Split-trial analysis reveals the information capacity of neural population codes

Dylan Le[1,3,4,5] and Xue-Xin Wei[1,2,3,4,5,*]

1) Department of Neuroscience, The University of Texas at Austin
2) Department of Psychology, The University of Texas at Austin
3) Center for Perceptual Systems, The University of Texas at Austin
4) Center for Learning and Memory, The University of Texas at Austin
5) Center for Theoretical and Computational Neuroscience, The University of Texas at Austin

* Correspondence: weixx@utexas.edu

Abstract:

Understanding how correlated neural noise affects neural population coding represents one basic question in computational and systems neuroscience. Recent theoretical work suggest that shared noise along the stimulus encoding direction is the primary factor that limits information encoding (i.e., information-limiting noise). Despite this theoretical insight, it has been difficult to test experimentally due to the challenges in inferring information-limiting noise from neural data. To overcome this challenge, we have developed a method (i.e., split-trial analysis) based on partitioning the neural population response of an individual trial. Results from extensive numerical simulations suggest that split-trial analysis substantially outperforms existing methods in accuracy, efficiency, and robustness. Notably, it does so without estimating the noise covariance matrix, which represents a major barrier for prior studies. Applications of split-trial analysis to a number of neurophysiological datasets reveals insights into the precision of the neural codes for several systems. First, it reveals a substantial amount of information-limiting noise in the head direction system in mice. Second, it suggests a small yet positive information-limiting noise in the orientation code in mouse V1. Third, it reveals that the information-limiting noise in the macaque prefrontal cortex is highly consistent over time during a simple saccade task. Split-trial analysis is a general technique that should be widely applicable in analyzing the properties of population codes in the brain.

1 Introduction

Neural noise is correlated among neurons [1, 2, 3]. Noise correlations (NCs) can vary depending on task conditions [4, 5], and may inform network computations [6, 7, 8]. One important question has been whether the presence of NCs benefits or diminishes neural coding capacity [9, 2, 10]. Somewhat confusingly, earlier work suggests that NCs can increase or decrease coding capacity depending on the detailed structure [2, 3, 9, 11, 12, 13, 14, 15]. A unifying theoretical view begins to emerge from recent work [16], that is, the key limiting factor for coding capacity is the information-limiting NCs, which arise due to shared noise along the stimulus encoding direction (i.e., information-limiting noise).

Despite this important theoretical insight, it remains difficult to measure information-limiting noise experimentally, due to its potentially small magnitude and the presence of other forms of NCs [16]. While some recent studies reported substantial information-limiting noise in several neural systems [17, 18, 19, 20, 21], others have found the opposite [22]. Some existing analysis methods rely on empirical estimate of Fisher information (FI), then extrapolating it to infer the information limit for large populations [18, 20]. However, direct estimation of FI requires estimation of the noise covariance matrix, which is notoriously difficult given a limited number of trials, although dimensionality reduction has been proposed to address this problem under certain conditions [17]. An alternative approach is to decode the stimulus and then use the Cramer-Rao bound [23] to convert the magnitude of the decoding error into an estimate of Fisher information [16]. However, prior methods require a large number of neurons and trials, and thus are severely limited in practice.

To address this problem, we develop a method to infer the information-limiting noise based on splitting an individual trial into multiple sub-trials. This split-trial method creates multiple repeated measurements within a single trial so that one can dissect (i) the noise shared by all neurons (a proxy of the information-limiting noise) and (ii) the noise private to a subset of neurons. Our method is simple to implement, yet is highly effective in inferring information-limiting noise given limited data. Comparison to existing methods reveals that split-trial analysis substantially outperforms existing methods in its sample efficiency, accuracy, and robustness. Applications of split-trial analysis to neural datasets reveal a substantial amount of information-limiting noise in the mice head direction system, and a small amount of information-limiting noise in mice V1. It also reveals the temporal structure of information-limiting noise in macaque prefrontal cortex during a simple saccade task.

2 Results

2.1 A signature of information-limiting noise based on split-trial analysis

Despite the emphasis of classic work on its importance [3, 4, 24], the magnitude of NCs may not directly determine the information-capacity of a neural population code [16]. Using simulations based on standard neural population coding models (Fig. 1a), we first demonstrate the lack of a systematic relationship

between the magnitude of the NCs and the information-limiting noise. We will study four different neural population codes. We start by considering a population code with independent Poisson noise (Code 1). For this baseline model, as expected, NCs are centered around 0, with some variability due to sampling noise (Fig. 1b). We then consider a second population code by adding population-wide gain fluctuations to the model, i.e., the response gain of neurons increases or decreases together on a trial-by-trial basis (Code 2). As expected, this leads to positive NCs (Fig. 1c). Importantly, the information-limiting noise is 0 for both neural populations, or equivalently, both populations have unbounded information as the number of neurons becomes large. In sum, these two neural population codes share the same information capacity, yet the magnitudes of their NCs differ substantially. We further construct a third neural population code by adding a small amount of information-limiting noise, i.e., by injecting noise to the stimulus (Fig. 1d). This stimulus noise bounds the information-limit of this population code, however, it minimally affects the distribution of NCs. This suggests that two neural populations codes (e.g., Code 1 & Code 3) can have similar magnitudes of NCs, but distinct information capacity. Finally, we consider a fourth neural population code with both population gain and the same stimulus noise (Fig. 1e). For this population, the distribution of NCs resembles the population code with no stimulus noise (Code 2). However, these two codes have distinct information capacities. Overall, these results demonstrate that the magnitude of NCs alone are not predictive of the information limit of the neural code for large neuronal populations.

Next, we identify a novel signature of information-limiting noise based on splitting the neural population response, which we will refer to as *the split-trial analysis*. The key idea is to create repeated measurements within a single trial by splitting the neural population into non-overlapping subsets, and then decoding the stimulus based on the activity from each sub-set (Fig. 2a). Our first main finding is that the correlation of the decoding errors between the two splits provides a signature of information-limiting noise. Intuitively, if the decoders across the two subsets are perfectly correlated, the decoder constructed based on all neurons will behave identically to the decoder based on each subset, thus doubling neurons used in decoding would not change the estimate or reduce the decoding errors. This implies that the decoding errors come from information-limiting noise. In the other extreme, if the estimates from the different splits are independent, there is likely no information-limiting noise– if there were, such shared noise would lead to positive correlation in the decoding errors of the two splits.

We used simulations to demonstrate this signature of information-limiting noise in the context of the four neural population codes defined earlier. Fig. 2b,c,f,g show that, when information-limiting noise is absent, the correlation of the decoding errors is generally close to 0. In contrast, in the presence of non-zero information-limiting noise, the decoding errors are positively correlated (Fig. 2d,e,h,i). Thus, unlike the magnitude of NCs, the correlation of decoding errors indicates the presence of information-limiting noise.

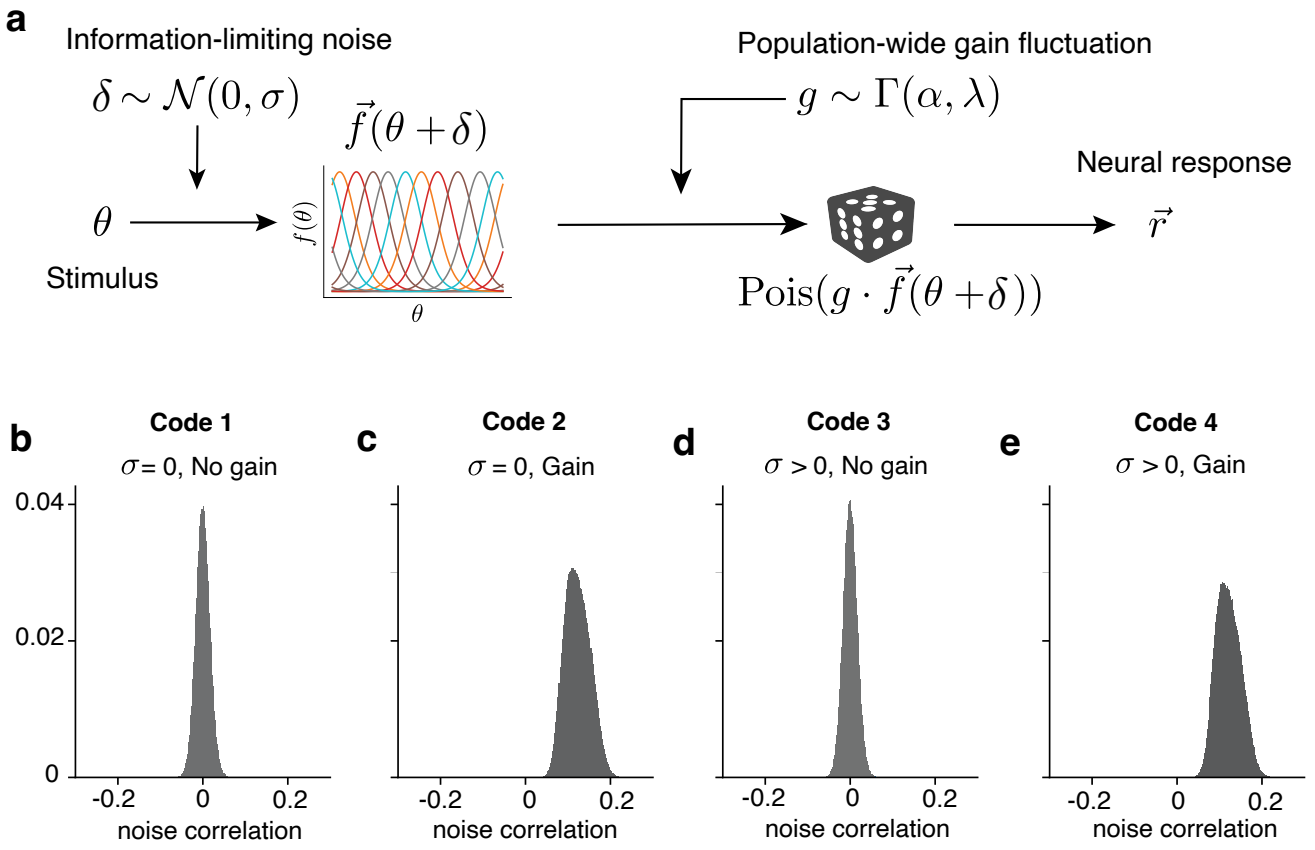

Figure 1: Simulation demonstrating the lack of correspondence between the magnitude of the noise correlations and the magnitude of the information-limiting noise. (a) Data generation pipeline for continuous estimation simulations. Stimuli are corrupted by gaussian noise to model information-limiting noise as input noise in the system. The corrupted stimuli are then used to estimate firing rates from the population (gaussian or von mises, depending on whether the stimulus variable was circular) tuning curves. After rates were determined, a population wide gain fluctuation was applied to the population on a trial by trial basis, adding a form of noise correlation that is not necessarily information limiting. The gain modulated rate is then used to generate spike counts with Poisson spiking statistics. **(b-e)** Distributions of pairwise Pearson's correlation coefficients (i.e., noise correlations) for neural populations generated under the scheme presented in **(a)** under different parameter regimes. Crucially, pairwise correlation coefficient distributions are not sufficient to distinguish between situations where the population has small ($\sigma = 5$ deg) information limiting noise from when the system is independent, as can be seen by comparing panels **b,c** to panels **d,e** respectively.

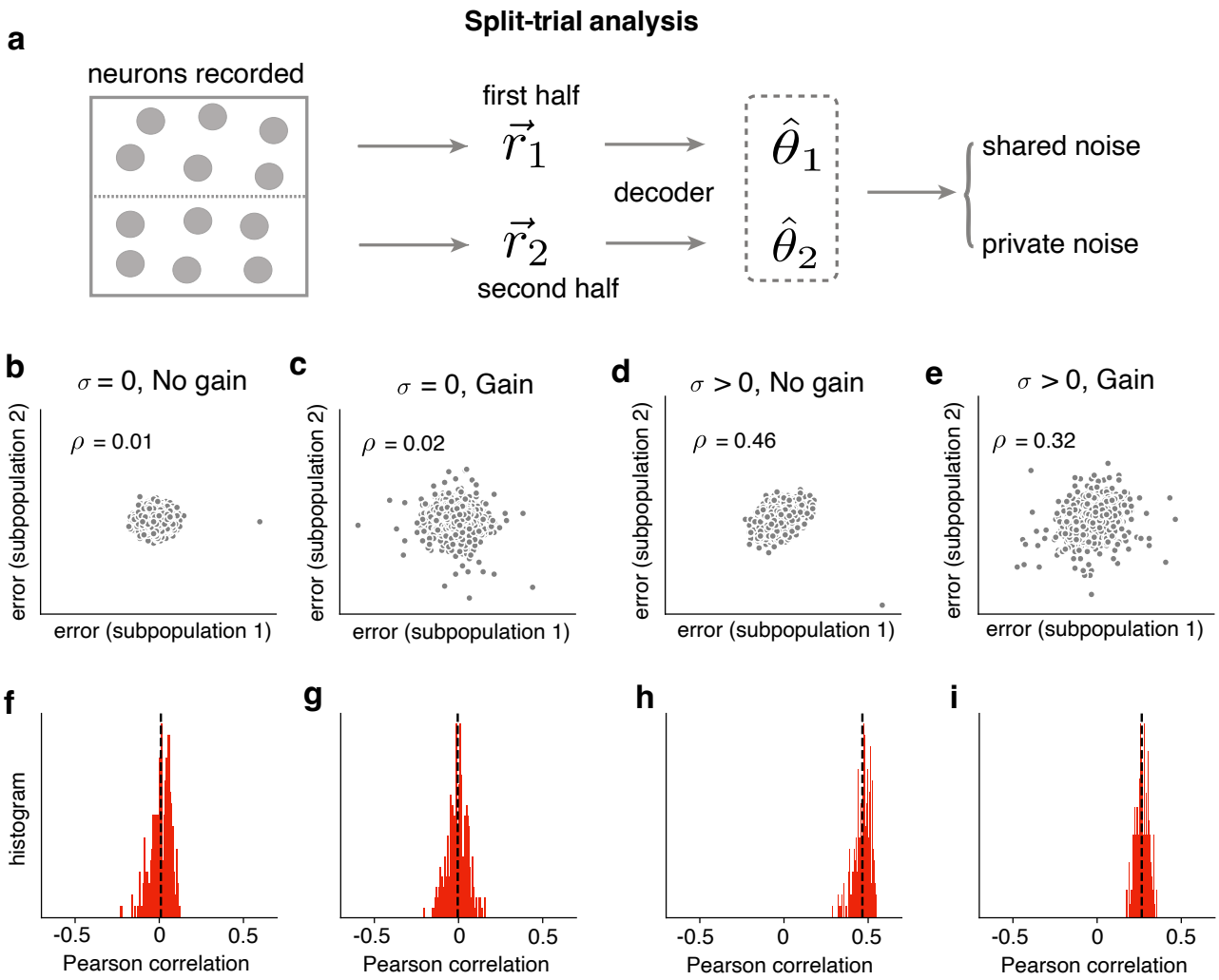

Figure 2: Split-trial analysis captures the information-limiting noise. (a) Overview of the split trial analysis. The neural population is first partitioned into two halves, which are both used to train individual decoders on the same set of trials. The decoding errors from both decoders are then leveraged against each other to separate the shared (information-limiting) noise from private noise (dependent on individual neuron noise). **(b)** Scatter plot of decoding error for two subpopulations of simulated neurons from an example simulation when the information-limiting noise is 0 and no gain fluctuations. (**c-e**) Similar to (**b**), but for other scenarios as indicated in the legend. Notably, when information-limiting noise is present ($\sigma = 5$ deg), clear correlations can be observed in the error distributions. **(f-i)** Histograms of Pearson correlation values between subpopulations across 500 different population splits for each condition, showing that the correlation between decoding error in subpopulations can be used to identify the presence of information-limiting noise correlations while being invariant under gain modulations.

