Dimensionality reduction for neural population decoding

Charles R. Heller[1,2] & Stephen V. David[2*]

[1] Neuroscience Graduate Program, Oregon Health and Science University
[2] Oregon Hearing Research Center, Oregon Health and Science University
[*]Corresponding author (davids@ohsu.edu)

Abstract

Rapidly developing technology for large scale neural recordings has allowed researchers to measure the activity of hundreds to thousands of neurons at single cell resolution *in vivo*. Neural decoding analyses are a widely used tool used for investigating what information is represented in this complex, high-dimensional neural population activity. Most population decoding methods assume that correlated activity between neurons has been estimated accurately. In practice, this requires large amounts of data, both across observations and across neurons. Unfortunately, most experiments are fundamentally constrained by practical variables that limit the number of times the neural population can be observed under a single stimulus and/or behavior condition. Therefore, new analytical tools are required to study neural population coding while taking into account these limitations. Here, we present a simple and interpretable method for dimensionality reduction that allows neural decoding metrics to be calculated reliably, even when experimental trial numbers are limited. We illustrate the method using simulations and compare its performance to standard approaches for dimensionality reduction and decoding by applying it to single-unit electrophysiological data collected from auditory cortex.

1 Introduction

Neural decoding analysis identifies components of neural activity that carry information about the external world (*e.g.* stimulus identity). This approach can offer important insights into how and where information is encoded in the brain. For example, classic work by Britten et al. demonstrated that the ability of single neurons in area MT to decode visual stimuli closely corresponds to animal's perceptual performance.[1] Thus, by using decoding the authors identified a possible neural substrate for detection of motion direction.[1] Yet, behavior does not depend solely on single neurons. In the years since this work, many theoretical frameworks have been proposed for how information might be pooled across individual neurons into a population code.[2–8] One clear theme that has emerged from this work is that stimulus independent, correlated activity (*i.e.* noise correlations) between neurons may substantially impact information coding.[2,4–8] This has now been confirmed *in vivo* using decoding analysis to measure the information content of large neural populations.[9–11] Therefore, covariability between neurons must be taken into account when measuring population coding accuracy.

Under most experimental conditions, estimates of pairwise correlation between neurons is unreliable due to insufficient sampling (*e.g.* too few stimulus repeats).[12] In these situations, traditional decoding algorithms are likely to over-fit to noise in the neural data. This issue becomes even more apparent as the number of pairwise interactions that must be estimated increases, a situation that is becoming more common due to the recent explosion in large-scale neurophysiology techniques.[13] In some cases, *e.g.* for chronic recording experiments and anesthetized preps, the number of trials can be increased to circumvent this issue. However,

in behavioral experiments, where the number of trials is often fundamentally limited by variables such as animal performance, new analytical techniques for decoding are required.

Here, we present decoding-based dimensionality reduction (*dDR*), a simple and generalizable method for dimensionality reduction that significantly mitigates issues around estimating correlated variability in experiments with a relatively low ratio of observations to neurons. Our method takes advantage of recent observations that population covariability is often low-dimensional[14–17] to define a subspace where decoding analysis can be performed reliably while still preserving the dominant mode(s) of population covariability. The *dDR* method can be applied to data collected across many different stimulus and/or behavior conditions, making it a flexible tool for analyzing a wide range of experimental data.

We motivate the requirement for dimensionality reduction by illustrating how estimates of a popular information decoding metric, d'^2,[4,5] can be biased by small experimental sample sizes. Building on a simple two-neuron example, we demonstrate that low-dimensional structure in the covariability of simulated neural activity can be leveraged to reliably decode stimulus information, even when the number of neurons exceeds the number of experimental observations. Finally, we use a dataset collected from primary auditory cortex to highlight the advantages of using *dDR* for neural population decoding over standard principal component analysis.

2 Results

2.1 Small sample sizes limit the reliability of neural decoding analysis

Linear decoding, a common analytical method in neuroscience, identifies a linear, weighted combination of neural activity along which distinct conditions (*e.g.* different sensory stimuli) can be discriminated. In neural state-space, this weighted combination is referred to as the decoding axis, $\boldsymbol{w}_{opt}$, and it is the line along which the distance between stimulus classes is maximized and trial-trial variance is minimized (Fig. 1a, b). To quantify decoding accuracy, single-trial neural activity is projected onto this axis and a decoding metric is calculated to quantify the discriminability of the two stimulus classes. Here, we use d'^2, the discrete analog of Fisher Information.[4,5] This discriminability metric has been used in a number of previous studies[6,9–11,18] and has a direct relationship to classical signal detection theory.[4,19]

Looking at the simulated data in Figures 1a and b, one can appreciate that an accurate estimate of $\boldsymbol{w}_{opt}$ requires knowledge of both the mean response evoked by each stimulus class ($\boldsymbol{\mu}_a$ vs. $\boldsymbol{\mu}_b$), as well the population covariance, Σ (summarized by the ellipses in Fig. 1a and b). Indeed, d'^2, is directly dependent on these features:

$$d'^2 = \Delta\boldsymbol{\mu}^T \boldsymbol{w}_{opt} \tag{1}$$

$$\boldsymbol{w}_{opt} = \Sigma^{-1}\Delta\boldsymbol{\mu} \tag{2}$$

$$\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_a - \boldsymbol{\mu}_b \tag{3}$$

Where $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_b$ are the Nx1 vectors describing the mean response of an N-neuron population to two stimuli, a vs. b, respectively, and Σ is the average NxN covariance matrix $\frac{1}{2}(\Sigma_a + \Sigma_b)$ (*e.g.* Fig. 1c).

In practice, the pairwise spike count covariance between neurons (often referred to as noise correlation, or r_{sc}) is reported to be very small – on the order of 10^{-1} or 10^{-2}.[20–22] As we can see from the shuffled distribution in Figure 1a (bottom), this can pose a problem for accurate estimates of the off-diagonal elements in Σ, and, as a consequence, $\boldsymbol{w}_{opt}$ itself. This difficulty is especially pronounced when sample sizes are relatively small (compare Fig. 1a to b). The estimates of covariance and stimulus discriminability improve with increasing sample size, but robust performance is not reached until $\approx$ 100 stimulus repetitions, even for this case with relatively strong covariance (Fig. 1d). The sample sizes (*e.g.* number of trials) in most experiments, especially those involving animal behavior, are typically much lower, raising the question: How can one reliably quantify coding accuracy in large neural populations observed over relatively few trials?

