Comparing high-dimensional neural recordings by aligning their low-dimensional latent representations

Max Dabagia, Konrad P Kording, and Eva L. Dyer

Abstract

As measurements in neuroscience have shifted from single neurons to large populations, analysis and modeling of neural activity has accordingly moved towards building models that describe how collections of neurons work together to build percepts. As we generate more data from more individuals, across different time points and through learning, we are faced with the dilemma of how we should go about comparing the activities of populations of neurons over all these differing conditions. A critical observation is that when the activity of many of neurons can be modeled as being driven by a smaller number of latent factors, then distinct neural recordings acquired from neurons that share the similar underlying latent space can be compared by finding and aligning their latent factors. In this article, we provide an overview of how latent space models can be generated, aligned, and compared. Being able to align neural representations promises meaningful ways of comparing high-dimensional neural activities across times, subsets of neurons, or individuals.

Introduction

The study of brain function and cognition relies on the measurement and interpretation of changes to the brain’s activity. Whether between states of disease and health [1], across novice- and expert-level performance on a complex task [2], or in wakefulness versus sleep [3], we need ways to compare neural activity patterns and draw conclusions about how they differ.

Historically, changes in neural activity patterns have been examined at the level of single cells, by characterizing how the firing rate (or tuning curves) of each cell shifts across conditions [4, 5, 6]. However, we know that the brain is a highly interconnected network of neurons [7], and thus the impact of different disease states or conditions is likely to have complex effects.

*MD is in the School of Computer Science at the Georgia Institute of Technology.
†KPK is in the Depts of Bioengineering and Neuroscience at the University of Pennsylvania.
‡ELD is in the Coulter Department of Biomedical Engineering at the Georgia Institute of Technology.
Corresponding author: evadyer@gatech.edu
across an entire circuit. Without modeling these higher-order interactions between neurons, it may be impossible to model the collective impact of a given perturbation.

Although improving technology allows simultaneous access to an ever-increasing number of neurons \([8, 9, 10]\), each recording is still restricted to a fraction of all the neurons in a circuit of interest. In most preparations, maintaining contact with specific neurons across two recording sessions is challenging or impossible, especially when studying brain activity from two individuals. At first glance, it is tempting to try to pair off similar neurons across recordings and find “correspondence” between specific neurons in each dataset. However, it is unclear whether this is a fruitful approach in most cases (although it has a remarkably successful history in simpler nervous systems, such as those of insects), or whether we should even expect to find correspondences at the level of individual units.

The alternative is to summarize each recording in a way that is invariant to the specific subset of neurons that are polled. Many existing approaches build invariance through the computation of metrics that capture essential properties of the pairwise interactions between neurons (like clustering or density of connections) \([11, 12, 13]\), or extract a global property of the network such as criticality \([14, 15]\). While these approaches can be informative when we know the right metrics to compare across conditions, having a data-driven approach to identify couplings across multiple units would provide a rich multi-dimensional picture of changes in neural circuits.

One potential approach is to first find a simpler representation of the data, or a factorization, and then find correspondence between the factors of this model (rather than individual neurons). The hope is that once in this latent space, datasets may agree with one another and comparisons may be more informative. Fortunately, neural activity can be approximated by such latent representations in many models and systems \([16, 17, 18]\). A substantial body of evidence now points to the fact that latent factors are stable across subsets of neurons, across time, and even individuals \([19, 20, 21, 22, 23, 24, 25]\), suggesting they may be a powerful tool for comparison.

In this article, we discuss strategies for comparing high-dimensional neural activity that leverage low-dimensional latent spaces present across multiple recordings. We then present the challenges that need to be overcome to compare recordings at larger scales and across more severe perturbations, and conclude with an overview of the prospective analyses and discoveries these approaches will make possible.

**Latent state space models for neural populations**

To build a population-level view of neural activity, we can start by thinking of the firing rates of all measured neurons as a point in a multi-dimensional state space, where each axis rep-
resents the firing rate of one of the recorded neurons (Figure 1A) [26]. Viewed this way, the
dimension of this space is the same as the number of neurons sampled in the recording. If
every neuron operated completely independently this would be the end of the story, and more-
over neuroscience would be in trouble, because understanding a neural circuit would require
the characterization of every single neuron involved in it. Fortunately, there is some evidence
that this is not the case. For one thing, neurons interact with each other through both excita-
tion and inhibition, which naturally leads to correlation in their activity [27, 28]. When many
neurons are correlated with each other, it is reasonable to expect that their activities might be
explainable by a smaller number of variables.