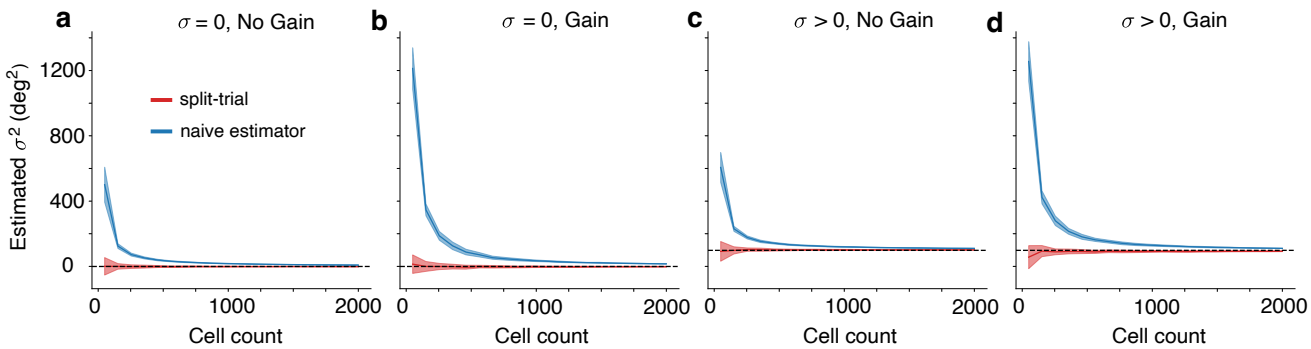

Figure 3: Benchmarking the performance of the split-trial analysis in estimating the variance of the information-limiting noise. Figures show the efficiency of the split trial technique as a function of total cell population size (blue), with a comparison against a naive estimator of information limiting noise (orange). The naive estimator consists of taking the variance of the decoding error as that quantity functions as an approximation of the cramer-rao bound of the estimator. The split trial analysis is capable of accurate estimation of the ground truth injected information limiting noise ($\sigma = 5$ deg, panels **c,d**) as well as detecting when there is no noise in the system ($\sigma = 0$, panels **a,b**) with relatively few cells simulated.

2.2 Accurate and reliable inference of information-limiting noise using the split-trial analysis

The correlation between the decoding errors of the two splits, while reflecting the presence of information-noise, is deemed to be bounded between -1 and 1. Thus, it cannot be used to directly estimate the information-limiting noise. However, we find that a simple modification to take into account the variability is all that is needed. Specifically, suppose that a stimulus variable θ is encoded in a large population of neurons from which we were able to measure the response of $2N$ neurons. We first split the neural population into two halves, then perform decoding analysis on each split (e.g., using Bayesian decoding). For a given stimulus θ, denote the decoded stimulus from the two splits as $\hat{\theta}_1$,$\hat{\theta}_2$, respectively. One of our main results is that the information-limiting noise can be estimated from their covariance, Cov$(\hat{\theta}_1, \hat{\theta}_2)$.

We now derive this estimator under a simple setting to provide some intuition behind this seemingly surprising result. First, notice that the decoded stimuli $\hat{\theta}_1$ and $\hat{\theta}_2$ are both random variables centered around the ground truth θ. Intuitively, their variability comes from two sources: (i) the information-limiting noise, denoted as δ; (ii) the private noise that is specific to that sub-population, denoted as ϕ_1, ϕ_2;. Thus, we can model the decoders for the two subsets using the following model: $\hat{\theta}_1 = \theta + \delta + \phi_1$; $\hat{\theta}_2 = \theta + \delta + \phi_2$. Under the assumption that the private decoding noise ϕ_1 and ϕ_2 are independent, it is straight-forward to show that

$$\text{Var}(\delta) = \text{Cov}(\hat{\theta}_1, \hat{\theta}_2).$$

Using simulated data, we compared the performance of our estimator with an existing method [16], for which the information-limiting noise is estimated based on a cross-validated decoder. Fig. 3 shows example results based on the four neural population codes studied above. These results suggest that, for cases where the information-limiting noise is 0, estimates from the split-trial analysis are generally centered around 0 (see Fig. 3a,b). Estimates from split-trial analysis can be smaller than 0 – this is a feature of our method and is important for an unbiased estimate of the variance of the information-limiting noise.

In contrast, the existing method estimates the noise to be larger than 0, which is an over-estimate of the information-limiting noise. Even in the regime of thousands of neurons, such over-estimation persists. When the information-limiting noise is non-zero (Fig. 3c,d), our method is able to accurately recover its magnitude. Again, the standard method over-estimates the magnitude of the information-limiting noise. Remarkably, the performance of the split-trial analysis with 100 neurons already outperforms the accuracy of the existing method with 1000 neurons, suggesting a dramatic improvement in its efficiency.

While the variance of the information-limiting noise is of interest, for some applications, it may also be useful to infer its entire distribution, based on which various statistics (such as median absolute error, standard deviation) can be extracted. This is particularly relevant when the distribution of information-limiting noise is heavy-tailed. We further developed a method to extract the entire distribution of information-limiting noise based on a deconvolution approach (see the Method section, and Fig. S1).

Generalization to binary classification In addition to continuous estimation tasks, information-limiting noise can also be studied in binary classification tasks [18, 17], a classical paradigm in psychophysics and neuroscience [25]. Suppose that we observe the neural activity corresponding to a pair of stimuli θ_1 and θ_2. One can ask how well one can discriminate the two stimuli (i.e., d-prime) from the response of a large neural population. The split-trial analysis can be generalized to this setting. The idea is to split the population into two sub-populations, and then apply a classifier to each sub-population to extract the decision variables for classification, DV_1, DV_2. These decision variables play similar roles as the decoders in the continuous case described earlier. We develop a procedure to use the covariance of the decision variables from the two splits to estimate the information-limiting noise.

We benchmark the performance of our method and three existing methods [16, 18, 17] for estimating the magnitude of the information-limiting noise from binary-classification tasks. The first baseline method [16] provides a naive estimate of the d-prime using linear classifiers with all neurons available. The second method [17] first performs dimensionality reduction, and then estimates the noise covariance, based on which the d-prime is estimated. The third method [18] directly estimates the generalized FI based on the noise covariance and then relies on extrapolation to infer the information-limiting noise for larger numbers of neurons. Based on extensive numerical simulations by varying the number of trials, neurons, and the ground truth d-prime values, we find that the split-trial analysis can more efficiently and reliably recover the ground truth. In particular, split-trial analysis substantially outperformed alternative methods under more challenging scenarios, *e.g.,* small information-limiting noise, limited amount of trials, and a relatively small number of neurons.

Fig. 4 shows example results from these benchmark experiments. More simulation results based on a range of other sample sizes are shown in Fig. S2. When neural variability is dominated by large information-limiting noise (Fig. 4a,d), estimating information-limiting noise is relatively easy. All four methods (including the naive method) perform well when d-prime is equal to 2. As the information-limiting noise becomes smaller, the problem becomes more difficult. For these more challenging cases, only our method can reliably infer the information-limiting noise with a sample of 500 per stimulus. The naive method and the method proposed in [17] substantially underestimate the magnitude of information-

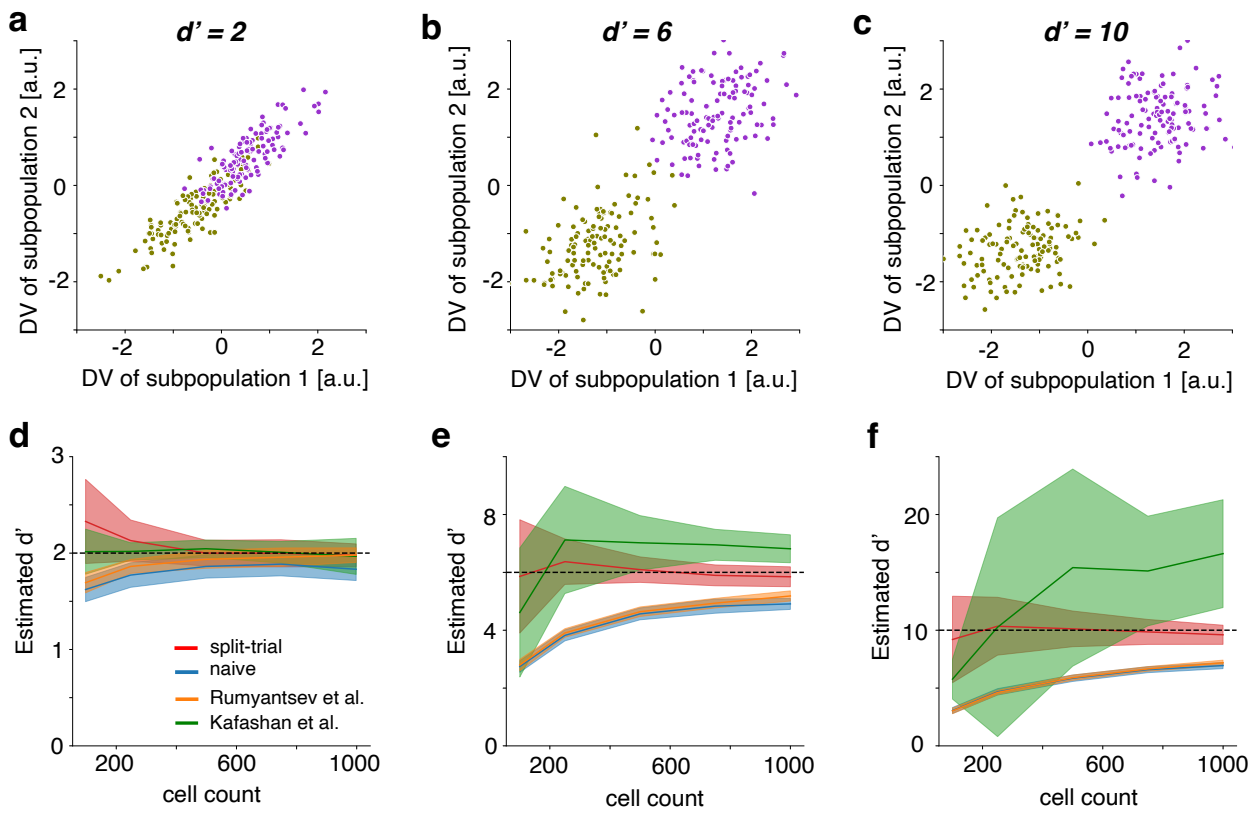

Figure 4: Benchmarking the performance of the split-trial analysis in recovering the ground truth d-prime in binary classification tasks. (a) Decision variable (DVs) scatters for classifiers trained on two separate subpopulations of simulated neural datasets when d-prime =2. The distributions for each class are shown in gold and purple, and it can be observed that in higher noise (low d') regimes, correlation between the two splits is more strongly observed. **(b,c)** Similar to **(a)**, but for d-prime =6, 10, respectively. **(d)** We estimated d' in simulated data using a variety of different approaches to estimate information limiting noise, using 500 samples/stimulus, when d-prime =2. **(e,f)** Similar to **(d)**, but for d-prime =4 and 6, respectively.

limiting noise (Fig. 4e,f). The method based on extrapolation by [18] exhibits substantial instability depending on the number of trials and the exact values of ground-truth d-prime values (Fig. 4e,f; also see additional results in Figs. S3-S5). For smaller sample sizes, split-trial analysis also under-estimates the d-prime values, but to a lesser extent compared to other methods.

2.3 Application 1: information-limiting noise in head direction system in mice

We first apply our method to study the head direction system [28], which encodes the brain's internal compass in multiple species [29, 30]. The head direction system has been modeled using continuous attractor networks [31, 32, 33]. This type of model predicts that neural activity forms an activity bump and that neural noise manifests itself as coherent shifts along the ring manifold, thus predicting the existence of information-limiting noise [16] in head direction system. To test this prediction, we re-analyzed a recently published calcium imaging dataset that simultaneously recorded 100-200 head direction cells simultaneously in free-moving mice [27]. Prior results based on these data showed that head direction

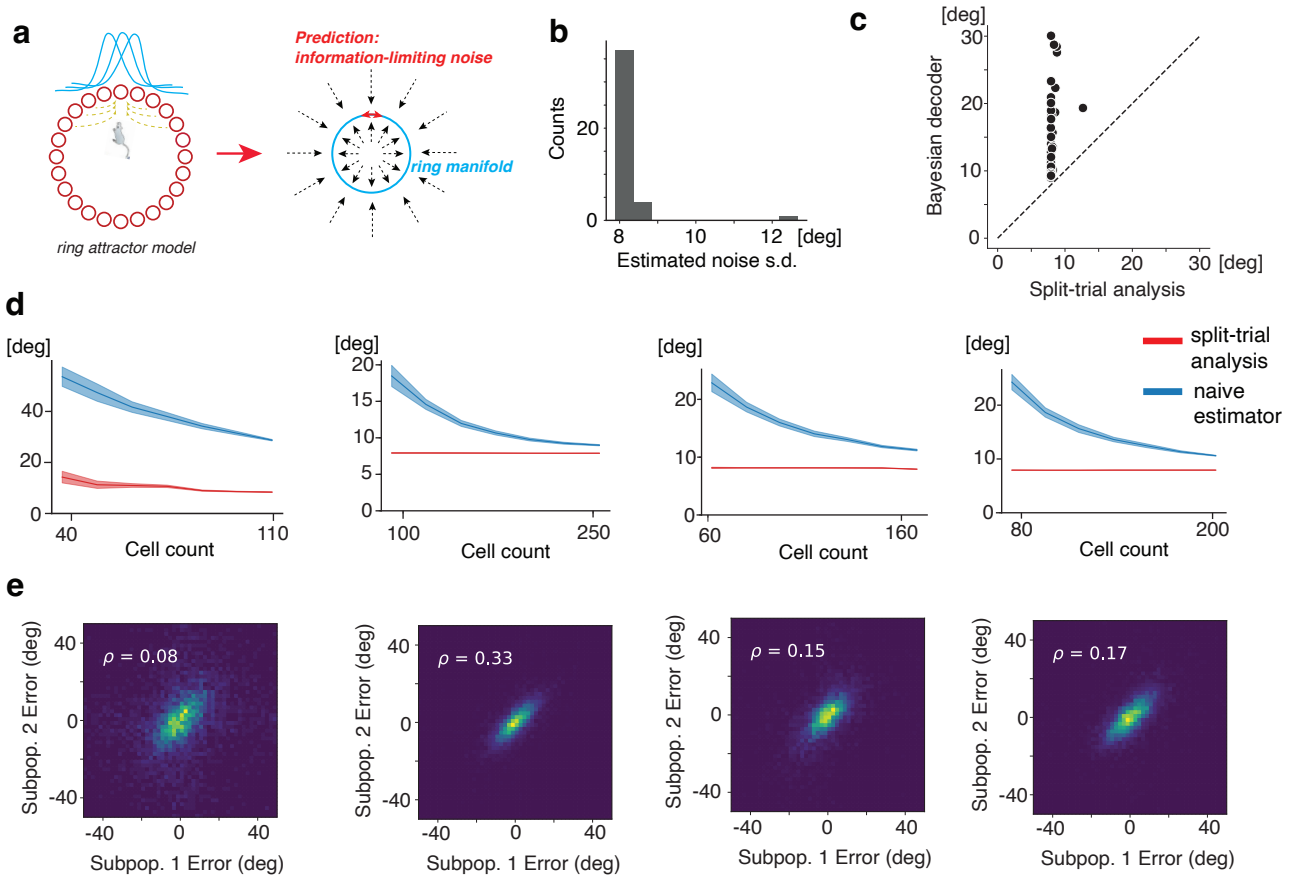

Figure 5: Results on head direction cells. Data reanalyzed from [26].**(a)** According to the ring attractor model commonly used to model head-direction cells, noise in the internal compass of the animal should manifest in a way identical to information limiting noise. **(b)** Histogram displaying the estimated information-limiting noise σ over 42 head direction population recording sessions from [27]. Decoding was done using a ZIG decoder as in the original work, and we estimate a mean information-limiting σ of ≈ 8.1 deg. **(c)** Plotting the estimated information limiting σ on the x-axis, and the estimate from a naive decoder (such as a Bayesian one) on the y-axis, we observe that our technique not only reports a generally lower information limit, but also that the variance of our estimator is substantially lower for this system. **(d)** Scaling in estimates of information-limiting σ as cell population size varies in four example sessions for both the split-trial analysis and a naive estimator. **(e)** 2D histograms of the error density across two subpopulations for a single iteration used to generate the plots in **(d)**, with Pearson's correlation value noted.

can be decoded with good precision [27]. However, it remains unclear whether the remaining decoding errors is due to information-limiting noise or non-information-limiting noise from limited numbers of neurons sampled. This question is challenging to address given existing techniques and sample sizes. Our method provides a unique opportunity to address this question.

Using a recently developed Bayesian decoding method for calcium imaging data [34], we first quantify how decoding errors vary with the number of neurons used in the decoding analysis by sub-sampling neurons. For most sessions, increasing the number of neurons used leads to a smaller error. Even when the number of neurons approaches the largest population size recorded in individual sessions, there does not appear to be a clear sign of saturation in the mean squared decoding error. These results suggest that inferring the magnitude of information-limiting noise using standard decoding analysis is difficult.

We then use the proposed split-trial analysis to analyze these data. We split the recorded neural population from each session into two halves, and perform decoding analysis on each of the two subpopulations. We find that the decoding errors between the two splits are only mildly correlated, which implies that the decoding error would be further reduced with a larger population size. Using the covariance of the decoding errors from the two splits to estimate the information-limiting noise, we find that the standard deviation of the estimated information-limiting noise is $\sim$ 8.15 deg (s.d. = 0.74 deg across sessions). Fig. 5b shows a scatter plot of the decoding errors inferred from the standard decoding approach and the information-limiting noise estimated from the split-trial analysis. The results show that, while there is substantial variability in the decoding error across different sessions, the inferred information-limiting noise is highly consistent across sessions. Furthermore, within individual sessions, we find that this value is also highly consistent when the number of neurons used is larger than 50. This observation further illustrates the robustness of our results. In addition, we used the deconvolution approach to directly estimate the full distribution of the information-limiting noise. We find that the estimated MAE is 4.4 deg, which is smaller than the MAE from a Bayesian decoding analysis (~ 6 deg; [26]) The magnitudes of inferred MAE from split-trial analysis are also highly consistent across sessions (s.d.=0.61 deg). Together, these results provide strong evidence for the existence of information-limiting noise in the head direction system. Such information-limiting noise is consistent with the predictions from continuous attractor models [31, 33, 16].