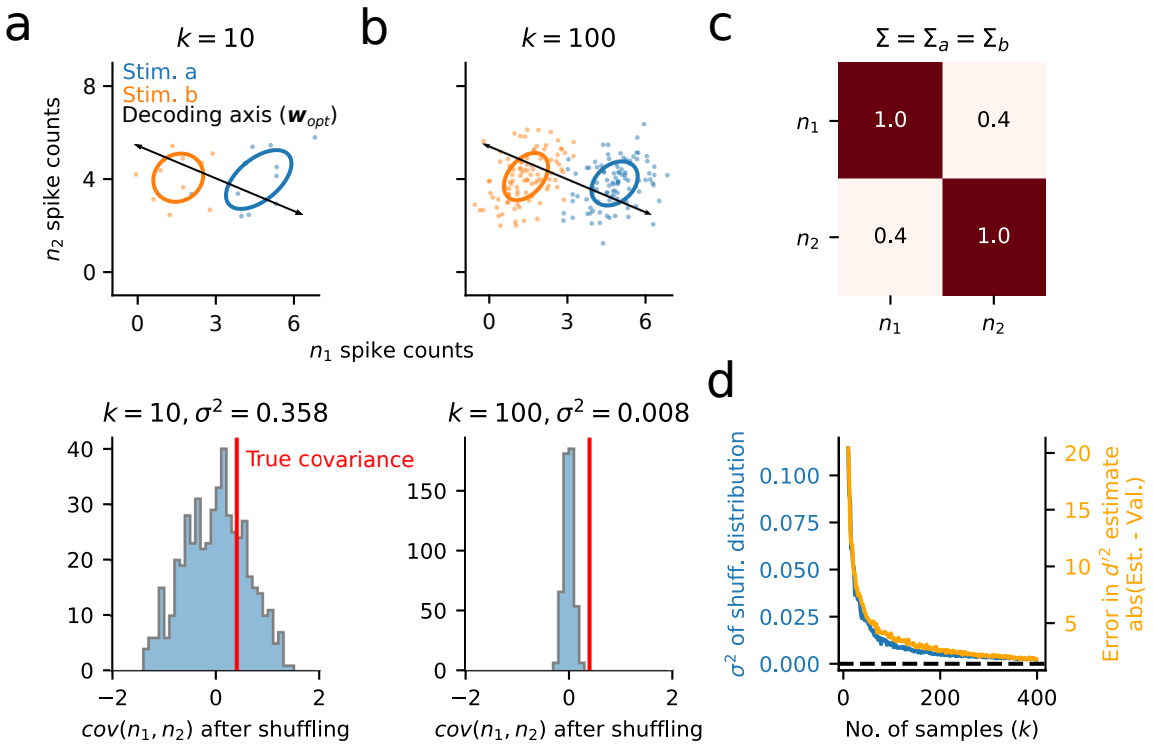

Figure 1: **Measurements of pairwise covariance and discriminability are unreliable when sampling is limited. a.** Top: $k = 10$ single trial spike count responses are drawn from standard multivariate Gaussians $\mathcal{N}(\boldsymbol{\mu}_a, \Sigma)$ and $\mathcal{N}(\boldsymbol{\mu}_b, \Sigma)$ corresponding to two different stimulus conditions, a and b. Ellipses show the standard deviation of spike counts across trials. Bottom: Reliability of the pairwise covariance estimate between neuron 1 (n_1) and neuron 2 (n_2) is calculated by shuffling values of n_1 500 times. The true covariance (red line) falls within this distribution, indicating that estimates of covariance are not reliable for $k = 10$. **b.** Same as in (a), but drawing $k = 100$ samples for each stimulus. The narrower distribution of permuted measures indicates a greater likelihood of identifying an accurate estimate of covariance. **c.** The covariance matrix, Σ, used to generate data in (a)/(b). The true pairwise covariance for this pair of simulated neurons has a value of 0.4. **d.** Variance (σ^2) of covariance estimates based on the permutation analysis in (a)/(b) for a range of sample sizes, k (blue). Variance decays as $\mathcal{O}(\frac{1}{k-1})$ (see Appendix). Overlaid is the difference in stimulus discriminability, d'^2 (Eqn. 1), between estimation and validation sets (50-50 split) estimated for each sample size (orange). Large values in the d'^2 difference for low k indicate overfitting of $\boldsymbol{w}_{opt}$ to the estimation data. This difference asymptotes toward zero as sample size increases and the estimate of covariance becomes reliable.

2.2 Neural activity is low-dimensional

Analysis of neural population data with dimensionality reduction has consistently revealed low-dimensional structure in neural activity.[23] Specifically, recent studies have found that stimulus-independent variability (*i.e.* noise correlations) is dominated by a small number of latent dimensions.[14,15,17,24] Noise correlations are thought to impact stimulus coding accuracy[7] and are known to depend on internal states, such as attention, that affect behavioral task performance.[15,16,20,25] These findings suggest that the space of neural activity relevant for understanding stimulus decoding, and its relationship to behavior, may be small relative to the total number of recorded neurons.

When population data exhibits low-dimensional structure, the largest eigenvector(s) of Σ (*i.e.* the top principal components of population activity) provides a reasonable, low-rank approximation to the full-rank covariance matrix. Importantly, these high variance dimensions of covariability can be estimated accurately

even from limited samples. To illustrate this, we simulated population spike counts, X, for $N = 100$ neurons by drawing k samples from a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance Σ (Eqn. 4).

$$X = \mathcal{N}(\boldsymbol{\mu}, \Sigma) + \epsilon_{indep.} \tag{4}$$

Where in Eqn. 4, $\epsilon_{indep.}$ represents a small amount of independent noise added to each neuron, effectively removing any significant structure in the smaller noise modes.