To discover these low-dimensional factors that may drive neural activity, dimensionality re-
duction methods can be used to convert observations of a large number of observed neurons
into a smaller number of latent variables. This identifies redundancies (such as correlation)
across many measured neural traces in a dataset. The simplest methods identify each latent
variable with a particular pattern of neural activity (Figure 1B), such that that any observation
can be approximated as a linear combination of these factors [29, 2, 30]. Principal Compo-
nents Analysis (PCA) is probably the most widely-used technique [21, 31, 32], and chooses
latent variables which retain the most variance of the observations possible. There are many
extensions of this approach that either relax PCA’s assumptions (e.g. the latent variables must
be orthogonal) [33], or incorporate additional structure into the solution (e.g., adding a non-
negativity constraint) [34, 35]. To leverage temporal information, a variety of methods model
the latent factors as a dynamical system and find mappings which preserve coherent dynam-
ics [36, 37, 19, 38]. More sophisticated manifold learning strategies (1C) can identify low-
dimensional structures embedded in non-linear ways in neural activity, and “flatten” them to
recover the latent factors [39, 40].

**The need for alignment and stitching**

An important feature of latent representations is that they summarize neural recordings in a
way that (in theory) is sensitive to only the factors driving the activity of a population, and
invariant to which particular neurons are tracked. However, in practice, both the dimension-
ality reduction technique used to generate the latent representation and the reality of having
access to only a fraction of all the neurons in a population can transform latent representations,
making it impossible to compare them directly without corrective action.

**Distribution alignment.** To tackle the issue of finding correspondence between two latent
spaces, one can pose this problem as finding a mapping between two datasets through dis-
btribution alignment (DA) (Figure 2A). In this setting, we define a class of transformations to
Figure 1: *Low-dimensional and dynamical latent space models.* In A, we depict two multi-unit neural recordings collected in different behavioral states (Left and Right). Below, a visualization of the firing rates of three units, where each axis corresponds to the firing rate of a different neuron. When we record from the same unit across different time points, the changes in firing rate can be observed in the collective firing rates. In B, we depict the formation of a latent space model which takes high-dimensional neural recordings and maps them into a 2D latent space. In C, from left to right, we show a linear subspace, nonlinear manifold, and union of subspaces.

explore and a divergence measure, and then look for a transformation that minimizes the divergence between the transformed source and target, effectively aligning them. A few different approaches for DA have been introduced, tailored to the neural decoding setting. The first, Distribution Alignment Decoding (DAD) [22], uses density estimation to infer the probability distributions in the latent space, and then, owing to the non-convexity of Kullback-Leibler divergence, employs a brute-force search to identify the optimal rotation which matches the two distributions. The second, Hierarchical Wasserstein Alignment (HiWA) [41] improves upon
Figure 2: Overview of different alignment methods for neural activity. In A, a distribution alignment technique (HiWA\cite{41}) is used to align one latent representation (source) to another structurally-similar one (target). The bar graph (lower right) shows improved cluster prediction error (NNE) based on the nearest neighbors of each source point within the labelled target, as well as the Wasserstein distance (WD)\cite{42} between the two distributions, for Monkey M and C. In B, the covariance inference approach is demonstrated; the objective is to complete the covariance matrix of the aggregated population of several recordings with neuronal or temporal overlap, which then allows construction of a common, covariance-based latent embedding. In C, coupled latent representations $f(x_1)$ and $g(x_2)$ are learned so that the resulting space is shared across the two recordings, which is the strategy of learning coupled mappings.
this strategy by leveraging the tendency of neural circuits to constrain their low-dimensional activity to clusters or low-dimensional subspaces. In conjunction with a more geometric formulation of distance drawn from optimal transport, HiWA is able to more quickly and robustly recover the correct rotation.

Recently, the potential of an online alignment strategy to enhance the stability of a brain-computer interface (BCI) controller has been demonstrated [43]. This method assumes a sufficiently substantial fraction of neurons are stable so that only a few coordinate axes of the neural manifold need to be adjusted. Although less general than distribution alignment because it assumes the underlying transformation to be a permutation of a few of the coordinate axes, the decrease in computational requirements makes this solution more practical in a realistic online BCI setting. However, all of these nascent distribution alignment approaches are only capable of dealing with the unitary transformations between distributions. In the case that the low-dimensional dynamics of neural circuits are intrinsically different, more sophisticated and nonlinear approaches will be required for alignment.