2.4 Application 2: The precision of orientation code in mouse V1

We next study a classic neural code, i.e., the orientation code in the primary visual cortex [35]. Recent studies have led to conflicting results regarding the existence as well as the magnitude of information-limiting noise in the orientation code of mouse V1 [17, 18, 22]. To further investigate this question, we apply split-trial analysis to the large-scale calcium imaging data reported in [22], which typically recorded more than 10,000 neurons simultaneously in individual sessions.

In one of the experiments conducted in [22], stimulus orientations were continuously sampled from the

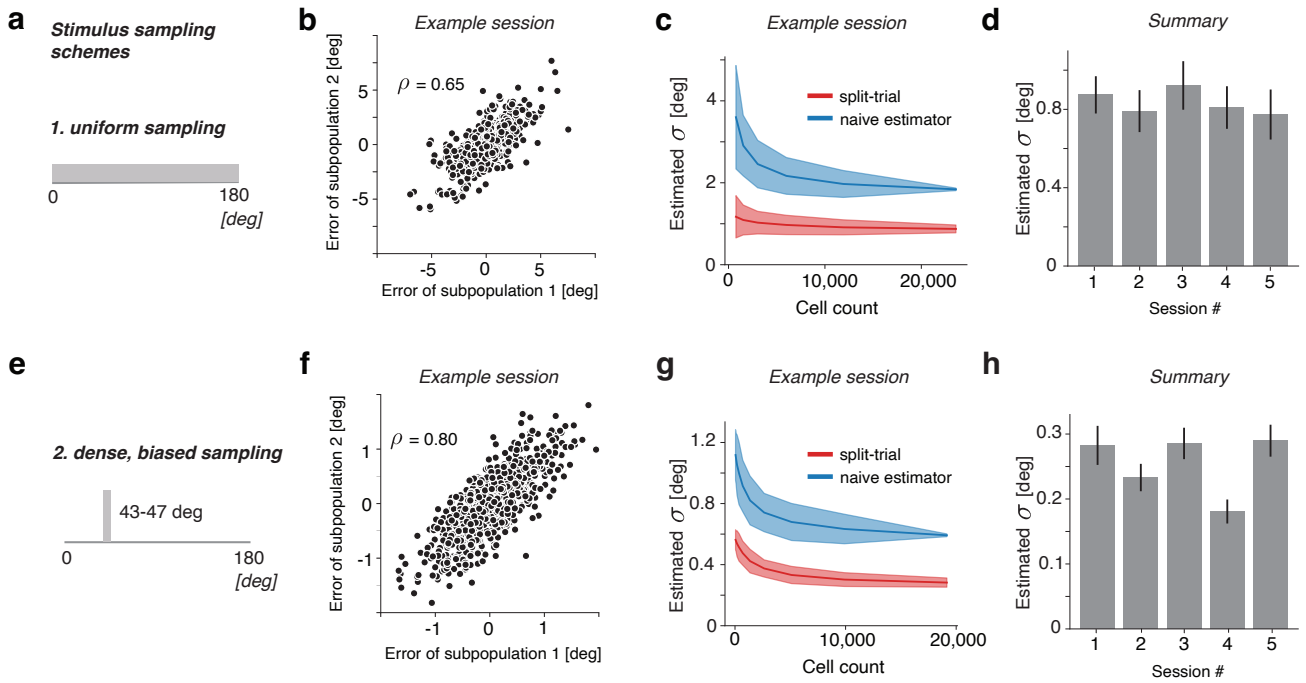

Figure 6: Results on the orientation code in mice V1 from two different experimental paradigms. Data reanalyzed from [22]. **(a)** Schematic of the uniform stimulus sampling method. **(b)** Example scatter plots of the subpopulation errors for sessions from the static orientation and dense discrimination datasets, respectively. The modest degree of correlation in both indicates the presence of information-limiting correlations in the considered datasets. Data from from [22]. **(c)** An example session's estimated information-limiting σ plotted against population size for static orientation and dense discrimination, respectively. **(d)** The estimated information-limiting σ for the static orientation and dense discrimination datasets, taken at the full population sizes for each session. One session was not visualized due to its large error bar. See Fig. S8 for results for all 6 individual sessions. **(e-g)**. Similar to **(a-d)**, but for another dataset that densely sample the stimuli from a small range, i.e., [43,47] deg.

whole circle. We first applied the standard decoding analysis to these data. Consistent with the report in [22], the mean squared error of the decoded orientation continuously decreases as the number of neurons used in decoding increases, and goes to about 2 deg when all neurons are used (typically 10k-20k neurons). However, it remains unclear whether the decoding error has saturated. We sought to address this question by applying the split-trial analysis to these data while varying the number of neurons used. We find that, using 10% of all neurons recorded in individual sessions, the inferred standard deviation of information-limiting noise is about 1 deg. The magnitude of the inferred information-limiting noise is largely stable when more neurons were included in the analysis. Using all neurons, the inferred information-limiting noise is only slightly reduced. The inferred value of the standard deviation of information-limiting noise is highly consistent with the 6 sessions (mean= 0.82 deg, sd =0.05 deg).

We further analyzed a second dataset collected in [22] using a different stimulus sampling paradigm. In this experiment, orientations of the gratings were densely sampled within a small range of 4 deg. Applying the split-trial analysis, the results suggest that the information-limiting noise is around 0.26 deg (sd= 0.04 deg). This is smaller than the magnitude of the decoding error inferred from the direct decoding approach in [22], but remains larger than 0. We notice that reliably inferring information-limiting noise is more challenging in this dataset, as evident in the slightly decreasing trend of inferred information-limiting noise beyond 5,000 neurons (Fig. 6e). This is likely due to the tiny amount of information-limiting noise in this data.

The results based on these two experiments provide evidence for the existence of a small amount of information-limiting noise in V1. Our inferred magnitudes of the information-limiting noise in mice V1 are substantially lower than those reported in [17, 18], which are generally larger than 5 deg. One reason for such differences might be the smaller cell counts used in [17, 18] (on the order $\sim$ 1000 neurons or less) together with the inaccuracy in the estimation of d-prime given the limited number of neurons. When we apply standard Bayesian decoding using cell counts that are comparable to [18] and [17] via subsampling the data from Stringer et al [22], the estimated s.d. of information-limiting noise is generally around 4 deg (Fig. S8), which is closer to the values reported in [18] and [17]. Another contributing factor could be the different paradigms used in estimating information-limiting noise. In particular, the d-prime values computed based on discriminating a pair of well separated stimuli may be difficult to link to the s.d. of information-limiting noise from continuous estimation tasks when the neural response manifold is highly curved [36].

2.5 Application 3: Temporal integration of stimulus information in frontal cortex

Neural computation involves the integration of information both over neurons and over time. As most previous studies examined correlated neural noise focused on the effect of integrating information over neurons in a fixed time window [9, 16], the degree to which temporal integration affects information encoding remains an important open question [37]. While noise correlations can be computed for neural activity across time, it remains unclear how they affect temporal integration for large neural populations. For simplicity, theoretical studies of neural population coding typically assume independent noise for different

time windows, which implies a large benefit of integrating over a longer period. Empirical studies have examined the effect of temporal integration for individual neurons, and found that integration over several hundred milliseconds remains beneficial [38, 39]. However, results obtained from individual neurons or small neural populations may not generalize to large neural populations. For a large neural population, the impact of temporal integration on the encoded information should mainly depend on the temporal dynamics of information-limiting noise. Consider the extreme case that information-limiting noise for two adjacent time windows is perfect correlated. In this case, integrating over these two time windows would lead to no improvement in information encoding. Conversely, if information-limiting noise at two adjacent timesteps is uncorrelated, temporal integration would substantially increase the information. So far, the temporal structure of information-limiting noise remains an open question.

We investigate this question in the context of monkey prefrontal cortex (PFC) when the monkeys were performing a simple left/right saccade task [20]. Briefly, on each trial (Fig. 7a), the monkey was tasked with fixating on a point in the center of a screen for a variable fixation period (400-800ms), after which the fixation point would disappear and a stimulus point would appear in either the animals' left or right visual field. The animal would then be required to fixate for 500ms on the stimulus point, before a reward was stochastically delivered on correct trials. Up to 800 neurons in PFC were recorded simultaneously in individual during the experiments (n=4 sessions from two monkeys). We focus on neural activity during the 300ms before the reward period. Performing standard decoding analysis on all neurons recorded, we find that the d-prime values are close to 7 (mean =6.82, s.d.= 0.8), consistent with the original report in [20]. Examining the correlation of the decision variables across the two splits, we find that they are only modestly correlated, which implies that even larger populations may lead to even larger d-prime values. Applying the split-trial analysis to the neural population response in the 300ms window defined above, we find that the inferred d-prime values for the four experimental sessions are 13.3, 13.5, 10.2, and 11.2, respectively. These values are substantially larger than those from the standard binary classification analysis. Meanwhile, they are substantially smaller than the values inferred in [20] using a method based on extrapolation. The inferred d-prime values are consistent for the two sessions of individual monkeys (Fig. 7c), and monkey V has larger d-prime values compared to monkey W.

We next turn to our key question of how correlated noise affects information during temporal integration. We first applied the standard decoding analysis on neural populations of different sizes, while varying the time-window of the decoding. We tested a number of time windows, ranging from 300ms to 20ms with a 20ms decrement (Fig. 7b). We find that for small neural populations (e.g., n=25 cells), d-prime increases substantially with increasing time window. When n=25, the d-prime values obtained at 300ms are on average 2.3 times larger compared to those from 20ms. As the number of neurons increases, the benefit of temporal integration reduces. is needed when all recorded neurons are considered. Applying the split-trial analysis to all neurons recorded in individual sessions, we find that there is a benefit of temporal integration below 40ms. However, beyond 40ms, there is little improvement in the information. On average, the d-prime values based on a 40ms window are almost as large as those obtained from the 300ms window (Fig. 7e).

Because the analysis above involves a gradually changing time window, one possible explanation is

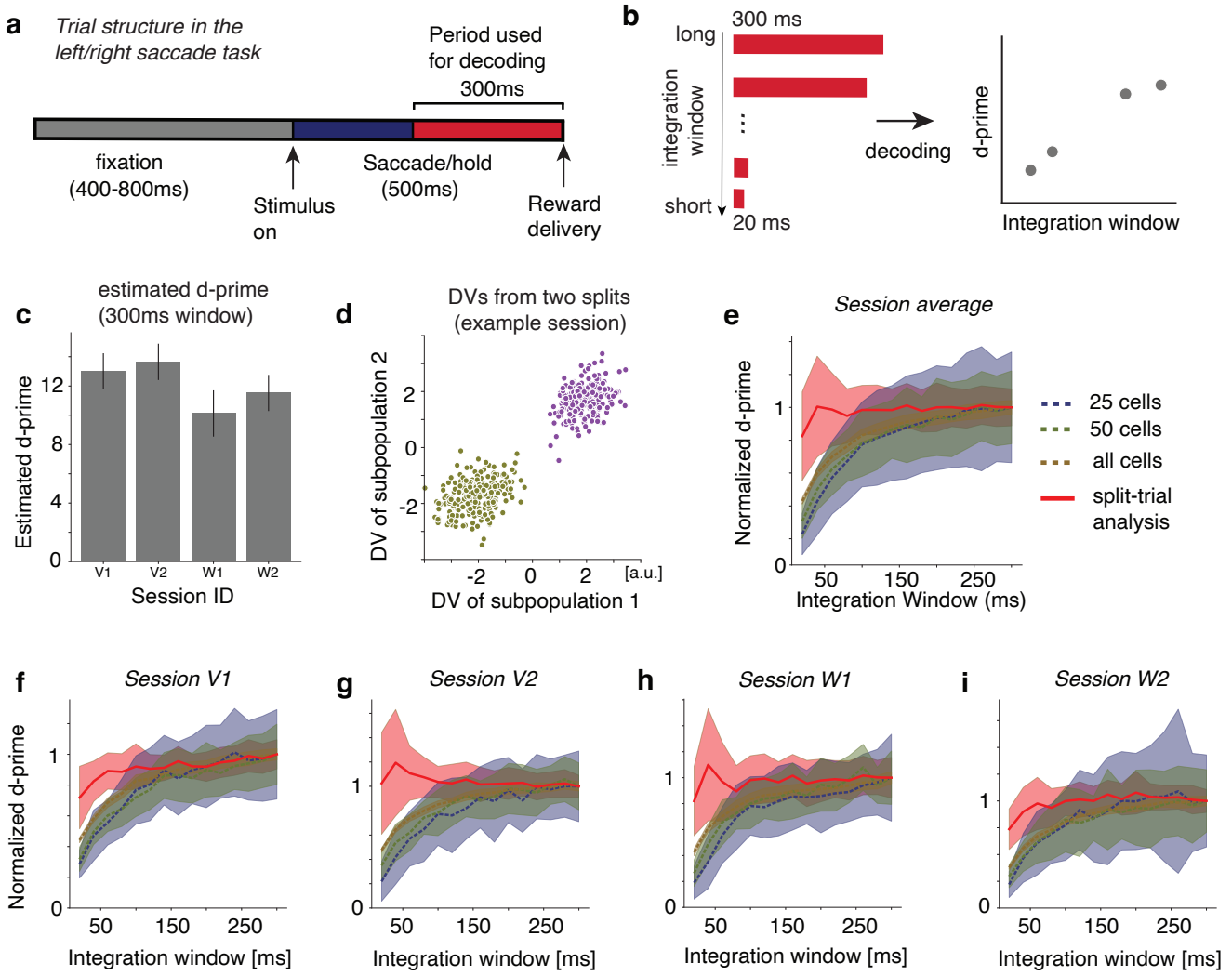

Figure 7: Results on the neural code in macaque PFC during a left/right saccade task. Data reanalyzed from [20]. **(a)** Schematic of the structure of an individual trial. **(b)** Schematic of the analysis procedure. We gradually decrease the length of the time window used for the analysis. For each time window, we inferred the d-prime value based on split trial analysis and the standard classification analysis (naive approach). **(c)** Estimated d' limit for PFC recording sessions from Bartolo et al., dataset consisted of two animals (labeled 'V' and 'W') across recording sessions for each animal. **(d)** Example DV scatter for a trial from session V1. **(e)** d-prime values in the limit of infinite number of neurons was estimated using both the split trial analysis (solid lines) and the naive approach (dashed lines) for increasingly large spike integration windows for the time period starting 300ms before reward delivery on successful trials. Each plot shows change in d' limit as integration window is widened across different total population sizes (colors in legends), normalized by the mean value at the widest integration window. The curves in this panel indicate that, in this dataset, longer integration times do not substantively change the split trial method's estimation of the information limit of the system in this task, as is the case for the naive approach. This indicates a strong and stable signal of information can be recovered in PFC with as little as 20ms spike data from the split analysis approach. **(f-i)** Individual plots for each session, in the same manner as panel **(e)**.As was the case in panel **(e)**, curves displayed estimated d' using 25, 50, and all cells recorded in the session.

that the saturation of information with larger time windows is due to the later period containing little or no information about the task variable. We perform additional analysis to rule out this possibility. Specifically, we use split-trial analysis to infer the d-prime values from each of the fifteen 20ms time windows during the 300ms period we analyzed. We find that there is still substantial amount of information in the later periods, although the information decreases modestly (Fig. S13). Together, these results suggest that, for this simple saccade task, information-limiting noise in PFC for the saccade location is highly correlated over time within individual trials. Given a sufficiently large neural population, each of the 20ms time window contains highly redundant information. One potential implication of this finding is that, for this simple behavioral task, if the downstream areas have access to a large population of neurons, integrating the response from PFC over a brief time period is sufficient to transmit most information about the task variable.

3 Discussion

We developed a split-trial analysis method to infer the magnitude of information-limiting noise (or equivalently, the information capacity) for large neural populations. This procedure does not involve computing NCs. Instead, it splits a neural population into two subsets, and then project the neural responses back to the stimulus axis via decoding. Split-trial analysis creates repeated measurements within a single trial, and it enables the partition of the noise of the neural population (captured by the decoding errors) into shared noise across the two splits and private noise specific to each split. Our results show that the estimated shared noise across two splits provides a surprisingly simple yet effective estimator of the information-limiting noise. Simulation results suggest that split-trial analysis outperforms previous methods in terms of both the number of neurons and the number of trials required to reliably estimate the information-limiting noise. While in this paper we have focused on information-limiting noise, the method can also be used to estimate the contribution of non-information limiting noise as a by-product for a finite neural population, which can be estimated by taking the difference between information-limiting noise and total noise.

