch Image Models. `https://github.com/rwightman/pytorch-image-models`.

[176] Deza, A., Jonnalagadda, A., and Eckstein, M. (2018). Towards Metamerism via Foveated Style Transfer. arXiv:1705.10041 [cs] ArXiv: 1705.10041.