A deep learning-driven solution to this problem [44] draws inspiration from generative adversarial networks (GANs), which have received some attention in the computational neuroscience community both for generative tasks like producing realistic neural activity [45, 46], or to aid in downstream tasks like decoding visual stimuli from fMRI scans [47]. In essence, the key insight behind GANs is that a network can be indirectly trained to sample from a distribution by training it to fool an observer into believing that the samples it generates came from the target distribution. To use this idea for distribution alignment, the generator just learns a transformation of neural activity which aligns it to the distribution of the training set. This approach is promising for brain-machine interfaces, where the priority is just correctly decoding target variables and less about interpreting the embedding. On the other hand, when using latent representations to study a neural circuit, it might be desirable for distinctive new features of the shifted distribution to be preserved, but a sufficiently powerful alignment network which has “overfit” to the target dataset will attenuate them in favor of matching the original dataset as closely as possible.

**Covariance inference.** An alternative approach for combining recordings is to treat the situation as a missing data problem; that is, the combined pool of all of the neurons from the recordings are assumed to come from the same population, and the challenge is to infer their covariance when only a smaller fraction of its neurons can be simultaneously observed [48, 12] (Figure 2B). Normally, finding the population’s latent space would only be possible when recordings made at different points in time contain overlapping groups of neurons, specifically when the size of the overlap is at least the dimensionality of the latent space. However, by leveraging the dynamics of the population, or the amount of information gained about the instantaneous firing rate of one neuron by observing the history of another, the covariance ma-
trix can be inferred even with no shared neurons across recordings, as long as the recordings are performed before the dynamics change too significantly [11]. These requirements limit this approach’s usefulness, especially in the acute (single-session implantation) paradigm where no neurons can be reliably matched across recordings and recordings are usually not made serially. Nevertheless, in many datasets (imaging in particular) such recording structure can be leveraged to recover the latent representations, and so this method is a useful counterpart to distribution alignment approaches.

**Learning independent mappings into a common latent space.** The final strategy is to join the process of dimensionality reduction across recordings such that the resulting latent representations are all embedded in the same space (Figure 2C). A straightforward approach for doing so that has withstood the test of time is canonical correlation analysis (CCA) [49, 50, 51, 25], and more sophisticated nonlinear extensions [52, 53, 54]. This technique is actually a generalization of principal component analysis, where rather than choosing a single component at a time which maximizes variance, it chooses one component for each dataset so that the covariance along these components is maximized. CCA reduces to PCA when the datasets being compared are the same. The problem is that if the recordings were not made simultaneously (or aligned to a stimulus), the correspondence of individual moments in time across the datasets is essentially arbitrary, and so the notion of covariance is poorly defined.

A recent method for learning nonlinear dynamical systems (LFADS) “stitches” together data from many recording sessions by learning a separate input and output mapping for each session along with a shared latent space for all sessions jointly [55]. This method develops the mappings between each recording session (dataset) and the latent space by first finding the trial-averaged responses for each unit and then learning a weight matrix (input map) to match these trial-averaged responses. This means that given training data consisting of the direction (or target) of the reach, all of the trajectories in the same direction can be utilized to improve the quality of the latent representation, which alleviates the need to maintain contact with the same neurons across recordings. The result is that task information can be leveraged to stitch together dynamics from different populations of neurons. In this case, a common latent space model is learned across all of the measured units over all days and used to decode on any one of the days.

**Comparison of various methods.** Given how challenging it can be to read out the same information when it has been encoded in multiple unknown complex dynamical systems, each of the methods discussed previously relies on a different set of assumptions to make the problem tractable. Distribution alignment exploits transformation-invariant similarities between latent spaces to bring them into agreement, which typically involves outside knowledge of the task structure (e.g. common behaviors, or movement trajectories). This assumption imposes
a hard requirement on the latent structure; if the similarities are absent, too few, or too im-
precise, the alignment procedure breaks down. However, when this latent structure exists,
these methods can align recordings separated by large amounts of time or even from different
individuals. On the other hand, covariance inference makes no assumption on the underlying
distributional properties of activity but instead assumes that the latent model can be inferred
solely from the covariance matrix.

In the absence of stable latent space geometry, assuming structure in the way that the
recordings are generated allows different datasets to be stitched together. For instance, CCA
and LFADS assume that there is some temporal alignment (e.g. trial structure) in the recordings
that can be used to align samples across different sets of neurons or individuals. Distribution
alignment methods make no such assumptions on temporal structure but instead require that
the distributional properties of the two datasets are similar enough to bring them into align-
ment. Covariance inference methods assume that there is some overlap between neurons, or
at least that dynamics are stable across sequential recordings, and this can be used to stitch
together recordings temporally.