The idea of splitting the population into subsets and performing decoding on each subset was mentioned in [22]. However, it was not recognized that it could be used as an effective way to infer information-limiting noise. Previously, it was proposed that the decoding errors of the two subpopulations should be generally correlated in the presence of correlated neural noise [22]. However, we have demonstrated that this is not the case (Fig. 1). For example, while coherent multiplicative gain fluctuations lead to large positive correlations in the neural population, the decoding errors from the two subpopulations are generally not correlated. Importantly, we find that the correlation of the decoding errors across the two splits serves as a signature of the information-limiting noise, rather than arbitrary forms of NCs. This is a critical insight that was not appreciated previously.

Split-trial analysis has several advantages compared to existing methods. First, it is simple and does not involve the estimation of noise covariance, which is a major hurdle in analyzing a large neural population.

Although the development of recording technology may enable the recording of even larger neuron populations in the future, the number of trials required to estimate the noise covariance matrix would remain a problem. Second, split-trial analysis does not suffer from severe estimation biases of previous methods. The standard method [16] based on directly decoding the stimulus overestimates the information limiting noise for a finite neural population. Our numerical results show that similar concerns apply to the method in [17], in particular when the information-limiting noise is small. Furthermore, our results show that a method based on extrapolation [18]), while being well motivated, is unstable in practice. Split-trial analysis method is substantially more robust than these prior methods, and remains generally stable for different numbers of trials and neurons. One exception is that, when only a small number of trials are available to infer a large d-prime, split-trial analysis exhibits an under-estimation. Interestingly, we did not observe such biases for continuous estimation tasks. It is possible that a refined version of split-trial analysis may remove these biases when estimating d-prime in binary classification tasks.

Application of our methods to neurophysiological data reveals several new insights. First, the results revealed a substantial amount of information-limiting noise (s.d. $\sim$ 8 deg in standard deviation) in the mouse head direction system. The magnitude of this information-limiting noise is highly consistent in three mice we analyzed. This result is consistent with attractor dynamics, and provides an example of the existence of substantial information-limiting noise in spatial navigation system, in additional to recent work on place cells [21]. It remains to be tested whether head direction systems in other species exhibit similar amounts of information-limiting noise and whether the information-limiting noise depends on the experimental conditions.

The magnitude of the information-limiting noise inferred from mice V1 data [22] is substantially smaller. When the stimuli were sampled continuously over orientation, the standard deviation of the information-limiting noise is slightly smaller than 1 deg. For the densely sampled dataset, the information-limiting noise is around 0.2 deg, yet remains larger than 0 deg. These values are substantially lower than those reported in two other recent studies [17, 18]. Although the exact reasons of this discrepancy remains unclear, we consider that the larger information-limiting noise inferred in [17, 18] could potentially due to a combination of the following: (i) a smaller number of neurons recorded and used in the analysis (10k v.s 1k); (ii) coarse sampling of the stimulus in their experimental design; (iii) an overestimation in the estimated information-limiting noise. Notably, we found evidence for a difference in the magnitude of the information-limiting noise when comparing the two experimental conditions in [22]. [22] considered that the difference in the decoding errors was due to the density of the trials sampled– an effect we also confirmed when downsampling the dense dataset. This difference may not be fully explained by the number of trials. When we subsample the second dataset (dense sampling) to match the number of samples in the first dataset (uniform sampling), we find that the magnitude of the inferred information-limiting noise is about 0.6 deg (Fig. S10), which remains generally smaller than the values inferred from the first experimental condition with uniform sampling of the stimuli. This raises the possibility that information-limiting noise may depend on the task conditions, such as stimulus statistics (i.e., sensory adaptation) and paradigms (i.e., binary classification v.s. continuous estimation).

Our method opens the door for studying the temporal dynamics of information-limiting noise. This was

a difficult question to study due to the entanglement of information-limiting noise and private noise in neural responses. Applying split-trial analysis to study monkey PFC data collected during a saccade task shows that temporal integration in PFC has little effect beyond 40 ms when a large population is involved, unlike what was reported for the cases of individual neurons or small neural populations [39]. This suggests a large temporal correlation of the information-limiting noise, i.e., information-limiting are highly correlated over time. While our analysis approach is general, it is possible that this result may be specific of this simple task. It would be interesting to apply our method to study whether temporal integration over longer timescales carries more substantial benefits in other decision tasks and other brain regions. In particular, for tasks that would benefit from longer integration and deliberation [40, 41], information-limiting noise may be less correlated over time.

Further extensions of our methods should be possible. First, we have focused on split the population into halves due to its simplicity. It would be interesting to consider a k-way split (k>2) of an individual trial. By considering the noise shared in all k splits, one might be able to get more accurate estimate of the information-limiting noise. Conceivably, one potential drawback of having too many splits would be that the decoding error for each split would be too noisy. Finding the optimal number of splits is an interesting question for future research. Second, while here we estimated the information-limiting noise for uni-dimensional variables, our method can be generalized to multivariate stimulus variables. One strategy is to estimate the information-limiting noise for one variable at a time, and then combine the noise from individual dimensions. Alternatively, one may perform joint decoding of multiple variables at the same time, then extract the shared component in the multivariate stimulus space.

One important yet under-investigated question is how information-limiting noise depends on the task conditions, e.g., attentional states [4] or learning [42]. Having multiple task conditions in a single experiment necessarily means fewer trials for each condition. The large number of trials required for inferring information-limiting noise under a given condition severely has limited the ability to task-dependence of information-limiting noise [43] in previous studies. With split-trial analysis, the ability to infer information-limiting noise reliably with much fewer trials should open up exciting opportunities for future experiments that investigate how network computations affect the capacity of the neural code and and temporal dynamics of information-limiting noise in neural circuits.

4 Methods

Simulations

For simulations, the tuning curve of an individual neuron for stimulus variable θ was assumed to be Gaussian (or von Mises function). For neuron i, the tuning curve is denoted as $f_i(\theta)$, which is centered at the neuron's preferred stimulus $\theta_{pref,i}$. The number of cells and tuning width were generally set for full coverage of the stimulus space. Neural spiking statistics were determined by the tasks that were being simulated (see below). Stimuli were sampled continuously over the stimulus range.

Continuous Estimation: When simulating neural activity responses for a continuous estimation task, the response was generated according to the folowing model

$$f(\theta) = \frac{A\exp(\kappa\cos(\theta - \theta_{pref}))}{2\pi I_0(\kappa)} \tag{1}$$

$$\tilde{f}(\theta) = gf(\theta + \delta) \tag{2}$$

$$r_i(\theta) \sim \text{Pois}(\tilde{f}_i(\theta)). \tag{3}$$

The neural tuning functions were modeled as von mises distributions centered at their preferred stimulus with parameter κ determining tuning curve width, parameter A controlling peak rate, and parameter $I_0(\kappa)$, the modified Bessel function of order 0, is essentially a normalizing factor that can be omitted for purposes of modeling neural tuning. The information-limiting noise was modeled as a shared input noise term $\delta \sim \mathcal{N}(0, \sigma)$ which corrupted the input stimulus shared across the entire neural population (i.e. δ is the same for each neuron). For some simulations, a term $g \sim \Gamma(\alpha, 1/\alpha)$ is introduced to model the trial-by-trial gain fluctuation.

Discrete Stimuli & Binary Classification: When generating simulated data for binary classification tasks, the neural response was generated according to the following expressions:

$$\vec{r}(\theta_i) \sim \mathcal{N}(f(\vec{\theta}_i), \Sigma) \tag{4}$$

$$\Sigma = \Sigma_0 + \epsilon \vec{f}' \vec{f}'^T \tag{5}$$

$$\vec{f}' = \frac{\vec{f}(\theta_1) - \vec{f}(\theta_2)}{\Delta\theta}. \tag{6}$$

Matrix Σ_0 is a covariance matrix for the population with no information-limiting component (i.e. a minimal projection along the $\vec{f}'$ direction). This matrix was generated by constructing a random positive semi-definite matrix via $\Sigma_0 = UDU^T$, where U is a randomly drawn orthogonal matrix (in our case the *ortho_group* function from the scipy.stats package in Python), and D is a diagonal matrix determining the eigenvalues of Σ_0. The entries of D were drawn from an exponential distribution with rate parameter one, but the largest values of D were multiplied by a factor of up to 5 to control for the level of noise correlations present in the neural population. Constructing Σ_0 in this way enables one to manipulate non-information-limiting noise in the system directly by changing columns of U and entries in D. The parameter ϵ is the magnitude of the information limiting correlations present in the population.

General data analysis pipeline

Continuous stimuli

For neural data (simulated or empirical) which came as a result presentation of continuous-valued stimuli, we used a naive Bayesian decoder to estimate the stimulus presented for each trial based on the neural population response [44], assuming neurons are noise independent. The decoded stimuli were compared to the ground truth to obtain the error for each trial. The decoding errors were used to perform the split-trial analysis.

Discrete stimuli

In the case of data generated from a discrete stimulus space, as in classical binary classification tasks, the neural response was used to train an linear support vector machine (SVM), although any well-trained classifier might suffice. Typically 70% of the data was held out for training and a 5-fold cross-validation procedure was used on the training data to obtain optimal model parameters. For the test data, rather than using the choice of decoded stimulus, we obtained the decision variables by taking the (1-dimensional) distance of each trial to the decision boundary from the trained SVM. These decision variables were then used to perform the split-trial analysis.

Split-trial analysis

Split-trial analysis begins by partitioning our total population of N neurons into two subsets of $N/2$ distinct cells each. Then we train decoders (or classifiers) on the data from each subset. Let $X_i = \hat{\theta}_i - \theta = \delta + \phi_i$ be the decoding error for both subpopulations. We assume that our decoders $\hat{\theta}_1$ and $\hat{\theta}_1$ are unbiased, i.e., $E[\phi_1] = E[\phi_2] = 0$. We will further assume that δ and the ϕ_i are all independent. Intuitively, δ describes the shared (information limiting noise) term, and the ϕ_i's describe the private noise unique to the individual subpopulations. In fact, the random variables do not need to be gaussian as the following argument only requires they be mean zero, finite variance, and that they be independent, the assumption of Gaussian distribution was simply chosen for notational convenience. From this decomposition of the error we can calculate: $X_1 + X_2 = \phi_1 + \phi_2 + 2\delta; X_1 - X_2 = \phi_1 - \phi_2$. Because ϕ_1, ϕ_2 and δ are independent, $Var(X_1 + X_2) = Var(\phi_1) + Var(\phi_2) + 4Var(\delta)$, and $Var(X_1 - X_2) = Var(\phi_1) + Var(\phi_2)$. Thus, the variance of the shared noise (denoted as σ^2) can be calculated as $\sigma^2 = Var(\delta) = \frac{1}{4}(Var(X_1 + X_2) - Var(X_1 - X_2)) = Cov(\phi_1, \phi_2)$. This result suggests that one can use $Cov(\phi_1, \phi_2)$ to estimate the variance of information-limiting noise.

For a single iteration of the split-trial analysis, the neural population is randomly divided into two halves by assigning each neuron to one of two subpopulations. Then the decoding error will be considered X_1, X_2 for the first and second subpopulations, respectively. We then calculate $Cov(X_1, X_2)$ as the estimator for the variance of the information-limiting noise obtained for a single iteration. Multiple iterations of split-trial analysis can be performed to get a distribution of the estimated variance of the information-limiting noise, from which the mean and the confidence interval of the estimated variance (or the standard deviation) of

the information-limiting noise can be obtained.

Discrete extension of split-trial analysis

We develop an algorithm to generalize split-trial analysis to the case of estimating d-prime values for binary classification tasks. Previous methods [17, 18] have also been developed to infer information-limiting noise based on binary classification. In this case, the decoding error $X_i = \hat{\theta}_i - \theta_i$ is no longer available. We find that the decision variables/logits (DVs) returned by the linear classifier for individual trials can be used similarly to decoded stimuli in the continuous estimation case. In this setting, the DVs are a measure of a given trials (signed) distance to the decision boundary, with sign determined by class label. According to the standard one-dimensional signal detection theory, one assumes that the decision variable for choosing the stimulus i follows a Gaussian distribution, $DV_i \sim \mathcal{N}(\mu_i, \xi)$, where $i = 1, 2$ for each subpopulation. Under this standard setting, d-prime can be defined as $\frac{|\mu_1 - \mu_2|}{\sqrt{\xi}}$. If these decision variables were directly extracted from a linear classifier, ξ would contain contribution from both information-limiting noise and non-information-limiting noise. The key idea is to use split-trial analysis to isolate the variability of the information-limiting noise.

Specifically, when extended to consider two subpopulations, we have twice as many distributions, but each DV can fall into only two distinct distributions. We then take the same decomposition as before to describe $DV_i = \mu_i + \phi_i + \delta$, for a given subpopulation and trial, and a similar calculation as described previously enables us to leverage shared variability in both neural population's DVs to estimate the shared information limiting noise in the decision space, σ. We first take the difference between DV_i and the (binarized) label for the ground truth stimulus class for the trial (typically ± 1). From there we can use the 'error' directly as we do in the continuous approach to obtain σ_i, for subpopulation i, then $\sigma = \frac{\sigma_1 + \sigma_2}{2}$. The information-limiting discriminability index (d') is then calculated: $\frac{|\mu_1 - \mu_2|}{\sigma}$.

In practice, μ_i is the empirical mean of the DV's for each stimulus, e.g. $\mu_1 = \frac{1}{T}\left(\sum_{t \in \text{stim 1}}^{T} DV_{1,t}, \sum_{t \in \text{stim 1}}^{T} DV_{2,t}\right)$. Since the only quantity of interest is ultimately the d', the class labels can be chosen somewhat freely in this framework, as the standard deviations of the DVs, ξ will scale with the selected labels, as the classifier is trained.

Estimating the distribution of information limiting noise using a deconvolution approach

The distribution of the information-limiting noise may be non-Gaussian, in which case variance alone may be insufficient to describe the properties of information-limiting noise. We develop an approach that directly estimated the information-limiting error distribution. Let $X_1 = \delta + \phi_1, X_2 = \delta + \phi_2$ be the decoding error for the first split and the second split, respectively. Assume that ϕ_1 and ϕ_2 are independent, and furthermore their density functions are symmetric. It follows that $X_1 + X_2 = 2 * \delta + \phi_1 + \phi_2$, and $X_1 - X_2 = \phi_1 - \phi_2$. Because ϕ_1 and ϕ_1 are independent and symmetric, $\phi_1 - \phi_2$ and $\phi_1 + \phi_2$ follows the same distribution. Therefore, the probability density of $X_1 + X_2$ is the result of convolving the density

of $X_1 - X_2$ and the density of 2δ. This suggests that density of 2δ can be estimated by solving a deconvolution problem [45]. Note that, for this deconvolution approach, it is important that the density of the decoding error is symmetric. This assumption of symmetric error distribution is not required when using the covariance of the decoders from the two splits to estimate the variance of information-limiting noise.

In our implementation, we assume that the density of 2δ has the form of $g = ae^{|x|^b}$, which includes Gaussian density function as an special case. The problem is to find a filter g for which $g * (X_1 - X_2) = X_1 + X_2$. To find the best fit for the filter g we did a grid search over a and b and fit with least squares, using the empirical distribution $X_1 + X_2$ and $X_1 - X_2$ computed based on the decoding errors of the two splits. The range of exponents, b, and the range of widths, a, were continuously sampled in the range $(0, 3]$, and $(0, 5]$, respectively. After finding the best filter, we treated it as a PDF for the information-limiting noise distribution and extracted the standard deviation σ and the median absolute deviation to characterize the noise for the dataset.

Dataset specific details

Ca^{2+} Imaging Dataset: In this study [22], calcium imaging was done in V1 for awake, head-fixed mice while they observed static oriented gratings whose orientations continuously spanned the circle. Analysis of the dataset followed the procedure laid out in the original author's report. In short, we performed decoding using the neural activity collected while animals observed a collection of continuously distributed oriented static gratings. This decoding was done using their so-called 'super-neuron' based decoding procedure for both subpopulations. 48 super-neurons were spaced evenly across the range $[0, 360]$ deg. Information limits were estimated using the deconvolution variant of the split analysis described below, as well as with the naive decoders trained from the original report from [22].

Head Direction Cell Dataset: In this dataset, calcium imaging was performed in the anterodorsal thalamic nucleus (ADN) of mice. The animals were kept in a chamber with a 360 deg LED screen. Vertical stripes were displayed at different regions of this screen during the baseline recording session, which is the period of the session that our analysis focused on. Our analysis pipeline for the head direction cell data from [27] also followed the original approach with small tweaks to fit our splitting procedure. The recorded responses from the head direction cells of the animal while it tracked the rotation of a moving landmark were fed into a ZIG decoder [34], as was done in the original paper by [27]. In fact, for each iteration, we trained two ZIG decoders, based on the activity of either of our two population splits. The head direction errors from both decoders for each trial were then used for split-trial analysis as previously described.