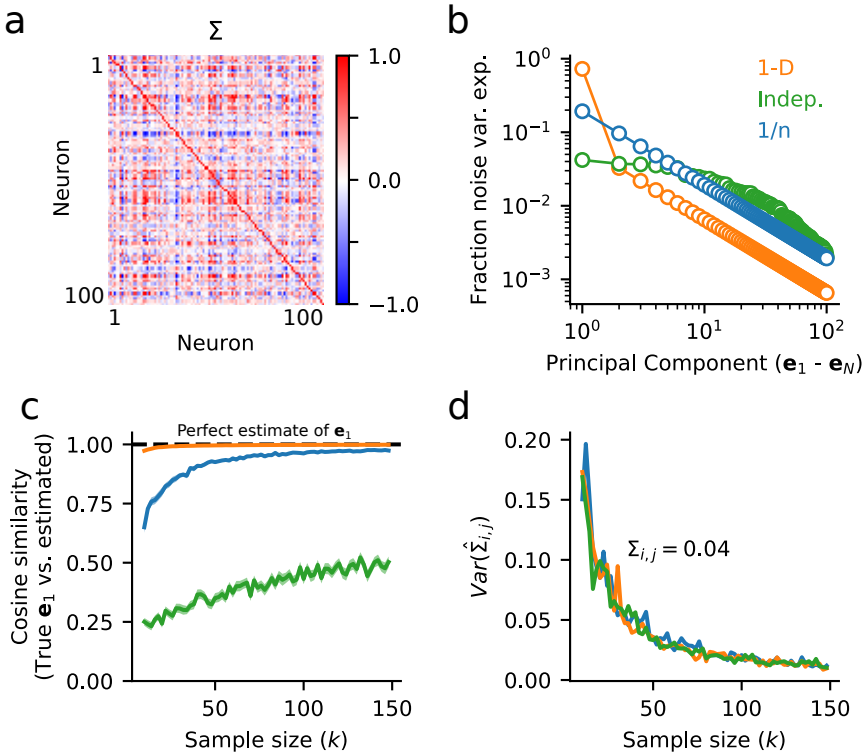

Figure 2: **Low-dimensional correlated activity can be estimated reliably for neural populations, even when pairwise covariance cannot. a.** Example covariance matrix, Σ, for a 100-neuron population with low-dimensional covariance structure. **b.** Scree plot shows the fraction of total population variance captured along each noise dimension, computed by PCA, for three different datasets with varying dimensionality. Orange: 1-dimensional noise (1-D), covariance matrix in (a); green: independent noise (Indep.); blue: power law decay ($1/n$). **c.** Surrogate datasets with varying numbers of samples, k, are drawn from the three noise distributions in (b). For each dataset, the cosine similarity between the estimate of the largest noise dimension, $\hat{\boldsymbol{e}}_1$, and the true noise dimension, $\boldsymbol{e}_1$, is plotted as function of sample size. For low-dimensional data, $\boldsymbol{e}_1$ can be estimated very reliably. **d.** Variance in the estimate of covariance, $\Sigma_{i,j}$, for two neurons with a true covariance of 0.04 is plotted as a function of the number of trials, as in Figure 1d. Even at sample sizes > 100, $Var(\hat{\Sigma}_{i,j}) \approx 0.02$, corresponding to a standard deviation of ≈ 0.14. Therefore, estimates of $\Sigma_{i,j}$, may be off by up to an order of magnitude. Note that the amount of uncertainty does not depend on the dimensionality of the data, and results for all three datasets overlap (see Appendix for an analytical derivation).

To investigate how different noise structures impact estimates of Σ, we simulated three different surrogate populations. First, we simulated data with just one large, significant noise dimension (Fig. 2, 1-D data, orange). In this case, the first eigenvector can be estimated reliably, even from just a few samples (Fig. 2c). However, when the noise is independent and shared approximately equally across all neurons, estimates of the first eigenvector are poor (Fig. 2, Indep. noise, green). These first two simulations represent extreme examples – in practice, population covariability tends to be spread across at least a few significant dimensions.[26] To investigate a scenario that more closely mirrors this structure, we simulated a third dataset

where the noise eigenspectrum decayed as $1/n$, where n goes from $n = 1$ to N. Recent studies of large neural populations suggest that this power law relationship is a reasonable approximation to real neural data.[26] In this case, by $k \approx 50$ trials, estimates of the first eigenvector are highly reliable, approaching a cosine similarity of ≈ 0.9 between the estimated and true eigenvectors (Fig. 2, $1/n$ noise, blue). In all simulations, regardless of dimensionality, we find that estimates of single elements of Σ (*i.e.* single noise correlation coefficients) are highly unreliable (Fig. 2d), as we see in the two-neuron example (Fig. 1d).

Collectively, these simulations demonstrate that accurate estimates of covariance need not necessarily be limited by uncertainty in estimates of individual noise correlation coefficients themselves. In the following sections we describe a simple decoding-based dimensionality reduction algorithm, *dDR*, that leverages low-dimensional structure in neural population activity to facilitate reliable measurements of neural decoding.

2.3 decoding-based Dimensionality Reduction (*dDR*)

The *dDR* algorithm operates on a pairwise basis. That is, given a set of neural data collected over S different conditions, a different *dDR* projection exists for each of the $\frac{S!}{2!(S-2)!}$ unique pairs. For simplicity, we will describe the case where $S = 2$, and consider these to be two unique stimulus conditions. However, note that the method can be applied in exactly the same manner to handle datasets with many different types and numbers of decoding conditions, where a unique *dDR* projection would then exist for each pair.

Let us consider the spiking response of an N-neuron population evoked by two different stimuli, S_a and S_b, over k-repetitions of each stimulus. From this data we form two response matrices, A and B, each with shape Nxk. Remembering that our goal is to estimate discriminability (d'^2, Eqn. 1), the *dDR* projection should seek to preserve information about both the mean response evoked by each stimulus condition, $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_b$, as well as the stimulus-independent noise covariance, Σ. Therefore, we define the first dimension of *dDR* to be the axis that maximally separate $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_b$. We call this the *signal* axis.

$$signal = \boldsymbol{\mu}_a - \boldsymbol{\mu}_b = \Delta\boldsymbol{\mu} \tag{5}$$

Next, we compute the first eigenvector of Σ, $\boldsymbol{e}_1$. This represents the largest noise mode of the neural population activity. Together, *signal* ($\Delta\boldsymbol{\mu}$) and $\boldsymbol{e}_1$ span the plane in state-space that is most optimized for reliable decoding. Finally, to form an orthonormal basis, we define the second *dDR* dimension as the axis orthogonal to $\Delta\boldsymbol{\mu}$ in this plane. As this second dimension is designed to preserve noise covariance, we call this the $noise_1$ axis.

$$noise_1 = \boldsymbol{e}_1 - \boldsymbol{e}_1\Delta\boldsymbol{\mu}^T \tag{6}$$

The process outlined above is schematized graphically in Figure 3.

Thus, the *signal* and $noise_1$ axes make up a 2xN set of weights, analogous to the loading vectors in standard *PCA*, for example. By projecting our Nxk data onto this new basis, we capture both the stimulus coding dimension ($\Delta\boldsymbol{\mu}$) and preserve the principal covariance dimension ($\boldsymbol{e}_1$), two critical features for measuring stimulus discriminability. Importantly, because $\boldsymbol{e}_1$ can be measured more robustly than Σ itself (Figure 2), performing this dimensionality reduction helps mitigate the issues we encounter due to small sample sizes and large neural datasets.

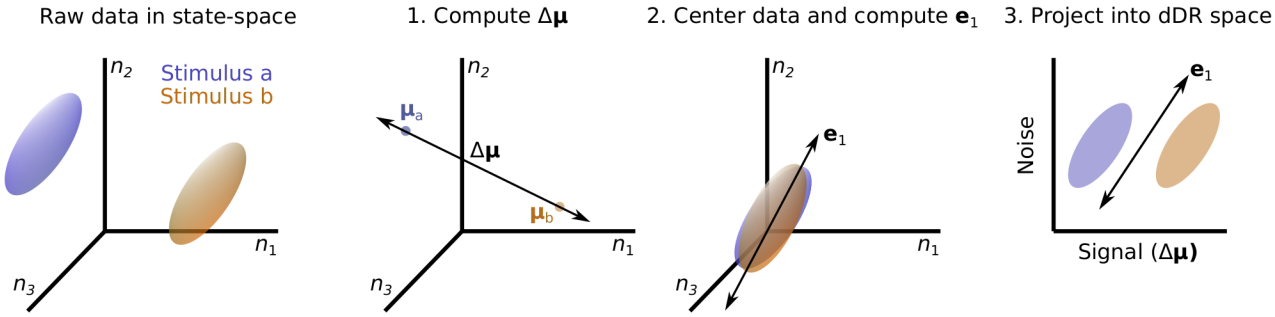

Figure 3: **decoding-based Dimensionality Reduction (dDR).** Left to right: Responses of 3 neurons (n_1, n_2, n_3) to two different stimuli are schematized in state-space. Ellipsoids illustrate the variability of responses across trials. **1.** To perform dDR, first the difference is computed between the two mean stimulus responses, $\Delta\boldsymbol{\mu}$. **2.** Next, the mean response is subtracted for each stimulus to center the data around 0, and PCA is used to identify the first eigenvector of the noise covariance matrix, $\boldsymbol{e}_1$ (additional noise dimensions $\boldsymbol{e}_m, m > 1$ can be computed, see text). **3.** Finally, the raw data are projected onto the plane defined by $\Delta\boldsymbol{\mu}$ and $\boldsymbol{e}_1$.