**Challenges**

Discovery of latent representations that captures key structures within different neural datasets
is an essential ingredient needed for alignment. However, using dimensionality reduction to
study neural population dynamics can be daunting because it is unsupervised and the result-
ing latent factors can be hard to interpret. Even with supervised labels or training data (like
movement kinematics, sleep or behavioral state), it can be hard to know when a latent space
model is sufficient to describe the structure of the data. Furthermore, it is unclear when la-
tent variables should be considered intrinsic to the functionality of a neural population, or
when they are just a concise description of activity. This is an instance of the classic statistical
conundrum of differentiating causation versus correlation. Even between neurons with corre-
lated activity, it is practically impossible to determine whether the correlation arises because
one neuron drives the activity of another, or if there is another hidden neuron driving both
of them. For example, it is not understood whether the ability to decode movement patterns
from activity in motor cortex is because neurons have been found which are designed to encode
movement, or because their activity is highly correlated with movement as a side-effect of their
true functional roles. Most dimensionality reduction techniques are agnostic to this distinction,
yet interpretable representations are really only those in which correlations have been reduced
to their underlying causes. Moving forward, careful design of experiments that can manipu-
late neural response properties (e.g., optogenetics) and examine the resulting changes in the
neural manifolds are needed to to understand latent space models in the context of underlying
circuit or systems-level mechanisms.
The methods described in the previous section assume that there is some persistent low-dimensional structure that can be observed across two recordings. However, there are many settings where changes in underlying brain states may occur due to sources of modulation like attention and engagement \cite{56, 57, 58, 59}, or shift as units die or electrodes move \cite{60, 61, 62}. To overcome these challenges, some supervised data or labels can be provided to the methods to help ground alignment when the shape of the two distributions is not sufficiently similar to find a common space for the datasets. Additionally, methods for partial domain adaptation and alignment may be utilized to facilitate matching in these conditions. Methods that can account for realistic shifts in neural data will allow for comparisons while also modeling state-dependent changes like attention that may modulate the neural distribution.

As the algorithms to perform these types of analyses grow more sophisticated and specialized, it will be important to have neural datasets that are generated across sufficiently diverse conditions to test methods on. In machine learning, common datasets used across the community have often spurred progress, from simple benchmarks like MNIST \cite{63} to more modern and powerful tools like the ImageNet dataset \cite{64}. In neuroscience, open-access brain imaging datasets and tools like those provided in the Allen Institute’s optophysiology-based Brain Observatory \cite{65} and their more recent high-density survey with Neuropixel probes \cite{66}, can be used to build a common accepted benchmark dataset to compare methods, as well as fertile ground for more in-depth studies. Moving forward, widespread data sharing like this will allow for more reproducible results and pave the way for more powerful and generalizable algorithms.

**Outlook**

In current brain machine interfaces (BMIs), decoders rely on stable mappings from neural activity to the variables of interest, which often degrade over time \cite{67}. Thus with the ability to directly compare latent representations and align them, it will be possible to correct for changes in neural distributions over time to develop more robust neural decoding interfaces. These controllers might be able to adapt to shifts in neural activity due to cognitive state or adaptation more effectively without collection of new training data, which is a critical feature for a controller which will be chronically implanted. We emphasize that latent space alignment should not be considered only a convenience for analysis, but also a tool for improving the weaknesses of current BMIs and neural prosthetics.

Many of the approaches that we discussed in this review deal primarily with linear methods for modeling the latent factors. However, computations and thus latent spaces that are present within small circuits and across larger neural populations, are likely to lie in spaces that linear models cannot appropriately capture. In these cases, more complex models like unions of subspaces and manifolds \cite{68, 69, 70} may be used to capture more complex low-dimensional
structures present in these data. By coupling multi-manifold models with matrix completion methods [71, 72], it could be possible to stitch together manifolds from many brain areas and start to fill in the blanks across recordings that are collected from different sites of contact in the brain spanning non-overlapping populations of neurons. Thus, through combining the techniques described in the review with more complex latent space models, we may be able to compare whole brain activity patterns even when individual recordings provide only a small view of the bigger picture.

While understanding of the physiological impact of neurological and neurodegenerative diseases has advanced tremendously in recent years, a description of the disruptions they cause to the functionality of neural circuits has not yet been possible. It is highly likely that when such a description is made, it will be through the study of shifts in the activity of many neurons in diseased brains. The alignment strategies outlined in this review and more sophisticated methods built on them, will provide new avenues to compare the latent spaces that the brain occupies in health and provide signatures of decline as it progresses into disease.

References