Macaque PFC Dataset: For the dataset from [20], animals were trained to fixate to a target on the center of a screen for as long as the fixation point stays on the screen (400-800ms). A stimulus would then appear on either the right or left of the animal's visual field and the animal was expected to saccade to the point and hold its gaze for 500ms on the stimulus, after which a reward would be delivered 70%

of the time. For our analysis, we trained an SVM on the neural activity in PFC from the first 300ms of the 500ms period preceding reward delivery or trial cessation. More specifically, we took the time a trial ended (due to either reward delivery or timing out) and considered the 500ms preceding that point to be the period when the decision was being encoded in the PFC. We then trained our SVMs on the first 300ms of that 500ms interval. The reason for this delineation is that it was often the case that animals could fixate on the stimulus for a period of time less than 500ms before saccading back to the stimulus and holding gaze for the required time. In parsing the data this way were were able to find a stable point of comparison across all the experimental trials. From the SVMs we recovered the decision variables as the linearized distance between the trials and the separating hyperplane. In this case, the 'errors' used for the split-trial analysis are the differences between the decision variables and the class labels which were set at values -1 and 1.

In order to assess the dynamics underlying information integration in PFC for this task, we also performed the split-trial analysis using data with increasingly wide time windows from the same starting time point. In this case we started with a 20ms wide window, estimating the information-limiting noise from the first 20ms of activity from our starting point. The window was increased by increments of 20ms up to the full 300ms wide window of activity used in the previously described analysis.

Implementation of alternative methods

The PLS reduction technique used in [17] was reimplemented following the description from the original paper and from their own original code. In short, neural response data was split by trial into thirds, where one third of trials were used to train a PLS Regression model, one third ($r_{transform}$) was used to calculate the optimal weight, $W_{opt} = \Delta\mu^T_{transform} \cdot \Sigma_{transform}$, and the final estimated d' was calculated by $W_{opt} \cdot \Delta\mu_{test}$ and taking the square root of the result.

The technique from [18] was reimplemented in Python from their original Matlab code. Neural population responses from two conditions were used to estimate $f' = \frac{\Delta\mu}{\Delta\theta}$, by taking the difference of their mean response normalized by the difference in stimulus. The covariance was estimated in a similar manner as well, except by taking the average of the two stimulus-specific covariances rather than normalizing by the difference in stimulus $\Delta\theta$. Linear fisher information was estimated from these quantities using partitioned matrix inversion techniques from [46]. To estimate the inferred information limit from the population response, a curve was fit via least-squares regression to the data $1/I_{fisher}$ vs $1/N$, discarding the data from the first 10% of cells (so the curve starts fitting from 10% of the total cell population size, rather than from the first cell). This was done to provide a better linear fit in line with the theory laid out in the authors' original work.

Parameters used for the figures

Figure 1

The histograms of NCs in Fig. 1 were generated by simulating activity from a population of neurons with $N = 1000$. The neural activity followed von mises (see 'Continuous Estimation' in Simulations section above) tuning with $\kappa = 0.75$, and activity parameter A = 15. Neurons were presumed to have von mises tuning curves with preferred stimuli chosen uniformly along $[0, 2\pi]$. For the conditions where information limiting noise was present ($\sigma > 0$), the standard deviation of the input noise was set at 5 deg. When gain fluctuations were present in the simulations, $\alpha = 2$. 5000 sample responses were generated from a uniform sampling of stimuli over the stimulus range $[0, 2\pi]$. The responses simulated neural activity within a time bin with width 120ms.

Figure 2

Parameters for the simulations used to generate the scatter plots and histograms in this figure were identical to those used for figure 1 listed above with the following differences: $\kappa = 3$, $A = 10$, and the number of iterations for each condition was 300. The decoding window for the decoder was 120ms, the same window size used to generate the activity, as in the previous figure.

Figure 3

For the simulations in this figure, the number of neurons along the x-axis was chosen with 20 steps with linearly spacing from 50 cells at the fewest to 2000 neurons at the largest population size. $\kappa = 2$, $A = 8$, $\sigma = 10$ deg. Neurons were presumed to have von mises tuning curves with preferred stimuli chosen uniformly along $[0, 2\pi]$. The number of iterations was 100, as in figure 1. Rather than sample activity from discrete bins, stimuli was generated by uniformly sampling 5000 samples from $\mathcal{U}(0, 2\pi)$ and the tuning curves for were used to estimate the continuous response to each stimuli, assuming a 120ms response window.

Figure 4

The number of neurons chosen for simulations were $[100, 250, 500, 750, 1000]$, and the number of samples per condition were $[25, 50, 100, 250, 500]$. Values were chosen to cover an empirically feasible range while being able to display some level of asymptotic behavior. Neurons were presumed to have von mises tuning curves with preferred stimuli chosen uniformly along $[0, 360]$ deg. The number of samples used for training the SVM was 75% of the total trials for the given iteration, of which there were 100 iterations per condition (cell count and sample count pair), with a 3-fold cross-validation procedure performed using the sklearn python package, selecting the best model while ranging over a SVC with parameters $max_iterations \in [100, 1000]$, and regularization parameter C$\in [10^{-4}, 10^{2}]$. The ground truth d' values were [2,6,10], as labeled. For the matrix D described in the simulation section above, the 5 largest eigenvalues were multiplied by 5 to increase the level of (non information-limiting) correlation in the population response, to provide more challenge to the analysis techniques used. For implementation of the method

from Rumyantsev et al. [17], data was reduced to 5 PLS dimensions, consistent with their original report.

Figure 5

For each dataset, neuron counts were selected with 10 steps with geometric spacing from 50 to the maximum cell count for each session (generally in the range of $\sim$14,000-20,000 cells). Each decoding split was performed with 50 iterations. The λ parameter for the ridge regression based method used in [22] was set to 1.

Figure 6

Cells counts were selected along the range of $[25, 50, 100, 250, Max]$. Maximum recorded cell counts for the sessions ranged from 500-850 cells. The widths of the integration windows were chosen from 20ms to 300ms in steps of 20ms. The number of iterations performed for each condition (cell count, integration window width) was 100. The number of samples used for training the SVM was 75% of the total trials for the given iteration, with a 3-fold cross-validation procedure performed using the sklearn Python package, selecting the best model while ranging over an SVC with parameters $max_iterations \in [100, 1000]$, and regularization parameter C$\in [10^{-4}, 10^{2}]$.

Figure 7

The number of neurons used in each iteration was selected with linear spacing from 10 to the maximum number of cells for the given session, with 10 steps in total. There were a total of 50 iterations performed for each cell population size in each experimental recording session. The parameters for the ZIG decoder [34], were $n_epochs = 1800, learning_rate = 0.01, gen_nodes = 30, thresh = 0, gam_shift = 0.0001, bin_len = 181$. The decoder was trained on 60% of the total data and the remaining 40% of data were split into half test and half validation.

References

[1] Marlene R Cohen and Adam Kohn. Measuring and interpreting neuronal correlations. *Nature neuroscience*, 14(7):811, 2011.

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

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

[4] Marlene R Cohen and John HR Maunsell. Attention improves performance primarily by reducing interneuronal correlations. *Nature neuroscience*, 12(12):1594, 2009.

[5] Jude F Mitchell, Kristy A Sundberg, and John H Reynolds. Spatial attention decorrelates intrinsic activity fluctuations in macaque area v4. *Neuron*, 63(6):879–888, 2009.

[6] Stefano Panzeri, Monica Moroni, Houman Safaai, and Christopher D Harvey. The structures and functions of correlations in neural population codes. *Nature Reviews Neuroscience*, 23(9):551–567, 2022.

[7] Ingmar Kanitscheider, Ruben Coen-Cagli, Adam Kohn, and Alexandre Pouget. Measuring fisher information accurately in correlated neural populations. *PLoS computational biology*, 11(6):e1004218, 2015.

[8] Guillaume Hennequin, Yashar Ahmadian, Daniel B Rubin, Máté Lengyel, and Kenneth D Miller. The dynamical regime of sensory cortex: stable dynamics around a single stimulus-tuned attractor account for patterns of noise variability. *Neuron*, 98(4):846–860, 2018.

[9] Larry F Abbott and Peter Dayan. The effect of correlated variability on the accuracy of a population code. *Neural computation*, 11(1):91–101, 1999.

[10] Alexander S Ecker, George H Denfield, Matthias Bethge, and Andreas S Tolias. On the structure of neuronal population activity under fluctuations in attentional state. *Journal of Neuroscience*, 36(5):1775–1789, 2016.

[11] Maoz Shamir and Haim Sompolinsky. Implications of neuronal diversity on population coding. *Neural computation*, 18(8):1951–1986, 2006.

[12] 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, 2016.

[13] 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, 2016.

[14] Alexander S Ecker, Philipp Berens, Andreas S Tolias, and Matthias Bethge. The effect of noise correlations in populations of diversely tuned neurons. *The Journal of Neuroscience*, 31(40):14272–14283, 2011. PMCID: PMC3221941.

[15] Rava Azeredo da Silveira and Fred Rieke. The geometry of information coding in correlated neural populations. *Annual Review of Neuroscience*, 44:403–424, 2021.

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

[17] 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*, 580(7801):100–105, 2020.

[18] Mohammad Mehdi Kafashan, Anna W Jaffe, Selmaan N Chettih, Ramon Nogueira, Iñigo Arandia-Romero, Christopher D Harvey, Rubén Moreno-Bote, and Jan Drugowitsch. Scaling of sensory information in large neural populations shows signatures of information-limiting correlations. *Nature communications*, 12(1):473, 2021.

[19] Klaus Wimmer, Duane Q Nykamp, Christos Constantinidis, and Albert Compte. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature neuroscience*, 17(3):431–439, 2014.

[20] Ramon Bartolo, Richard C Saunders, Andrew R Mitz, and Bruno B Averbeck. Information-limiting correlations in large neural populations. *Journal of Neuroscience*, 40(8):1668–1678, 2020.

[21] Omer Hazon, Victor H Minces, David P Tomàs, Surya Ganguli, Mark J Schnitzer, and Pablo E Jercog. Noise correlations in neural ensemble activity limit the accuracy of hippocampal spatial representations. *Nature communications*, 13(1):4276, 2022.

[22] Carsen Stringer, Michalis Michaelos, Dmitri Tsyboulski, Sarah E Lindo, and Marius Pachitariu. High-precision coding in visual cortex. *Cell*, 184(10):2767–2778, 2021.

[23] C Radhakrishna Rao et al. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc*, 37(3):81–91, 1945.

[24] Michael N Shadlen, Kenneth H Britten, William T Newsome, and J Anthony Movshon. A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *Journal of neuroscience*, 16(4):1486–1510, 1996.

[25] David Marvin Green, John A Swets, et al. *Signal detection theory and psychophysics*, volume 1. Wiley New York, 1966.

[26] Zaki Ajabi, Alexandra T Keinath, Xue-Xin Wei, and Mark P Brandon. Population dynamics of the thalamic head direction system during drift and reorientation. *bioRxiv*, 2021.

[27] Zaki Ajabi, Alexandra T Keinath, Xue-Xin Wei, and Mark P Brandon. Population dynamics of head-direction neurons during drift and reorientation. *Nature*, 615(7954):892–899, 2023.

[28] Jeffrey S Taube, Robert U Muller, and James B Ranck. Head-direction cells recorded from the post-subiculum in freely moving rats. i. description and quantitative analysis. *Journal of Neuroscience*, 10(2):420–435, 1990.

[29] Jonathan Green, Atsuko Adachi, Kunal K Shah, Jonathan D Hirokawa, Pablo S Magani, and Gaby Maimon. A neural circuit architecture for angular integration in drosophila. *Nature*, 546(7656):101–106, 2017.

[30] Daniel Turner-Evans, Stephanie Wegener, Herve Rouault, Romain Franconville, Tanya Wolff, Johannes D Seelig, Shaul Druckmann, and Vivek Jayaraman. Angular velocity integration in a fly heading circuit. *Elife*, 6:e23496, 2017.

[31] Kechen Zhang. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *Journal of Neuroscience*, 16(6):2112–2126, 1996.

[32] A David Redish, Adam N Elga, and David S Touretzky. A coupled attractor model of the rodent head direction system. *Network: computation in neural systems*, 7(4):671, 1996.

[33] Rani Ben-Yishai, R Lev Bar-Or, and Haim Sompolinsky. Theory of orientation tuning in visual cortex. *Proceedings of the National Academy of Sciences*, 92(9):3844–3848, 1995.

[34] Xue-Xin Wei, Ding Zhou, Andres Grosmark, Zaki Ajabi, Fraser Sparks, Pengcheng Zhou, Mark Brandon, Attila Losonczy, and Liam Paninski. A zero-inflated gamma model for post-deconvolved calcium imaging traces. *Neurons, Behavior, Data analysis, and Theory*, 3(2):1–21, 2020.

[35] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

[36] Nikolaus Kriegeskorte and Xue-Xin Wei. Neural tuning and representational geometry. *Nature Reviews Neuroscience*, 22(11):703–718, 2021.

[37] Annemarie Wolff, Nareg Berberian, Mehrshad Golesorkhi, Javier Gomez-Pilar, Federico Zilio, and Georg Northoff. Intrinsic neural timescales: temporal integration and segregation. *Trends in cognitive sciences*, 26(2):159–173, 2022.

[38] Yuzhi Chen, Wilson S Geisler, and Eyal Seidemann. Optimal temporal decoding of neural population responses in a reaction-time visual detection task. *Journal of neurophysiology*, 99(3):1366–1379, 2008.

[39] Robbe LT Goris, Corey M Ziemba, J Anthony Movshon, and Eero P Simoncelli. Slow gain fluctuations limit benefits of temporal integration in visual cortex. *Journal of vision*, 18(8):8–8, 2018.

[40] Jamie D Roitman and Michael N Shadlen. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of neuroscience*, 22(21):9475–9489, 2002.

[41] Natalie Steinemann, Gabriel M Stine, Eric Trautmann, Ariel Zylberberg, Daniel M Wolpert, and Michael N Shadlen. Direct observation of the neural computations underlying a single decision. *Elife*, 12:RP90859, 2024.

[42] Amy M Ni, Douglas A Ruff, Joshua J Alberts, Jen Symmonds, and Marlene R Cohen. Learning and attention reveal a general relationship between population activity and behavior. *Science*, 359(6374):463–465, 2018.

[43] Joao Barbosa, Heike Stein, Rebecca L Martinez, Adrià Galan-Gadea, Sihai Li, Josep Dalmau, Kirsten Adam, Josep Valls-Solé, Christos Constantinidis, and Albert Compte. Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. *Nature Neuroscience*, 23(8):1016–1024, 2020.

[44] Kechen Zhang, Iris Ginzburg, Bruce L McNaughton, and Terrence J Sejnowski. Interpreting neuronal population activity by reconstruction: unified framework with application to hippocampal place cells. *Journal of neurophysiology*, 79(2):1017–1044, 1998.

[45] Jianqing Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, pages 1257–1272, 1991.

[46] Charles A Rohde. Generalized inverses of partitioned matrices. *Journal of the Society for Industrial and Applied Mathematics*, 13(4):1033–1035, 1965.

[47] Erich Leo Lehmann and George Casella. *Theory of point estimation*. Springer, 1998.

[48] Peggy Seriès, Peter E Latham, and Alexandre Pouget. Tuning curve sharpening for orientation selectivity: coding efficiency and the impact of correlations. *Nature neuroscience*, 7(10):1129, 2004.

[49] Steve Cherry. Singular value decomposition analysis and canonical correlation analysis. *Journal of Climate*, 9(9):2003–2009, 1996.

Acknowledgments

This research uses datasets from a number of previously published studies. We would like to thank the authors of these studies for generously sharing the data or making their data publicly available. We thank Robbe Goris for helpful discussions. D.L. is supported by an GRFP fellowship from U.S. National Science Foundation (NSF). X.X.W. is supported by a Sloan Research Fellowship from the Alfred P. Sloan Foundation.

Supplementary Information

A. Implementations of three alternative methods

A1. Standard (naive) decoding approach

According to the Cramer-Rao bound, the variance of an unbiased estimator approaches the limit of $1/I(\theta)$ where $I(\theta)$ is the Fisher information contained in the population response about the variable θ. When we take the variance of our estimators' decoding error as neuronal population sizes increase, we are essentially measuring the convergence to the information-limiting fisher information. Without using tailor-made tools and techniques, this is the most intuitive approach to estimate information limits in empirically recorded neural data, and so was chosen as our 'naive' approach for comparison purposes in this work. The idea of using a cross-validated decoder to estimate information-limiting noise was proposed in [16]. Results showing the performance of this method are shown in Fig. 3.