As mentioned in the previous section, neural data often contains more than one significant dimension of correlated trial-trial variability. To account for this, dDR can easily be extended to include more noise dimensions. To include additional dimensions, we deflate the spike count matrix, X, by subtracting out the $signal$ and $noise_1$ dimensions identified by standard dDR, then perform PCA on the residual matrix to identify m further $noise$ dimensions. Note, however, that for increasing m the variance captured by each dimension gets progressively smaller. Therefore, estimation of these subsequent noise dimensions becomes less reliable and will eventually become prone to over-fitting, especially with small sample sizes. For this reason, care should be taken when extending dDR in this way.

To demonstrate the performance of the dDR method, we generated three sample datasets containing $N = 100$ neurons and $S = 2$ stimulus conditions. Each of the three datasets contained unique noise covariance structure: 1. Σ contained one significant dimension (Fig. 4a) 2. Σ contained two significant dimensions (Fig. 4b) 3. Noise variance decayed as $1/n$ (Fig. 4c). For each dataset, we measured cross-validated d'^2 between stimulus condition a and stimulus condition b using standard dDR with one noise dimension (dDR_1), with two noise dimensions (dDR_2), or with three noise dimensions (dDR_3). We also estimated d'^2 using the full-rank data, without performing dDR. Figure 4 plots the decoding performance of each method as a function of sample size (*i.e.* number of stimulus repetitions). In each case, d'^2 is normalized to the asymptotic performance of the full-rank approach, when the number of samples is >> than the number of neurons. This provides an approximate estimate of true discriminability for the population.

In contrast to the full-rank data where overfitting leads to dramatic underestimation of d'^2 on the test data for most sample sizes (Fig. 4 grey lines), we find that d'^2 estimates after performing dDR are substantially more accurate and, critically, more reliable across sample sizes. That is, asymptotic performance of the dDR method is reached much more quickly than for the full-rank method.

For the one-dimensional noise case, note that there is no benefit of including additional dDR dimensions (Fig. 4a), while for the higher dimensional data shown in Figure 4b-c, we see some improvements with dDR_2 and dDR_3. However, these benefits don't begin to appear until k becomes large and they diminish with increasing noise dimensions – the improvement of dDR_2 over dDR_1 is larger than that of dDR_3 to dDR_2 Fig. 4b-c. This is because subsequent noise dimensions are, by definition, lower variance and therefore more difficult to estimate reliably from limited sample sizes.

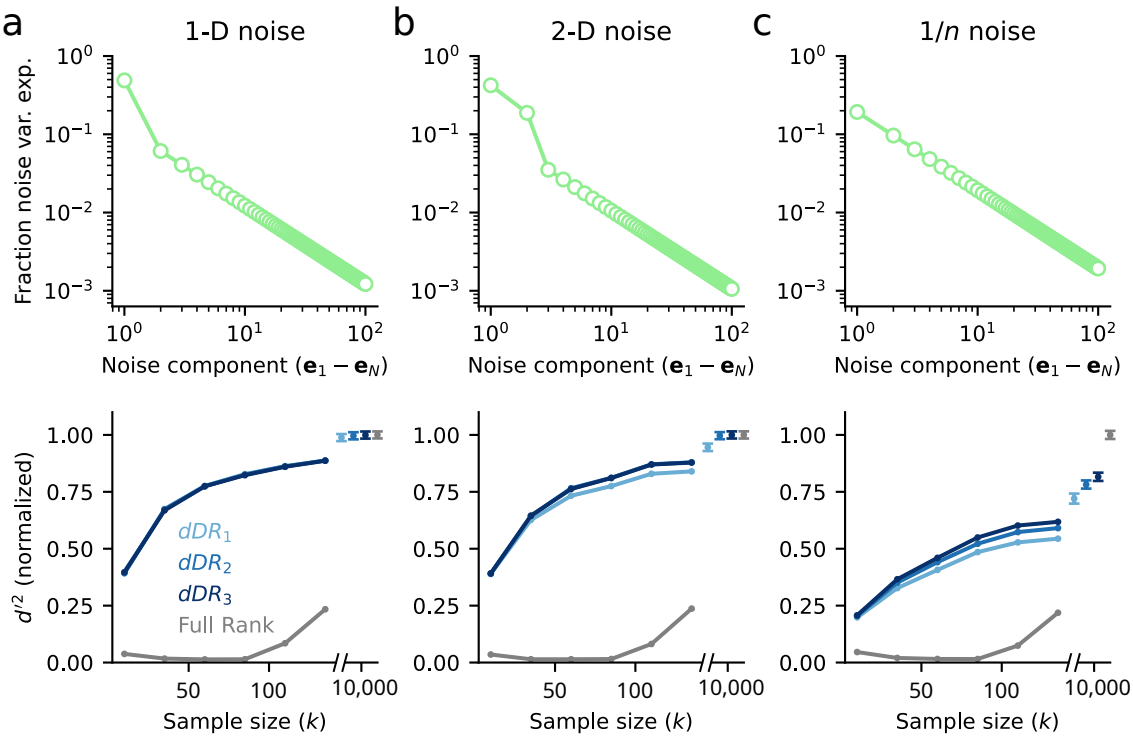

Figure 4: **Evaluation of decoding accuracy and reliability with dDR. a.** Analysis of data with one-dimensional (1-D) noise covariance. For each sample size, k, 100 datasets were generated from the same multivariate Gaussian distribution (Eqn. 4) where Σ was a rank-one covariance matrix and the mean response vector, $\boldsymbol{\mu}$, corresponded to one of two stimulus conditions, a or b. Top: Scree plot of noise covariance. Bottom: Cross-validated discriminability, d'^2, between a and b computed with full-rank data and with dDR using one (dDR_1), two (dDR_2) or three (dDR_3) noise dimensions, as a function of sample size. Mean d'^2 across all 100 surrogate datasets is shown here. For $k >> N$, the dDR results converge to the asymptotic value of the full-rank d'^2. However, even for small k, the dDR analyses estimates are much more accurate than the full-rank approach. **b.** Same as in (a), but for two-dimensional noise covariance data. In this case, dDR_2 captures the second noise dimension and outperforms the standard 1-D approach (dDR_1) **c.** Same as in (a) and (b), but for $1/n$ noise covariance.

2.4 *dDR* recovers more decoding information than standard principal component analysis

One popular method for dimensionality reduction of neural data is principal component analysis (PCA).[23] Generally speaking, PCA can be implemented on neural data in one of two ways: single trial PCA or trial-averaged PCA. In the single trial approach ($stPCA$), principal components are measured across all single trials and all experimental conditions. The resulting PCs capture variance both across single trials and across different *e.g.* stimulus conditions. In trial-averaged PCA ($taPCA$), single trial responses are averaged per experimental condition first, and PCs are measured over the resulting N-neuron x S-condition spike count matrix. In this case, for different stimulus conditions, the PCs specifically capture variance of stimulus-evoked activity rather than trial-trial variability, making this a more logical choice for many decoding applications. In the case of $S = 2$, as we have outlined above for the dDR illustration (Fig. 3), $taPCA$ is equivalent to $\Delta\boldsymbol{\mu}$, the first dDR dimension. Thus, dDR can roughly be thought of as a way to combine $taPCA$ and $stPCA$ – $taPCA$ identifies the *signal* dimension and $stPCA$ identifies the *noise* dimension(s).