In a discrete classification task, we approximate the Fisher information instead using a measure of discriminability (i.e., d-prime), which can be viewed as $\sqrt{I_{lin}(\theta)}$, the square root of the linear Fisher information [16, 7]. The intuitive justification follows that of the continuous case otherwise. The results on the performance of this method are shown in Fig. S2.

Overall, the main problem with this standard decoding approach is that it exhibits a bias for over-estimating of the magnitude of the information-limiting noise. For the case of estimating d-prime, it generally under-estimates the ground truth d-prime values. Such estimation biases reduce as the number of neurons used increases.

A2. Method of linear extrapolation

The implementation of the technique from [18] was done by directly adapting the Matlab code (shared by the authors in their original report [18]) to a re-implementation in Python. In short, fisher information was calculated from population responses as neurons were added one at a time, and then a line was fit to the data when plotted $1/N$ vs $1/I(\theta)$. The y-intercept of the line, when extrapolated, would give the I_∞ estimate. In practice we chose to start projection from the first 10% of the population (so after the first 100-200 cells), for reasons to be described below.

The approach of [18] was essentially to note that after a sufficient amount of neurons are present in the population, the amount of information incremented by each additional neuron can be viewed as essentially constant, such that increasing population size leads to linear increases in information (as measured by fisher information). The intuition behind needing a "sufficient" amount of neurons before linearity can be observed primarily arises from the fact that one first needs enough neurons to have a complete (coarse) representation of the stimulus space. Once the stimulus space is properly spanned by the existing neurons' tuning curves, then any stimulus can be represented by population activity and

each additional neuron adds a constant amount of additional information.

A potential point of concern for the method is that the linear projection to the I_∞ value may be unstable depending on the size of the population and the magnitude of the (information-limiting) noise. If the population is too small, then additional neurons grant non-constant increases in population fisher information as they potentially cover unique portions of stimulus space. If the limiting noise is quite small, then the linear extrapolation can have difficulties getting the precise value because the fit is performed in $1/I(\theta)$, and when I_∞ is large, the ensuing fraction can be quite small and numerical instability may occur. In our simulations, we find that this method based on extrapolation [18] is numerically unstable and it depends on the number of trials and the number of neurons. Simulation results to demonstrate these points are shown in Fig. 4, Fig. S2, and Figs. S3- S5.

A3. Method of dimensionality reduction

The approach from [17] for classification between two stimuli was implemented as described in the methods section of their own report. In essence, the neural response data is split into 3 partitions, which we'll denote $X_{train}, X_{weight}, X_{test}$ which are used for training the transformer, obtaining the optimal weight, and calculating the final d' value, respectively. The first partition X_{PLS} is used to train a partial least-squares regression model, which reduces data down to a lower dimension. The authors discovered n=5 was optimal for their data, which we followed in our implementation as well. Then the model was used to transform another data partition, X_{train} into the low dimensional space. The resulting low dimensional data was used to estimate $\Sigma = 0.5\,(\Sigma_1 + \Sigma_2)$, the low dimensional noise covariance matrix, where Σ_i were the noise covariance matrices for the dimension reduced neural responses from either of the two stimulus conditions. The optimal transformation weight was then found calculating $w_{opt} = \Sigma^{-1} \cdot \Delta\mu_{weight}$. The final partition is used to calculate d' by performing $\Delta\mu_{test}^T \cdot w_{opt}$.

In general, the method operates by assuming that the main drivers of neuronal noise can be captured in a low dimensional subspace, and by reducing the dimensionality, the noise covariance matrix estimation is much more numerically feasible. It is in particular effective considering theories that information-limiting noise is restricted to a single direction tangent to the signal manifold [16]. However, it should be noted that the method, as described, only seeks to estimate the information for a finite number of neurons, namely, the neurons present in the population. Specifically, the method does *not* seek to estimate I_∞. As a result, its primary use seems to be to best estimate the amount of information encoded by an empirically recorded neural population. As a result, the method is not optimal for detecting information-limits in a population, and is best used to estimate the maximal information that can be extracted from a given finite population. In practice, this allows the method to perform better than naive estimation procedures (SI Fig. 23), but falls short of estimating the information-limiting noise in general, in particular when the information-limiting noise is small (Fig. 4, Fig. S2).

B. Model justification when only information-limiting noise and priviate neural noise is present

In this section, we provide a justification for the split-trial analysis in the context of a standard neural population model in which only information-limiting noise causes noise correlations.

Model setting Assume that a one-dimensional stimulus variable θ is represented by a neural population code with N neurons (N may approach infinity). Denote the tuning curve of the i_{th} neuron as $f_i(\theta)$, which is a smooth function. $f_i(\theta)$ does not need to be Gaussian function. We will assume that neurons exhibit Poisson spiking noise, although the arguments below may also work for other types of noise (such as negative binomial noise or Gaussian noise). Thus, the code can be expressed as the following:

$$r_i \sim \text{Poisson}(f_i(\theta)\Delta t),$$

where r_i represents the activity of the i_{th} neuron.

We model information-limiting noise by injecting Gaussian noise to the stimulus

$$r_i \sim \text{Poisson}(f_i(\theta + \delta)\Delta t),$$

where random variable δ represents information-limiting noise and is shared by all neurons. Note that, in this model, noise correlations are only induced by information-limiting noise. The Poisson spiking noise is private to individual neurons. Although this setting is likely simplified, as in practice there may exist other types of shared neural variability that are not information-limiting, our simulation results in Fig. 3 show that it is already challenging to use standard decoding methods to estimate information-limiting noise.

Analysis of split-trial analysis Suppose that we collect neural responses from this population of neurons (N neurons), and perform split-trial analysis. Our goal is recover the variance of information-limiting noise δ. When performing split-trial analysis, we split the neural population into two halves and perform Bayesian decoding on each split to obtain the Bayesian estimator of the stimulus. In performing Bayesian decoding, we simply assume that neurons exhibit independent neural noise.

Assume that information-limiting noise δ is small and the tuning curve is smooth, given a sufficient amount of data, the tuning curves for individual neurons can be estimated accurately. For a sufficiently large neural population, the Bayesian estimator based on each split, $\hat{\theta}_i$ (i =1,2), is expected to be approximately Gaussian distributed [47], centered at $\theta + \delta$. The reason is because Bayesian decoding in this cases is essentially decoding $\theta + \delta$ from a large neural population with independent neural noise. Thus, we can write

$$\hat{\theta}_1 = \theta + \delta + \phi_1, \hat{\theta}_2 = \theta + \delta + \phi_2,$$

where ϕ_1 and ϕ_2 represent the decoding error of $\theta+\delta$ for the two splits. Furthermore, because that neural noise is independent in the neural population (when treating the decoding problem as one that decodes

$\theta + \delta$), the decoding error from the two splits should be independent. Therefore, under the assumptions above, ϕ_1 and ϕ_2 are independent random variables. Given these, a straightforward calculation (as shown in the Methods section) shows that the covariance of $\hat{\theta_1} - \theta_1$ and $\hat{\theta_2} - \theta_2$ can be used to estimate the variance of information-limiting noise δ.

Suppose that variance of information-limiting noise is σ^2. As the variances of ϕ_1 and ϕ_2 should be nearly identical when N is large, for simplicity we will assume that they are both σ_p^2. When using all N neurons to perform Bayesian decoding analysis, the variance of Bayesian estimator should be of the order of $\frac{\sigma_p^2}{2} + \sigma^2$. This is because integrating information from the two splits should reduce the variance of the estimator from the private noise by a factor 2, while preserving the variance due to the shared noise δ. Thus, using Bayesian decoding to estimate the variance of information-limiting noise suffers the problem of over-estimation, with an upward bias of $\frac{\sigma_p^2}{2}$. While this bias will converge to 0 when the number of neurons is infinite, practically the bias can be substantial, as we showed in our simulation results. Increasing the number of trials will not eliminate this bias.

For split-trial analysis, empirically the covariance estimator $Cov(\hat{\theta_1} - \theta_1, \hat{\theta_2} - \theta_2) = Var(\delta) + Cov(\phi_1, \phi_2)$ may exhibit statistical fluctuations, due to noisy empirical estimates of $Cov(\phi_1, \phi_2)$ (see Fig. 2). However, the estimator should generally be unbiased as $Cov(\phi_1, \phi_2)$ is expected to be 0. Our simulation results in Fig. 3a,c show that when the number of neurons becomes larger, the variability of this estimator reduces and becomes more accurate. For finite samples, the estimator is off by $Cov(\phi_1, \phi_2)$. Practically, this amounts to $\rho\sigma_p^2$, where ρ is the Pearson correlation of the samples of ϕ_1, ϕ_2. From Fig. 2f, we see that ρ is generally small with most values fall between [-0.1, 0.1]. For the sake of argument, suppose that ρ is 0.1, the error in estimating the information limiting should be about a fifth of the error of the standard (naive) method based on Bayesian decoding. Furthermore, our estimator does not suffer from the systematic overestimation of the naive method.

Remarks: The analysis above assumes that all noise correlations are caused by information-limiting noise. When other sources shared neural variable other than the information-limiting noise are present, one can still perform Bayesian decoding on the two splits. When the number of neurons is large, one may still approximately model the Bayesian estimator using a similar approach as above, i.e., $\hat{\theta_1} = \theta + \delta + \phi_1, \hat{\theta_2} = \theta + \delta + \phi_2$. However, for a finite neuron number N, ϕ_1 and ϕ_2 may be correlated depending on the exact structure of the shared, non-information-limiting noise in the neural population. When N is large, $Cov(\phi_1, \phi_2)$ should approach 0, otherwise there would be non-zero information-limiting noise shared by ϕ_1, ϕ_2 (which would contradict to the model assumptions). Such non-zero correlation between ϕ_1 and ϕ_2 may increase the error of our split-trial estimator. However, practically, we find that, at least for the case of gain fluctuations, our split-trial estimator still works well even when we use a decoder that assumes independent neural noise (see Fig. 3). In the next section, we perform additional analysis based on general Gaussian noise correlation structures, in the context of a neural population coding model assuming Gaussian tuning curves. These analysis provides additional justification regarding why we expect the split-trial analysis to work well in the more general setting.

C. Model justification under more general assumptions on noise covariance

In this section, we will investigate the properties of split-trial analyis under more general noise assumptions. We will do so in the context of neural populations with Gaussian tuning curves. We assume that individual neurons in the neural population exhibit Guassian tuning curves for stimulus variable θ:

$$f_i(\theta) = \frac{r_{max}}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\theta - \theta_{\text{pref,i}}}{\sigma_i}\right)^2\right),$$

where $\theta_{\text{pref},i}$ is the preferred stimulus for neuron i, and r_{max} is a firing rate parameter which determines the scaling of the neuronal firing rate. We let $\vec{\mu}(\theta) = \vec{f}(\theta) = [f_1(\theta), f_2(\theta)....f_n(\theta)]^T$. We will, for each stimulus parameter θ, let the inverse covariance matrix $\Sigma^{-1}(\theta)$ be a diagonal matrix, diag(σ_i^{-2}) , reflecting the fact that decoding of empirical data often assumes independence in activity across the population. Then population responses $\vec{x}$ are drawn according to $\vec{x} \sim \mathcal{N}(\vec{\mu}, \Sigma)$, omitting μ and Σ's dependence on θ.

C.1 Properties of the locally optimal linear estimator

First, we will investigate the mean and variance in the error of the locally optimal linear estimator (LOLE). We will use these results for further calculations and for assessment of the split-trial technique. Note that by using the locally optimal linear estimator, we are assuming that the decoder has knowledge of the noise correlation structure that is induced by non-information-limiting noise. The locally optimal linear estimator for $\hat{\theta}$ can be written as:

$$\hat{\theta} = \theta + \frac{\vec{f}'(\theta)^T\Sigma^{-1}(\theta)(\vec{x} - \vec{f}(\theta))}{\vec{f}'(\theta)^T\Sigma^{-1}(\theta)\vec{f}'(\theta)}$$

known to be optimal when estimating a parameter value near θ [48, 16]. Starting with the denominator of the fraction:

$$\begin{aligned}\vec{f}'(\theta)^T\Sigma^{-1}(\theta)\vec{f}'(\theta) &= f'^T\text{diag}(\sigma_i^{-2})f' \\ &= \sum_{i=1}^{n} \frac{f_i'(\theta)^2}{\sigma_i^2}.\end{aligned}$$

We evaluate the numerator similarly:

$$\begin{aligned}\vec{f}'(\theta)^T\Sigma^{-1}(\theta)(\vec{x} - \vec{f}(\theta)) &= f'^T\text{diag}(\sigma_i^{-2})\left(\vec{x} - \vec{f}(\theta)\right) \\ &= \begin{bmatrix} f_1'(\theta)/\sigma_1^2 \\ \vdots \\ f_n'(\theta)/\sigma_n^2 \end{bmatrix} \cdot \left(\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} - \begin{bmatrix} f_1(\theta) \\ \vdots \\ f_n(\theta) \end{bmatrix}\right) \\ &= \sum_{i=1}^{n} \frac{x_i f_i'(\theta)}{\sigma_i^2} - \sum_{i=1}^{n} \frac{f_i(\theta) f_i'(\theta)}{\sigma_i^2} \\ &= \sum_{i=1}^{n} \frac{(x_i - f_i(\theta)) f_i'(\theta)}{\sigma_i^2}\end{aligned}$$

We can then combine terms and evaluate the decoding error:

$$\hat{\theta} - \theta = \frac{\sum_{i=1}^{n} \frac{(x_i - f_i(\theta)) f_i'(\theta)}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{f_i'(\theta)^2}{\sigma_i^2}}$$

From which we can see that, were we to take the expectation of $\hat{\theta} - \theta$, we'd find the error has mean 0 since $\mathsf{E}[x_i] = f_i(\theta)$.

We proceed with the variance calculation:

$$\begin{aligned}
\mathsf{Var}\left[\hat{\theta} - \theta\right] &= \mathsf{E}\left[\left(\hat{\theta} - \theta\right)^2\right] - \mathsf{E}\left[\hat{\theta} - \theta\right]^2 \\
&= \mathsf{E}\left[\left(\hat{\theta} - \theta\right)^2\right] \\
&= \mathsf{E}\left[\left(\frac{\sum_{i=1}^{n} \frac{(x_i - f_i) f_i'}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{f_i'^2}{\sigma_i^2}}\right)^2\right] \\
&= \frac{\mathsf{E}\left[\sum_{i=1}^{n} \left(\frac{(x_i - f_i) f_i'}{\sigma_i^2}\right)^2 + 2 \sum_{i<j}^{n} \frac{(x_i - f_i)(x_j - f_j) f_i' f_j'}{\sigma_i^2 \sigma_j^2}\right]}{\left(f'^T \Sigma^{-1} f'\right)^2}.
\end{aligned}$$

We'll evaluate each term individually starting with the diagonal term in the numerator:

$$\begin{aligned}
\mathsf{E}\left[\sum_{i=1}^{n} \left(\frac{(x_i - f_i) f_i'}{\sigma_i^2}\right)^2\right] &= \mathsf{E}\left[\sum_{i=1}^{n} \frac{x_i^2 f_i'^2 - 2 x_i f_i f_i'^2 + f_i^2 f_i'^2}{\sigma_i^4}\right] \\
&= \sum_{i=1}^{n} \frac{\mathsf{E}[x_i^2] f_i'^2 - 2\mathsf{E}[x_i] f_i f_i'^2 + f_i^2 f_i'^2}{\sigma_i^4} \\
&= \sum_{i=1}^{n} \frac{\mathsf{E}[x_i^2] f_i'^2 - f_i^2 f_i'^2}{\sigma_i^4} \\
&= \sum_{i=1}^{n} \frac{(\sigma_i^2 + f_i^2) f_i'^2 - f_i^2 f_i'^2}{\sigma_i^4} \\
&= \sum_{i=1}^{n} \frac{\sigma_i^2 f_i'^2}{\sigma_i^4} \\
&= \sum_{i=1}^{n} \frac{f_i'^2}{\sigma_i^2}.
\end{aligned}$$

Now we address the numerator cross-term:

$$
\begin{aligned}
2\mathsf{E}\left[\sum_{i<j}^{n}\frac{(x_i-f_i)(x_j-f_j)f_i'f_j'}{\sigma_i^2\sigma_j^2}\right] &= 2\mathsf{E}\left[\sum_{i<j}^{n}\frac{x_ix_jf_i'f_j'-x_if_jf_i'f_j'-x_jf_if_i'f_j'+f_if_jf_i'f_j'}{\sigma_i^2\sigma_j^2}\right] \\
&= 2\sum_{i<j}^{n}\frac{\mathsf{E}[x_ix_j]f_i'f_j-\mathsf{E}[x_i]f_jf_i'f_j'-\mathsf{E}[x_j]f_if_i'f_j'+f_if_jf_i'f_j'}{\sigma_i^2\sigma_j^2} \\
&= 2\sum_{i<j}^{n}\frac{(\mathsf{Cov}(x_i,x_j)+\mathsf{E}[x_i]\mathsf{E}[x_j])f_i'f_j'-f_if_jf_i'f_j'}{\sigma_i^2\sigma_j^2} \\
&= 2\sum_{i<j}^{n}\frac{(\mathsf{Cov}(x_i,x_j)+f_if_j)f_i'f_j'-f_if_jf_i'f_j'}{\sigma_i^2\sigma_j^2} \\
&= 2\sum_{i<j}^{n}\frac{\mathsf{Cov}(x_i,x_j)f_i'f_j'}{\sigma_i^2\sigma_j^2}.
\end{aligned}
$$

So we have that

$$
\begin{aligned}
\mathsf{Var}[\hat{\theta}-\theta] &= \frac{\sum_{i=1}^{n}\frac{f_i'^2}{\sigma_i^2}+2\sum_{i<j}^{n}\frac{\mathsf{Cov}(x_i,x_j)f_i'f_j'}{\sigma_i^2\sigma_j^2}}{\left(f'^T\Sigma^{-1}f'\right)^2} \\
&= \frac{\vec{f'}^T\Sigma^{-1}\vec{f'}+2\sum_{i<j}^{n}\frac{\mathsf{Cov}(x_i,x_j)f_i'f_j'}{\sigma_i^2\sigma_j^2}}{\left(f'^T\Sigma^{-1}f'\right)^2} \\
&= \frac{1}{\vec{f'}^T\Sigma^{-1}\vec{f'}}+\frac{2\sum_{i<j}^{n}\frac{\mathsf{Cov}(x_i,x_j)f_i'f_j'}{\sigma_i^2\sigma_j^2}}{\left(f'^T\Sigma^{-1}f'\right)^2}
\end{aligned}
$$

The denominator of this expression is the linear Fisher information arising if the population exhibits a covariance that aligns with the assumption of the estimator (which in this case was that the activity is independent), whereas the numerator contains the $\mathsf{Cov}(x_i,x_j)$ term which would track the true covariance values. If the estimator has perfect knowledge of the ground truth then $Var[\hat{\theta}-\theta]=\frac{1}{I_{fisher}}$ and the Cramer-Rao bound is achieved as $\mathsf{Cov}(x_i,x_j)=0,\forall i\neq j$. If there is a mismatch in ground truth and presumed covariance structure, then the variance is larger by the second term in the above equation. The second term cannot be negative as that would violate the Cramer-Rao bound. For the intuition behind this, consider: if $\mathsf{Cov}(x_i,x_j)<0$, then $f_i'>0 \implies f_j'<0$, and the numerator of the sum remains positive. All other terms are squares and forced positive. In other words, if two variables are negatively correlated, then one must increase firing rate when the other decreases, thus the product of the f' terms is negative whenever $\mathsf{Cov}(x_i,x_j)$ is negative, ensuring positivity of the second term, the case where variables are positively correlated follows.

C2. Covariance calculations

In this section, we will examine the performance of split-trial analysis under different noise regimes, characterized by assumptions on the ground truth noise covariance matrices, Σ. Let $Y=\hat{\theta}_1-\theta$ and

$Z = \hat{\theta}_2 - \theta$ be the error estimate from the first and second population splits given stimulus θ, respectively. Additionally, let each (gaussian) tuning curve be defined by mean g_i, h_j with $\sigma^2 = \eta^2, \nu^2$ (i.e. tuning curves of subpopulation 1 have the form $f_i(\theta) = \frac{r_{max}}{\eta_i\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\theta - g_i}{\eta_i}\right)^2\right)$) respectively as well. We will use only a two-fold split for proceeding calculations. This is equivalent to considering Σ as a block matrix $\Sigma = \begin{bmatrix} \Sigma_1 \Sigma_{12} \\ \Sigma_{12}^T \Sigma_2 \end{bmatrix}$ where e.g. Y is calculated as above using Σ_1. Specifically, we would represent $\Sigma_0 = \begin{bmatrix} \Sigma_{0,Y} \Sigma_{0,YZ} \\ \Sigma_{0,YZ}^T \Sigma_{0,Z} \end{bmatrix}$. Additionally, we will presume that our covariance matrix has the general structure of $\Sigma = \Sigma_0 + \epsilon f' f'^T$, where ϵ is the magnitude of the information limiting noise, and Σ_0 is a covariance matrix with no information-limiting components, following the convention of [16].

C2.1. Split-trial analysis when only information-limiting noise is present

We will start with special case for which $\Sigma_0 = \vec{\sigma}\mathbf{I}$, a diagonal matrix. In this case, all noise correlations are caused by information-limiting noise.

$$\begin{aligned}
\text{Cov}(Y, Z) &= \text{E}[YZ] - \text{E}[Y]\text{E}[Z] \\
&= \text{E}[YZ] \\
&= \frac{1}{I_Y I_Z} \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{\text{Cov}(y_i, z_j) g_i' h_j'}{\eta_i^2 \nu_j^2} \\
&= \frac{1}{I_Y I_Z} \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{\epsilon g_i' h_j' g_i' h_j'}{\eta_i^2 \nu_j^2} \\
&= \frac{\epsilon}{I_Y I_Z} \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{g_i'^2 h_j'^2}{\eta_i^2 \nu_j^2} \\
&= \epsilon
\end{aligned}$$

which confirms the validity of our estimator in the case where information-limiting noise is added to an otherwise independent population, in the case of a population of neurons with Gaussian tuning curve and Gaussian noise. These results complementary to the results reported in Section B.

C2.2 Split-trial Analysis with general noise covariance matrix

In this section, we will relax assumptions on Σ_0 and examine the general case. Specially, we allow other forms of noise correlations that are note caused by information-limiting noise. Let k_{ij}^Y be the element of Σ_Y^{-1} in the ith row and jth column for the following calculation:

$$
\begin{aligned}
\mathsf{Cov}(Y,Z) &= \mathsf{E}\left[\left(\frac{\vec{g}'^T\Sigma_Y^{-1}(\vec{y}-\vec{g})}{\vec{g}'^T\Sigma_Y^{-1}\vec{g}'}\right)\left(\frac{\vec{h}'^T\Sigma_Z^{-1}(\vec{z}-\vec{h})}{\vec{h}'^T\Sigma_Z^{-1}\vec{h}'}\right)\right] \\
&= \frac{1}{I_{0,Y}I_{0,Z}}\mathsf{E}\left[\left(\vec{g}'^T\Sigma_Y^{-1}(\vec{y}-\vec{g})\right)\left(\vec{h}'^T\Sigma_Z^{-1}(\vec{z}-\vec{h})\right)\right] \\
&= \frac{1}{I_{0,Y}I_{0,Z}}\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{j=1}^{n}\sum_{i=1}^{n}\mathsf{Cov}(y_j,z_l)g_i'h_k'k_{ij}^Y k_{kl}^Z \\
&= \frac{1}{I_{0,Y}I_{0,Z}}\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{j=1}^{n}\sum_{i=1}^{n}(C_{jl}+\epsilon g_j'h_l')g_i'h_k'k_{ij}^Y k_{kl}^Z \\
&= \frac{\epsilon}{I_{0,Y}I_{0,Z}}\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{j=1}^{n}\sum_{i=1}^{n}g_j'k_{ij}^Y g_i'h_k'k_{kl}^Z h_l' + \frac{1}{I_{0,Y}I_{0,Z}}\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{j=1}^{n}\sum_{i=1}^{n}C_{jl}g_i'h_k'k_{ij}^Y k_{kl}^Z \\
&= \epsilon + \frac{1}{I_{0,Y}I_{0,Z}}\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{j=1}^{n}\sum_{i=1}^{n}C_{jl}g_i'h_k'k_{ij}^Y k_{kl}^Z \\
&= \epsilon + \frac{\vec{g}'^T\Sigma_{0,Y}^{-1}\Sigma_{0,YZ}\Sigma_{0,Z}^{-1}\vec{h}'}{I_{0,Y}I_{0,Z}}
\end{aligned}
$$

where $I_{0,Y}, I_{0,Z}$ are the fisher information values calculated using $\Sigma_{0,Y}, \Sigma_{0,Z}$, respectively. We examine this second term causing deviations away from ϵ in the following section.

C.3 Examination of the error term in split-trial analysis with general noise covariance matrix

Here we attempt to develop an understanding of the second term in the above calculations for $\mathsf{Cov}(Y,Z)$, $\frac{g'\Sigma_{0,Y}^{-1}\Sigma_{0,YZ}\Sigma_{0,Z}^{-1}h'}{I_{0,Y}I_{0,Z}}$. Specifically we will be focusing on the numerator for this expression: $g'\Sigma_{0,Y}^{-1}\Sigma_{0,YZ}\Sigma_{0,Z}^{-1}h'$. It will be convenient to work with block matrices for this section, so we will be representing Σ_0 as $\Sigma_0 = \begin{bmatrix}\Sigma_{0,Y} & \Sigma_{0,YZ}\\ \Sigma_{0,YZ}^T & \Sigma_{0,Z}\end{bmatrix}$. As stated previously, the magnitude of information-limiting noise present in a population of neurons with covariance matrix Σ is $\vec{f}'^T\Sigma\vec{f}'$, for tuning curve derivatives $\vec{f}'$. We will also be representing these tuning curves in a block format:$\vec{f}' = \begin{bmatrix}\vec{g}'\\ \vec{h}'\end{bmatrix}$.

Intuition and the error term as a whitening transform

To proceed we will be transforming $\vec{f}'$ by multiplication with $\Sigma_{0,ind}^{-1} = \begin{bmatrix}\Sigma_{0,Y}^{-1} & O\\ O & \Sigma_{0,Z}^{-1}\end{bmatrix}$ to obtain $\vec{f}_{tr}' = \Sigma_{0,ind}^{-1}\vec{f}'$. In effect this is the inverse of Σ_0 assuming Σ_0 had off-diagonal block $\Sigma_{0,YZ}$ = O, the zero matrix. Then

calculation of the information-limiting noise term results in (note that $\Sigma_{0,ind}$ is symmetric):

$$\begin{aligned}
\vec{f}'^T \Sigma_{0,ind}^{-1} \Sigma_0 \Sigma_{0,ind}^{-1} \vec{f}' &= \begin{bmatrix} \vec{g}'^T & \vec{h}'^T \end{bmatrix} \begin{bmatrix} \Sigma_{0,Y}^{-1} & O \\ O & \Sigma_{0,Z}^{-1} \end{bmatrix} \begin{bmatrix} \Sigma_{0,Y} & \Sigma_{0,YZ} \\ \Sigma_{0,YZ}^T & \Sigma_{0,Z} \end{bmatrix} \begin{bmatrix} \Sigma_{0,Y}^{-1} & O \\ O & \Sigma_{0,Z}^{-1} \end{bmatrix} \begin{bmatrix} \vec{g}' \\ \vec{h}' \end{bmatrix} \\
&= \begin{bmatrix} \vec{g}'^T & \vec{h}'^T \end{bmatrix} \begin{bmatrix} I & \Sigma_{0,Y}^{-1} \Sigma_{0,YZ} \\ \Sigma_{0,Z}^{-1} \Sigma_{0,YZ}^T & I \end{bmatrix} \begin{bmatrix} \Sigma_{0,Y}^{-1} & O \\ O & \Sigma_{0,Z}^{-1} \end{bmatrix} \begin{bmatrix} \vec{g}' \\ \vec{h}' \end{bmatrix} \\
&= \begin{bmatrix} \vec{g}'^T & \vec{h}'^T \end{bmatrix} \begin{bmatrix} \Sigma_{0,Y}^{-1} & \Sigma_{0,Y}^{-1} \Sigma_{0,YZ} \Sigma_{0,Z}^{-1} \\ \Sigma_{0,Z}^{-1} \Sigma_{0,YZ}^T \Sigma_{0,Y}^{-1} & \Sigma_{0,Z}^{-1} \end{bmatrix} \begin{bmatrix} \vec{g}' \\ \vec{h}' \end{bmatrix} \\
&= \vec{g}'^T \Sigma_{0,Y}^{-1} \vec{g}' + \vec{h}'^T \Sigma_{0,Z}^{-1} \vec{h}' + \vec{g}'^T \Sigma_{0,Y}^{-1} \Sigma_{0,YZ} \Sigma_{0,Z}^{-1} \vec{h}' + \vec{h}'^T \Sigma_{0,Z}^{-1} \Sigma_{0,YZ}^T \Sigma_{0,Y}^{-1} \vec{g}' \\
&= I_{0,Y} + I_{0,Z} + 2 \left(\vec{g}'^T \Sigma_{0,Y}^{-1} \Sigma_{0,YZ} \Sigma_{0,Z}^{-1} \vec{h}' \right)
\end{aligned}$$

where the latter two terms on the second to last line were combined as they are transposes of each other and constants. Whichever term we obtain from calculations is based on the ordering of Cov(Y, Z) vs Cov(Z, Y) so they are identical in value as covariance is commutative in ordering. Ordering the terms we find

$$\vec{g}'^T \Sigma_{0,Y}^{-1} \Sigma_{0,YZ} \Sigma_{0,Z}^{-1} \vec{h}' = 0.5 \left(\vec{f}_{tr}'^T \Sigma_0 \vec{f}_{tr}' - I_{0,Y} - I_{0,Z} \right).$$

In order to interpret the RHS and enable us to understand the LHS term, consider the case where $\Sigma_0 = \begin{bmatrix} \Sigma_{0,Y} & O \\ O & \Sigma_{0,Z} \end{bmatrix}$, then $\vec{f}'^T \Sigma_{0,ind}^{-1} \Sigma_0 \Sigma_{0,ind}^{-1} \vec{f}' = \vec{f}'^T \Sigma_0^{-1} \vec{f}' = I_{indep}$. The RHS from above would then read $0.5\,(I_{indep} - I_{0,Y} - I_{0,Z})$, we assume that our two split populations are statistically identical so $I_{0,Y} = I_{0,Z} = I_{half}$, which results in $\vec{g}'^T \Sigma_{0,Y}^{-1} \Sigma_{0,YZ} \Sigma_{0,Z}^{-1} \vec{h}' = I_{indep}/2 - I_{half}$. This indicates that the quantity being calculated is the difference between the fisher information calculated from half of the available population against half the information calculated from the entire population. Of course in this current situation where we assumed a purely independent Σ_0 there is no such difference as there is no additional change in information due to correlations between the two population halves ($\Sigma_{0,YZ} = O$ here).

In the general case, rather than I_{indep} exactly, the quantity $\vec{f}_{tr}'^T \Sigma_0 \vec{f}_{tr}'$ can be described as the projection of the tuning curve derivatives onto the data after a transformation into an independent subspace, as if the whitening were applied only between the two subpopulations but no whitening applied within the subpopulations. To see this we note $\Sigma_{0,ind}^{-1} \Sigma_0 \Sigma_{0,ind}^{-1} =$ Var$(\Sigma_{0,ind}^{-1} X)$, where $X \sim \mathcal{N}(\mu, \Sigma_0)$ (this follows from Var$(ax) = a^2$Var(x)). Then we observe $\vec{f}'^T$Var$(\Sigma_{0,ind}^{-1} X) \vec{f}' = \vec{f}_{tr}'^T \Sigma_0 \vec{f}_{tr}'$.

Establishing bounds on recovered information-limiting noise values

Limiting the influence of the above mentioned error term is necessary in order for the split population analysis to return the ground truth ϵ noise, and in this section we shall endeavor to establish a bound on this error. **Assumption 1**: we will need to make the assumption that the recorded neural population has some base level private neuronal noise, σ_{min}. This guarantees that all of the covariance matrices we work with are strictly positive definite (rather than merely semidefinite), which ensures that all of the quadratic forms encountered are strictly positive (e.g. $\vec{a}^T M \vec{a} > 0$), and that all eigenvalues are also strictly positive. **Assumption 2** will be to also carry on the assumption from the previous section that the subpopulation splits have identical statistical properties such that $I_{0,Y} = I_{0,Z} = I_{half}$.