To demonstrate the relative decoding performance achieved using each method, we applied each to a dataset collected from primary auditory cortex in an awake, passively listening ferret. $N = 52$ neurons were recorded simultaneously using a 64-channel laminar probe[27] as in.[28–30] Auditory stimuli consisting of narrowband (0.3 octave bandwidth) noise bursts were presented alone (-Inf dB) or with a pure tone embedded at

varying SNRs (0 dB, −5 dB, −10 dB) in the hemifield contralateral to the recording site (see Experimental Methods). Each stimulus was repeated 50 times. For *stPCA* and *dDR*, we selected only the top $m = 2$ total dimensions, and for *taPCA*, we selected the single dimension, $\Delta\boldsymbol{\mu}$, that exists for $S = 2$. This dataset allowed us to investigate how each dimensionality reduction method performs for two distinct, behaviorally relevant neural decoding questions: How well can neural activity perform fine discriminations (*tone-in-noise detection*), discriminating noise alone vs. noise with tone? How well can it perform coarse discriminations (*frequency discrimination*), discriminating noise centered at frequency A vs. noise at frequency B?

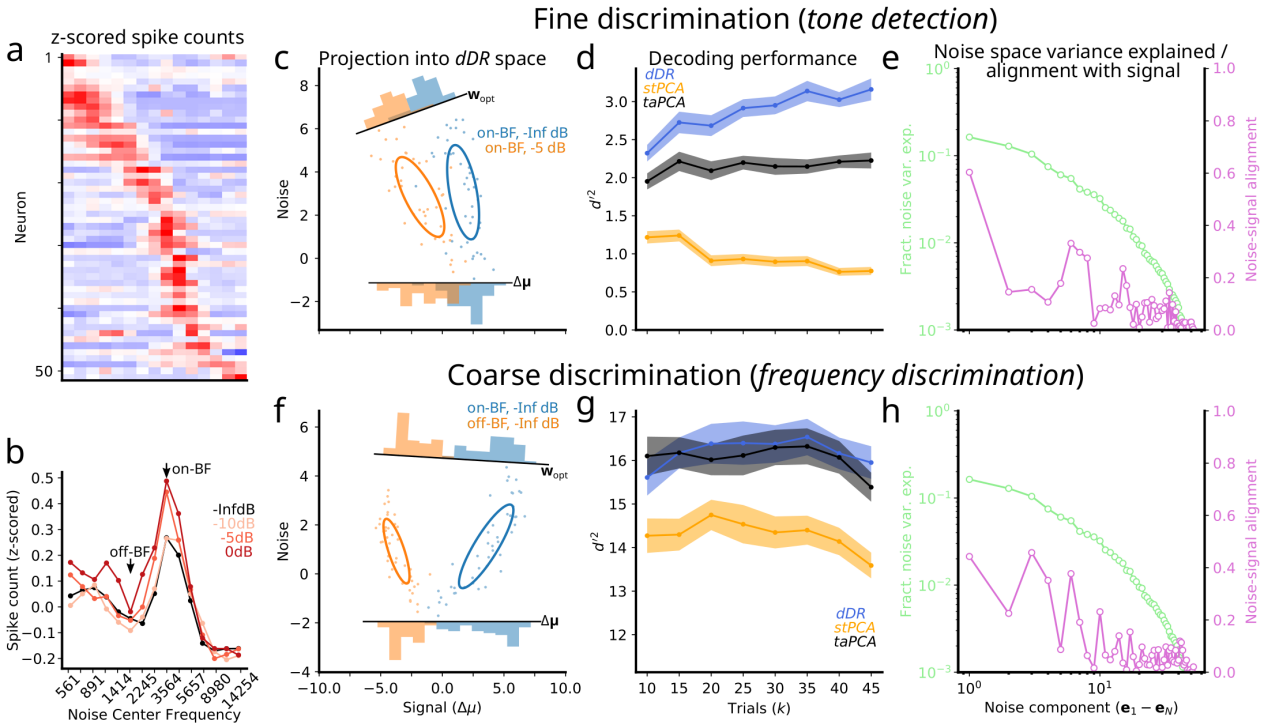

Figure 5: *dDR* **outperforms** *PCA* **for fine sensory discrimination. a.** Heatmap shows mean z-scored spike counts of $N = 52$ simultaneously recorded units for 15 different narrowband noise bursts (0.3 octave bandwidth tiling 5 octaves, x-axis). Each row shows tuning for one unit, with red indicating higher firing rate response. **b.** Population tuning curve for noise alone (black, data from panel a) and noise plus −10, −5, and 0 dB tones (light to dark red), computed by averaging tuning curves across neurons. **c-e.** Decoding analysis for tone-in-noise detection. **c.** Scatter plot compares single trial responses to noise alone at best frequency (on-BF, blue) vs. noise + −5dB tone (orange), projected into *dDR* space. Ellipses show standard deviation across trials, marginal histograms show projection of data onto optimal decoding axis ($\boldsymbol{w}_{opt}$) or onto $\Delta\boldsymbol{\mu}$ (equivalent to performing trial-averaged *PCA*). **d.** Estimate of d'^2 as a function of sample size (number of trials, k) using each dimensionality reduction method. For each point, d'^2 was averaged over 100 random samplings of k trials, drawn without replacement. Shading indicates standard error. **e.** Fraction variance explained by each noise component (green) computed by performing *PCA* on mean-centered single trial data. The alignment of each noise component with the signal axis is shown in purple. **f-h** Same as panels (c)-(e), for noise alone on-BF vs. noise along off-BF (see panel b).

The A1 dataset displayed a range of frequency tuning (Fig. 5a), with the majority of units tuned to ≈ 3.5 kHz. We therefore defined this as the best frequency of the recording site (on-BF, Fig. 5b). For *tone detection*, we measured discriminability (d'^2, Eqn. 1) between on-BF noise alone (on-BF, -Inf dB) and on-BF noise plus tone (on-BF, −5 dB), which each drove similar sensory responses (Fig. 5b-c). For *frequency discrimination*, we measured discriminability between the neural responses to on-BF noise and off-BF noise, where off-BF was defined as ≈ 1 octave away from BF, and drove a very different population response (Fig. 5b, f). In both cases, *taPCA* and *dDR* outperformed *stPCA* (Fig. 5d, g). This first

result is unsurprising due to the fact that *stPCA* is the only method not explicitly designed to capture variability in the sensory response. The top *PCs* are dominated by dimensions of trial-trial variability that do not necessarily contain stimulus information and thus underestimate d'^2 relative to the other two methods.

We also find that *dDR* consistently performs as well or better than *taPCA*. For the *tone detection* data, the sensory signal ($\Delta\boldsymbol{\mu}$) is small (*i.e.*, trial-averaged responses to the two stimuli were similar) and covariability is partly aligned with $\Delta\boldsymbol{\mu}$. Under these conditions, *dDR* makes use of correlated activity to optimize the decoding axis ($\boldsymbol{w}_{opt}$) and improve discriminability. *taPCA*, on the other hand, has no information about these correlations and is therefore equivalent to projecting the single trial responses onto the *signal* axis, $\Delta\boldsymbol{\mu}$. Thus, it underestimates d'^2 (Fig. 5c, d). In the *frequency discrimination* example, $\Delta\boldsymbol{\mu}$ is large. The covariability has similar magnitude to the previous example, but it is not aligned to the discrimination axis, and thus has no impact on $\boldsymbol{w}_{opt}$. In this case, *dDR* and *taPCA* perform similarly (Fig. 5f-g). These examples highlight that under behaviorally relevant conditions, *dDR* can offer a significant improvement over standard *PCA*, even with as few as 10 trials.

3 Discussion

We have described a new, simple method for dimensionality reduction of neural population data, *dDR*. This approach combines strategies for both trial-averaged *PCA* and single-trial *PCA* to identify important dimensions of population activity that govern neural coding accuracy. Using both simulated and real neural data, we demonstrated that the method performs robustly for neural decoding analysis in low experimental trial count regimes where the performance of full-rank methods break down. Across a range of behaviorally relevant stimulus conditions, *dDR* consistently performs as well or better than standard principal component analysis.

3.1 Applications

dDR is designed to optimize the performance of linear decoding methods in situations where sample sizes are small. This is often the case for neurophysiology data collected from behaving animals, where the number of stimulus and/or behavior conditions are fundamentally limited by task performance. In these situations, using full-rank decoding methods is unfeasible as it leads to dramatic overfitting and unreliable performance.[12] Dimensionality reduction methods, such as *PCA*, can be used to mitigate overfitting issues. However, the correct implementation of *PCA* in neural data is often ambiguous, and multiple different approaches to dimensionality reduction have been proposed.[23] We suggest *dDR* as a simple, standardized alternative that captures the strengths of different *PCA* approaches. Unlike conventional *PCA*, the *signal* and *noise* axes that comprise the *dDR* space have clear interpretations with respect to neural decoding. Importantly, *dDR* components explicitly preserve stimulus-independent population covariability. In addition to being important for overall information coding, this covariability is known to depend on behavior state[15,16,20,25,31] and stimulus condition.[21,32–34] Therefore, approaches that do not preserve these dynamics, such as trial-averaged *PCA*, may not accurately characterize how information coding changes across varying behavior and/or stimulus conditions.

3.2 Interpretability and visualization

A key benefit of *dDR* is that the axes making up the *dDR* subspace are easily interpretable: The first axis (*signal*) represents the dimension with maximal information about the difference in evoked activity between the two conditions to be decoded, and the second (*noise*) axis captures the largest mode of condition-independent population covariability in the data. Therefore, within the *dDR* framework it is straightforward to investigate how this covariability interacts with discrimination, an important question for neural information coding. Further, standard *dDR* (with a single noise dimension) can be used to easily visualize high-dimensional population data, as in Fig. 5. For methods like *PCA*, it can be difficult to dissociate signal and noise dimensions, as the individual principal components can represent an ambiguous mix of task conditions, stimulus conditions, and trial-trial variability.[35] Moreover, with *PCA* the number of

total dimensions is typically selected based on their cumulative variance explained, rather than by selecting the dimensions that are of interest for decoding, as in *dDR*.

3.3 Extensions

Latent variable estimation:

dDR makes the assumption that latent sources of low-dimensional neural variability can be captured using simple, linear methods, such as *PCA*. While these methods often seem to recover meaningful dimensions of neural variability,[16] a growing body of work is investigating new, alternative methods for estimating these latent dynamics,[15,17,36,37] and this work will continue to lead to important insights about the nature of shared variability in neural populations.

We suggest that *dDR* can be extended to incorporate these new methods. For example, rather than defining *dDR* on a strictly per decoding pair basis, a global noise axis could be identified across all experimental conditions using a custom latent variable method. This could then be applied to the decoding-based dimensionality reduction such that the resulting *dDR* space explicitly preserves activity in this latent space to investigate how it interacts with coding.

Incorporating additional *dDR* dimensions:

In this work we have described *dDR* primarily as a transformation from N-dimensions to two dimensions, *signal* and *noise*, with the exception of Figure 4. In our code repository, `https://github.com/crheller/dDR`, we include examples that demonstrate how the *dDR* method can be extended to include additional dimensions. However, as discussed in the main text, it is important to remember that estimates of neural variability beyond the first principal component may become unreliable as variance along these dimensions gets progressively smaller, especially in low trial regimes. In short, while information may be contained in dimensions $> m = 2$, caution should be used to ensure that these dimensions can be estimated reliably.

3.4 Code availability

We provide Python code for *dDR* which can be downloaded and installed by following the instructions at `https://github.com/crheller/dDR`. We also include a short demo notebook that highlights the basic work flow and implementation of the method to simulated data. All code used to generate the figures in this manuscript is available in the repository.

4 Experimental Methods

4.1 Surgical procedure

All procedures were performed in accordance with the Oregon Health and Science University Institutional Animal Care and Use Committee (IACUC) and conform to standards of the Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC). The surgical approach was similar to that described previously.[38] Adult male ferrets were acquired from an animal supplier (Marshall Farms). Head-post implantation surgeries were then performed in order to permit head-fixation during neurophysiology recordings. Two stainless steel head-posts were fixed to the animal along the midline using bone cement (Palacos), which bonded to the skull and to stainless steel screws that were inserted into the skull. After a two-week recovery period, animals were habituated to a head-fixed posture and auditory stimulation. At this point, a small (0.5 - 1 mm) craniotomy was opened above primary auditory cortex (A1) for neurophysiological recordings.

4.2 Neurophysiology

Recording procedures followed those described previously.[28,29] Briefly, upon opening a craniotomy, 1 - 4 tungsten micro-electrodes (FHC, 1-5 MΩ) were inserted to characterize the tuning and response latency

of the region of cortex. Sites were identified as A1 by characteristic short latency responses, frequency selectivity, and tonotopic gradients across multiple penetrations.[39] Subsequent penetrations were made with a 64-channel silicon electrode array.[27] Electrode contacts were spaced 20 μm horizontally and 25 μm vertically, collectively spanning 1.05 mm of cortex. Data were amplified (RHD 128-channel headstage, Intan Technologies), digitized at 30 KHz (Open Ephys[40]) and saved to disk for further analysis.

Spikes were sorted offline using Kilosort2 (`https://github.com/MouseLand/Kilosort2`). Spike sorting results were manually curated in phy (`https://github.com/cortex-lab/phy`). For all sorted and curated spike clusters, a contamination percentage was computed by measuring the cluster isolation in feature space. All sorted units with contamination percentage less than or equal to 5 percent were classified as single-unit activity. All other stable units that did not meet this isolation criterion were labeled as multi-unit activity. Both single and multi-units were included in all analyses.

4.3 Acoustic stimuli

Digital acoustic signals were transformed to analog (National Instruments), amplified (Crown), and delivered through a free-field speaker (Manger) placed 80 cm from the animal's head and 30° contralateral to the the hemisphere in which neural activity was recorded. Stimulation was controlled using custom MATLAB software (`https://bitbucket.org/lbhb/baphy`), and all experiments took place inside a custom double-walled sound-isolating chamber (Professional Model, Gretch-Ken).

Auditory stimuli consisted of narrowband white noise stimuli with $\approx$ 0.3 octave bandwidth. In total, we presented fifteen distinct, non-overlapping noise bursts spanning a 5 octave range. Each noise was presented alone (-Inf dB) condition, or with a pure tone embedded at its center frequency for a range of different signal to noise ratios ($-$10dB, $-$5dB, 0dB). Thus, each experiment consisted of 60 unique stimuli (4 SNR conditions X 15 center frequencies). Overall sound level was set to 60 dB SPL. Stimuli were 300ms in duration with 200ms ISI and each sound was repeated 50 times per experiment in a pseudo-random sequence.

5 Appendix

5.1 Variance of parameter estimates

In this work, we approximate the spike counts of a neural population as being drawn from a multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance Σ. The accuracy of our estimates of these respective parameters depends on how large the sample size is. That is, if we draw just two samples from the distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, our estimates of $\boldsymbol{\mu}$ and Σ will be highly variable across repeated iterations of this sampling. This means that when sample size is small we can't be certain of the measured parameter values. Here, we provide a brief derivation showing how the uncertainty in each of these parameter values depends on sample size, k.

Mean (μ):
We will investigate the mean of just a single neuron, μ, for simplicity. Here, and in the following cases, we assume the data has been centered such that the mean response across all trials for each neuron is zero. Consider repeated samples of a random variable, x_i, drawn from $\mathcal{N}(0, \sigma^2)$. Let us define the variable Y to be the mean of a random sequence of i.i.d. numbers, $x_i ... x_n$ with $E[x_i] = \mu$ and $Var(x_i) = \sigma^2$.

$$Y = \frac{1}{k} \sum_{i=1}^{k} x_i$$

Next, we can ask how *variable* our estimates of Y are with increasing sample size.

$$Var(Y) = Var\Big(\frac{1}{k} \sum_{i=1}^{k} x_i\Big)$$

$$Var(Y) = \frac{1}{k^2}\sum_{i=1}^{k} Var(x_i)$$

$$Var(Y) = \frac{1}{k^2}\sum_{i=1}^{k} \sigma^2$$

$$Var(Y) = \frac{\sigma^2}{k}$$

Thus, estimates of the mean spike count for a single neuron, μ, decay with increasing sample size as:

$$\mathcal{O}(\frac{1}{k}) \tag{7}$$

Single neuron variance (Σ_{diag}):
For the variance of single neurons, *i.e.* the diagonal elements of Σ, we can similarly derive their uncertainty as a function of k by defining Y as:

$$Y = \frac{1}{k-1}\sum_{i=1}^{k} x_i^2$$

$$Var(Y) = Var\Big(\frac{1}{k-1}\sum_{i=1}^{k} x_i^2\Big)$$

$$Var(Y) = \frac{1}{(k-1)^2}\sum_{i=1}^{k} Var(x_i^2)$$

$$Var(Y) = \frac{1}{(k-1)^2}\sum_{i=1}^{k} 2\sigma^4$$

$$Var(Y) = \frac{2\sigma^4}{k-1}$$

Thus, the uncertainty in single neuron variance depends the neuron's true variance σ^2, and decays as a function of sample size k.

$$\mathcal{O}(\frac{1}{k-1}) \tag{8}$$

Covariance (Σ):
And finally, for uncertainty of the covariance between two correlated neurons x and y, *i.e.* the off-diagonal elements of Σ, we define Y as:

$$Y = \frac{1}{k-1}\sum_{i=1}^{k} x_i y_i$$

As above, can write:

$$Var(Y) = Var\Big(\frac{1}{k-1}\sum_{i=1}^{k} x_i y_i\Big)$$

$$Var(Y) = \frac{1}{(k-1)^2}\sum_{i=1}^{k} Var(x_i y_i)$$

Then, using the three following identities:

$$Var(xy) = E[x^2y^2] - E[xy]^2$$

$$E[x^2y^2] = cov(x^2, y^2) + E[x^2]E[y^2]$$
$$E[XY]^2 = \big(cov(x, y) + E[x]E[y]\big)^2$$

We can write the following expression for the $Var(Y)$, taking $E[x] = E[y] = 0$:

$$Var(Y) = \Big(\frac{2(\Sigma_{x,y}^2)^2 + \sigma_x^2\sigma_y^2 - \Sigma_{x,y}}{k-1}\Big)$$

where $\Sigma_{x,y}$ is the true covariance between neurons x and y, and σ_x^2 and σ_y^2 represent each neuron's respective independent variance. Thus, as for single neuron variance, the uncertainty in covariance decays with sample size, k (Eqn. 9). Note, though, that typical covariance values are much smaller than single neuron variance, making this a much more difficult parameter to estimate given a particular sample size.

$$\mathcal{O}\Big(\frac{1}{k-1}\Big) \tag{9}$$

Acknowledgements

This work was supported by a National Science Foundation Graduate Research Fellowship (NSF GRFP, GVPRS0015A2) (CRH), the National Institute of Health (NIH, R01 DC0495) (SVD), Achievement Rewards for College Scientists (ARCS) Portland chapter (CRH), and by the Tartar Trust at Oregon Health and Science University (CRH).

References

[1] K. H. Britten, M. N. Shadlen, W. T. Newsome, and J. A. Movshon. The analysis of visual motion: a comparison of neuronal and psychophysical performance. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 12(12):4745–4765, December 1992.

[2] Ehud Zohary, Michael N. Shadlen, and William T. Newsome. Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature*, 370(6485):140–143, July 1994.

[3] Michael N. Shadlen and William T. Newsome. The Variable Discharge of Cortical Neurons: Implications for Connectivity, Computation, and Information Coding. *Journal of Neuroscience*, 18(10):3870–3896, May 1998.

[4] L. F. Abbott and Peter Dayan. The Effect of Correlated Variability on the Accuracy of a Population Code. *Neural Computation*, 11(1):91–101, January 1999.

[5] Peter Dayan and L. F. Abbott. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. Computational neuroscience. Massachusetts Institute of Technology Press, Cambridge, Mass, 2001.

[6] Bruno B. Averbeck and Daeyeol Lee. Effects of noise correlations on information encoding and decoding. *Journal of Neurophysiology*, 95(6):3633–3644, June 2006.

[7] Bruno B. Averbeck, Peter E. Latham, and Alexandre Pouget. Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, 7(5):358–366, May 2006.

[8] Xaq Pitkow, Sheng Liu, Dora E. Angelaki, Gregory C. DeAngelis, and Alexandre Pouget. How Can Single Sensory Neurons Predict Behavior? *Neuron*, 87(2):411–423, July 2015.

[9] Ramon Bartolo, Richard C. Saunders, Andrew R. Mitz, and Bruno B. Averbeck. Information limiting correlations in large neural populations. *The Journal of Neuroscience*, pages 2072–19, January 2020.

[10] MohammadMehdi Kafashan, Anna Jaffe, Selmaan N. Chettih, Ramon Nogueira, Iñigo Arandia-Romero, Christopher D. Harvey, Rubén Moreno-Bote, and Jan Drugowitsch. Scaling of information in large neural populations reveals signatures of information-limiting correlations. *bioRxiv*, page 2020.01.10.902171, January 2020.

[11] Oleg I. Rumyantsev, Jérôme A. Lecoq, Oscar Hernandez, Yanping Zhang, Joan Savall, Radosław Chrapkiewicz, Jane Li, Hongkui Zeng, Surya Ganguli, and Mark J. Schnitzer. Fundamental bounds on the fidelity of sensory cortical coding. *Nature*, March 2020.

[12] Ingmar Kanitscheider, Ruben Coen-Cagli, Adam Kohn, and Alexandre Pouget. Measuring Fisher Information Accurately in Correlated Neural Populations. *PLOS Computational Biology*, 11(6):e1004218, June 2015. Publisher: Public Library of Science.

[13] Ian H. Stevenson and Konrad P. Kording. How advances in neural recording affect data analysis. *Nature Neuroscience*, 14(2):139–142, February 2011. Number: 2 Publisher: Nature Publishing Group.

[14] Benjamin R. Cowley, Adam C. Snyder, Katerina Acar, Ryan C. Williamson, Byron M. Yu, and Matthew A. Smith. Slow Drift of Neural Activity as a Signature of Impulsivity in Macaque Visual and Prefrontal Cortex. *Neuron*, 108(3):551–567.e8, November 2020. Publisher: Elsevier.

[15] Neil C Rabinowitz, Robbe L Goris, Marlene Cohen, and Eero P Simoncelli. Attention stabilizes the shared gain of V4 populations. *eLife*, 4:e08998, November 2015.

[16] A. M. Ni, D. A. Ruff, J. J. Alberts, J. Symmonds, and M. R. Cohen. Learning and attention reveal a general relationship between population activity and behavior. *Science*, 359(6374):463–465, January 2018.

[17] Alexander S. Ecker, Philipp Berens, R. James Cotton, Manivannan Subramaniyan, George H. Denfield, Cathryn R. Cadwell, Stelios M. Smirnakis, Matthias Bethge, and Andreas S. Tolias. State Dependence of Noise Correlations in Macaque Primary Visual Cortex. *Neuron*, 82(1):235–248, April 2014.

[18] Rubén Moreno-Bote, Jeffrey Beck, Ingmar Kanitscheider, Xaq Pitkow, Peter Latham, and Alexandre Pouget. Information-limiting correlations. *Nature Neuroscience*, 17(10):1410–1417, October 2014.

[19] David M. Green and John A. Swets. *Signal detection theory and psychophysics.* Signal detection theory and psychophysics. John Wiley, Oxford, England, 1966. Pages: xi, 455.

[20] Marlene R. Cohen and John H. R. Maunsell. Attention improves performance primarily by reducing interneuronal correlations. *Nature Neuroscience*, 12(12):1594–1600, December 2009.

[21] Marlene R Cohen and Adam Kohn. Measuring and interpreting neuronal correlations. *Nature Neuroscience*, 14(7):811–819, July 2011.

[22] Alexander S. Ecker, Philipp Berens, Georgios A. Keliris, Matthias Bethge, Nikos K. Logothetis, and Andreas S. Tolias. Decorrelated neuronal firing in cortical microcircuits. *Science (New York, N.Y.)*, 327(5965):584–587, January 2010.

[23] John P Cunningham and Byron M Yu. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509, November 2014.

[24] Robbe L T Goris, J Anthony Movshon, and Eero P Simoncelli. Partitioning neuronal variability. *Nature Neuroscience*, 17(6):858–865, June 2014.

[25] J. D. Downer, M. Niwa, and M. L. Sutter. Task Engagement Selectively Modulates Neural Correlations in Primary Auditory Cortex. *Journal of Neuroscience*, 35(19):7565–7574, May 2015.

[26] Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D. Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365, July 2019.

[27] Jiangang Du, Timothy J. Blanche, Reid R. Harrison, Henry A. Lester, and Sotiris C. Masmanidis. Multiplexed, High Density Electrophysiology with Nanofabricated Neural Probes. *PLOS ONE*, 6(10):e26204, October 2011.

[28] Daniela Saderi, Zachary P Schwartz, Charles R Heller, Jacob R Pennington, and Stephen V David. Dissociation of task engagement and arousal effects in auditory cortex and midbrain. *eLife*, 10:e60153, February 2021. Publisher: eLife Sciences Publications, Ltd.

[29] Charles R. Heller, Zachary P. Schwartz, Daniela Saderi, and Stephen V. David. Selective effects of arousal on population coding of natural sounds in auditory cortex. *bioRxiv*, page 2020.08.31.276584, December 2020. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[30] Jacob Pennington and Stephen David. Complementary effects of adaptation and gain control on sound encoding in primary auditory cortex. preprint, Neuroscience, January 2020.

[31] Martina Valente, Giuseppe Pica, Caroline A. Runyan, Ari S. Morcos, Christopher D. Harvey, and Stefano Panzeri. Correlations enhance the behavioral readout of neural population activity in association cortex. preprint, Neuroscience, April 2020.

[32] Joel Zylberberg, Jon Cafaro, Maxwell H. Turner, Eric Shea-Brown, and Fred Rieke. Direction-Selective Circuits Shape Noise to Ensure a Precise Population Code. *Neuron*, 89(2):369–383, January 2016.

[33] Felix Franke, Michele Fiscella, Maksim Sevelev, Botond Roska, Andreas Hierlemann, and Rava Azeredo da Silveira. Structures of Neural Correlation and How They Favor Coding. *Neuron*, 89(2):409–422, January 2016. Publisher: Elsevier.

[34] Douglas A. Ruff and Marlene R. Cohen. Stimulus Dependence of Correlated Variability across Cortical Areas. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 36(28):7546–7556, July 2016.

[35] Dmitry Kobak, Wieland Brendel, Christos Constantinidis, Claudia E Feierstein, Adam Kepecs, Zachary F Mainen, Xue-Lian Qi, Ranulfo Romo, Naoshige Uchida, and Christian K Machens. Demixed principal component analysis of neural population data. *eLife*, 5:e10989, April 2016. Publisher: eLife Sciences Publications, Ltd.

[36] Matthew R. Whiteway, Bruno Averbeck, and Daniel A. Butts. A latent variable approach to decoding neural population activity. *bioRxiv*, page 2020.01.06.896423, January 2020. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[37] Byron M. Yu, John P. Cunningham, Gopal Santhanam, Stephen I. Ryu, Krishna V. Shenoy, and Maneesh Sahani. Gaussian-Process Factor Analysis for Low-Dimensional Single-Trial Analysis of Neural Population Activity. *Journal of Neurophysiology*, 102(1):614–635, July 2009.

[38] S. J. Slee and S. V. David. Rapid Task-Related Plasticity of Spectrotemporal Receptive Fields in the Auditory Midbrain. *Journal of Neuroscience*, 35(38):13090–13102, September 2015.

[39] Jennifer K. Bizley, Fernando R. Nodal, Israel Nelken, and Andrew J. King. Functional organization of ferret auditory cortex. *Cerebral Cortex (New York, N.Y.: 1991)*, 15(10):1637–1653, October 2005.

[40] Joshua H Siegle, Aarón Cuevas López, Yogi A Patel, Kirill Abramov, Shay Ohayon, and Jakob Voigts. Open Ephys: an open-source, plugin-based platform for multichannel electrophysiology. *Journal of Neural Engineering*, 14(4):045003, August 2017.