We know that a lower bound for $\frac{\vec{f'}^T\Sigma_{0,ind}^{-1}\Sigma_0\Sigma_{0,ind}^{-1}\vec{f'}}{I_{half}}$ must be a positive value since the numerator of the fraction is a quadratic form on positive definite matrices (so by assumption 1 it is greater than 0), and the denominator is an information value (also non-negative as a rule, though in this case strictly greater than 0 once again due to the first assumption). We therefore naturally have a lower bound of $0 \leq \frac{\vec{f'}^T\Sigma_{0,ind}^{-1}\Sigma_0\Sigma_{0,ind}^{-1}\vec{f'}}{I_{half}}$. Which enables us to also write

$$\begin{aligned}\vec{g'}^T\Sigma_{0,Y}^{-1}\Sigma_{0,YZ}\Sigma_{0,Z}^{-1}\vec{h'} &= 0.5\left(\vec{f'}^T\Sigma_{0,ind}^{-1}\Sigma_0\Sigma_{0,ind}^{-1}\vec{f'} - 2I_{half}\right)\\ &\geq 0.5\left(0 - 2I_{half}\right)\\ &\geq -I_{half}\end{aligned}$$

this naturally provides $-\frac{1}{I_{half}} \leq \frac{\vec{g'}^T\Sigma_{0,Y}^{-1}\Sigma_{0,YZ}\Sigma_{0,Z}^{-1}}{I_{half}^2}$. We can run a similar argument for an upper bound U, that is, assuming $\frac{\vec{f'}^T\Sigma_{0,ind}^{-1}\Sigma_0\Sigma_{0,ind}^{-1}\vec{f'}}{I_{half}} \leq U \implies \vec{f'}^T\Sigma_{0,ind}^{-1}\Sigma_0\Sigma_{0,ind}^{-1}\vec{f'} \leq UI_{half}$.

$$\begin{aligned}\vec{g'}^T\Sigma_{0,Y}^{-1}\Sigma_{0,YZ}\Sigma_{0,Z}^{-1}\vec{h'} &= 0.5\left(\vec{f'}^T\Sigma_{0,ind}^{-1}\Sigma_0\Sigma_{0,ind}^{-1}\vec{f'} - 2I_{half}\right)\\ &\leq 0.5\left(UI_{half} - 2I_{half}\right)\\ &\leq \left(\frac{U-2}{2}\right)I_{half}\end{aligned}$$

This provides our general bounds:

$$\begin{aligned}-\frac{1}{I_{half}} \leq \frac{\vec{g'}^T\Sigma_{0,Y}^{-1}\Sigma_{0,YZ}\Sigma_{0,Z}^{-1}\vec{h'}}{I_{half}^2} &\leq \left(\frac{U-2}{2}\right)\frac{1}{I_{half}}\\ \implies -1 \leq \frac{\vec{g'}^T\Sigma_{0,Y}^{-1}\Sigma_{0,YZ}\Sigma_{0,Z}^{-1}\vec{h'}}{I_{half}} &\leq \frac{U-2}{2}\end{aligned}$$

if we assume, for example, that $U = 4$, we obtain

$$-1 \leq \frac{\vec{g'}^T\Sigma_{0,Y}^{-1}\Sigma_{0,YZ}\Sigma_{0,Z}^{-1}\vec{h'}}{I_{half}} \leq 1$$

and the error term reveals itself as a correlation value (discussed further below).

In order to obtain control over U, we will primarily be examining the numerator of the error term, $\vec{g'}^T\Sigma_{0,Y}^{-1}\Sigma_{0,YZ}\Sigma_{0,Z}^{-1}\vec{h'}$. It is at this juncture that we wish to point out the remarkable similarity our error term $\left(\frac{\vec{g'}^T\Sigma_{0,Y}^{-1}\Sigma_{0,YZ}\Sigma_{0,Z}^{-1}\vec{h'}}{I_{0,Y}I_{0,Z}}\right)$ has to the so-called "canonical correlation": $\rho = \frac{a^T\Sigma_W^{-1/2}\Sigma_{W,V}\Sigma_V^{-1/2}b}{\sqrt{a^Ta}\sqrt{b^Tb}}$. The similarities in the two expressions are in some sense unsurprising as canonical correlation analysis (CCA) is interested in assessing the correlations between two distinct datasets, a scenario that our split-trial method constructs by design. This connection and its repercussions will be explored further in the next section. For now, we shall primarily focus on establishing a finite upper bound to our error term.

We first will rewrite $\vec{g}'^T \Sigma_{0,Y}^{-1} \Sigma_{0,YZ} \Sigma_{0,Z}^{-1} \vec{h}'$ as $\vec{g}'^T \Sigma_{0,Y}^{-1/2} \Sigma_{0,Y}^{-1/2} \Sigma_{0,YZ} \Sigma_{0,Z}^{-1/2} \Sigma_{0,Z}^{-1/2} \vec{h}'$. We can then obtain

$$\begin{aligned}
\vec{g}'^T \Sigma_{0,Y}^{-1} \Sigma_{0,YZ} \Sigma_{0,Z}^{-1} \vec{h}' &= \vec{g}'^T \Sigma_{0,Y}^{-1/2} \Sigma_{0,Y}^{-1/2} \Sigma_{0,YZ} \Sigma_{0,Z}^{-1/2} \Sigma_{0,Z}^{-1/2} \vec{h}' \\
&\leq ||\Sigma_{0,Y}^{-1/2} \Sigma_{0,YZ} \Sigma_{0,Z}^{-1/2}||_{op} \cdot ||\Sigma_{0,Y}^{-1/2} \vec{g}'|| \cdot ||\Sigma_{0,Z}^{-1/2} \vec{h}'|| \\
&\leq ||\Sigma_{0,Y}^{-1/2} \Sigma_{0,YZ} \Sigma_{0,Z}^{-1/2}||_{op} \cdot \sqrt{I_{0,Y}} \cdot \sqrt{I_{0,Z}} \\
&\leq \rho_{max} \cdot I_{half}
\end{aligned}$$

where the second line is an application of Cauchy-Schwarz,$|| \quad ||_{op}$ is the operator norm, and ρ_{max} is the maximum canonical correlation between subpopulations Y and Z. The equivalence between the canonical correlations and the operator norm in the last line comes from established results on CCA and singular values as shown in [49]. From here we can establish $\frac{\vec{g}'^T \Sigma_{0,Y}^{-1} \Sigma_{0,YZ} \Sigma_{0,Z}^{-1} \vec{h}'}{I_{half}} \leq \rho_{max} \leq 1$ and we conclude $-1 \leq \frac{\vec{g}'^T \Sigma_{0,Y}^{-1} \Sigma_{0,YZ} \Sigma_{0,Z}^{-1} \vec{h}'}{I_{half}} \leq 1$. Assuming $I_{half} \to \infty$ as $n \to \infty$ (true by hypothesis as we are working with the non information-limited matrices Σ_0 here), gives that the error term $\frac{\vec{g}'^T \Sigma_{0,Y}^{-1} \Sigma_{0,YZ} \Sigma_{0,Z}^{-1} \vec{h}'}{I_{half}^2} \to 0$.

Limitations and connections to canonical correlation analysis (CCA)

The bounds on our error term we identified above appear to in some sense measure the (non information-limiting) correlations between our two subpopulations. The appearance of the canonical correlations in the calculation of the bound provides us the insight that what is being measured more specifically is the extent to which the maximal singular values in both subpopulations align. Since these calculations specifically concern the covariance matrices built out of Σ_0, the directions considered are not the information-limiting ones (typically considered the direction that is tangent to the signal response manifold). Put another way, the magnitude of our error term is maximal when the (decoding) error distributions from both subpopulations is perfectly aligned/correlated.

Because the split-trial technique relies upon examining correlated error between two subpopulations to estimate the information-limiting noise (parameterized as shared noise across the total population), any highly aligned noise mode across the two subpopulation splits would be counted in the estimation of the information limit. This also explains why our technique displays an error term even when the decoder is given full information of the covariance structure, the inference being performed on the individual population decoding errors cannot leverage the known cross-population structure because neither decoder ever has access to the information.

As a result, a limitation of our technique is that it cannot leverage any given information on the correlation structure between the subpopulations used for a given split. Additionally, highly aligned non information-limiting noise modes can corrupt our estimate, even if they do not lie along the differential direction. In practice these issues are mitigated by repeating the procedure multiple times to obtain many different population splits and averaging across the estimates obtained. Across multiple repeats, any correlation structure between neurons will be captured within a given split, and any highly aligned noise modes which appear consistently across different population shuffles would be appropriately counted as shared information-limiting noise.

C3. Further comparison to the standard approach based on Bayesian decoding

Astute readers may note that the split-trial technique's estimate of the information limiting noise generally takes the form of $\frac{1}{I_{private}} + \epsilon$, which is of the same form as that of the optimal decoder's estimate of the same quantity [18]. In this section we will provide some points of discussion that make the split-trial method preferable to standard decoding based approaches for the task of estimating information-limiting noise.

First, we noted in previous sections that our fluctuations around ϵ are dependent on the alignment of the (high-dimensional) error distributions between both subpopulation splits. The deviations away from the ground truth ϵ are dependent on how the private noise for the neural population (recall the error term is based on covariance matrices with no information-limiting noise) happens to coincidentally be correlated between both subpopulations. The worst case for the error term occurs when both subpopulations exhibit perfect correlation (i.e. the canonical correlation $\rho = 1$). If this were to occur systematically, then such coordinated error fluctuations would be nothing short of information limiting noise, which by hypothesis is contained entirely within the ϵ term, and so can be dismissed. This could,in principle, occur by random chance for the particular neurons chosen for the subpopulation splits, however this is highly unlikely and by averaging over many different subpopulation permutations would have minimal impact. Additionally, it is a known result that two random vectors in high dimension are almost surely orthogonal, so this error term will, in general, be very small.

Second, arguments of the minimality of the split-trial error term aside, the chief advantage of the technique over standard optimal decoders is that it offers an explicit estimate of the information limiting noise at finite neural populations. Even if one were to make a Cramer-Rao bound based argument for the standard approach, the simple decoders would not be able to offer an estimate of information limiting noise until the variance of the decoder error (as neural population size increases) saturated. The advantage of the split-trial technique, then, is that it offers a putative information limiting noise correlation value at finite population sizes with confidence intervals. This allows experimenters to decide heuristically for themselves whether to accept an information limit based on the level of confidence the method gives for the estimate. Such decisions are not available short of several thousand recorded neurons using standard approaches typically [22].

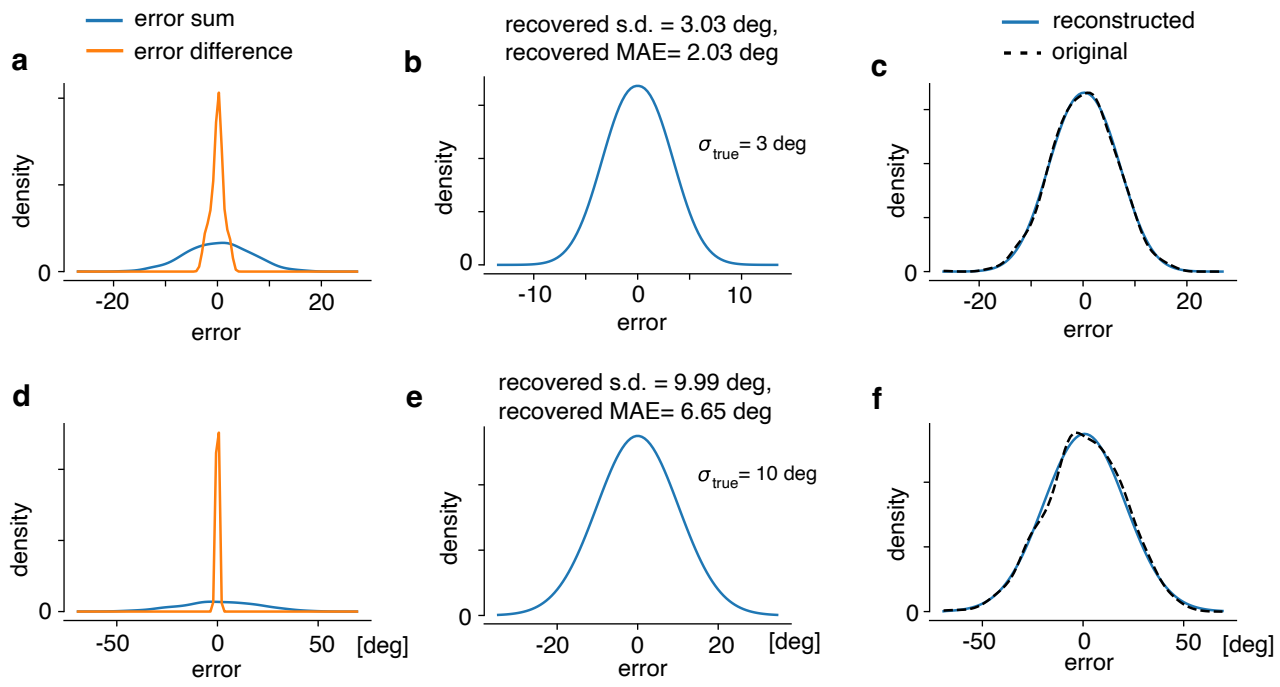

Figure S1: Validation of deconvolution based approach for estimating information-limiting noise. Data is simulated in similar schemes as **Fig. 3**, but the information limiting σ is estimated using a deconvolution approach. *Top row:* data generated with a ground truth information-limiting $\sigma = 3$, *bottom row:* data generated with a ground truth information-limiting $\sigma = 10$. **(a,d)** The sum and difference of errors across both splits. **(b,e)** The estimated error distributions for the two datasets, with the estimated σ and MAE listed. **(c,f)** Using the deconvolving filter estimating the error distribution, we can reconstruct the original error sum distribution (solid blue) and compare it against the actual original distribution (dashed).

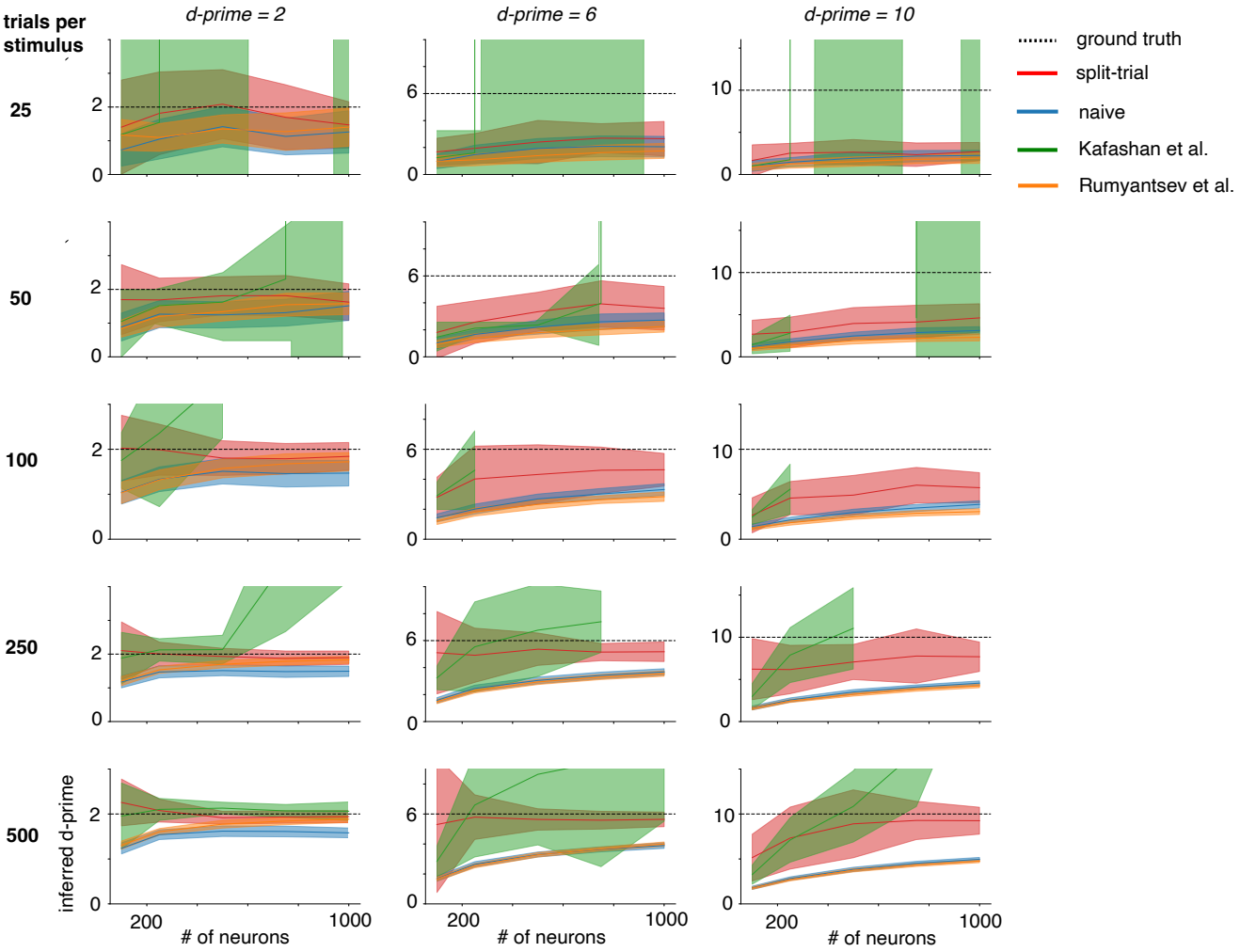

Figure S2: Benchmarking the performance of split-trial analysis and three alternative methods on inferred the ground truth d-prim. This figure shows the results for all sample/stim and cell population combinations used (related to Fig. 4). Note that the portions of the curves for which the method from [18] disappears due to a preponderance of unbounded information estimates.

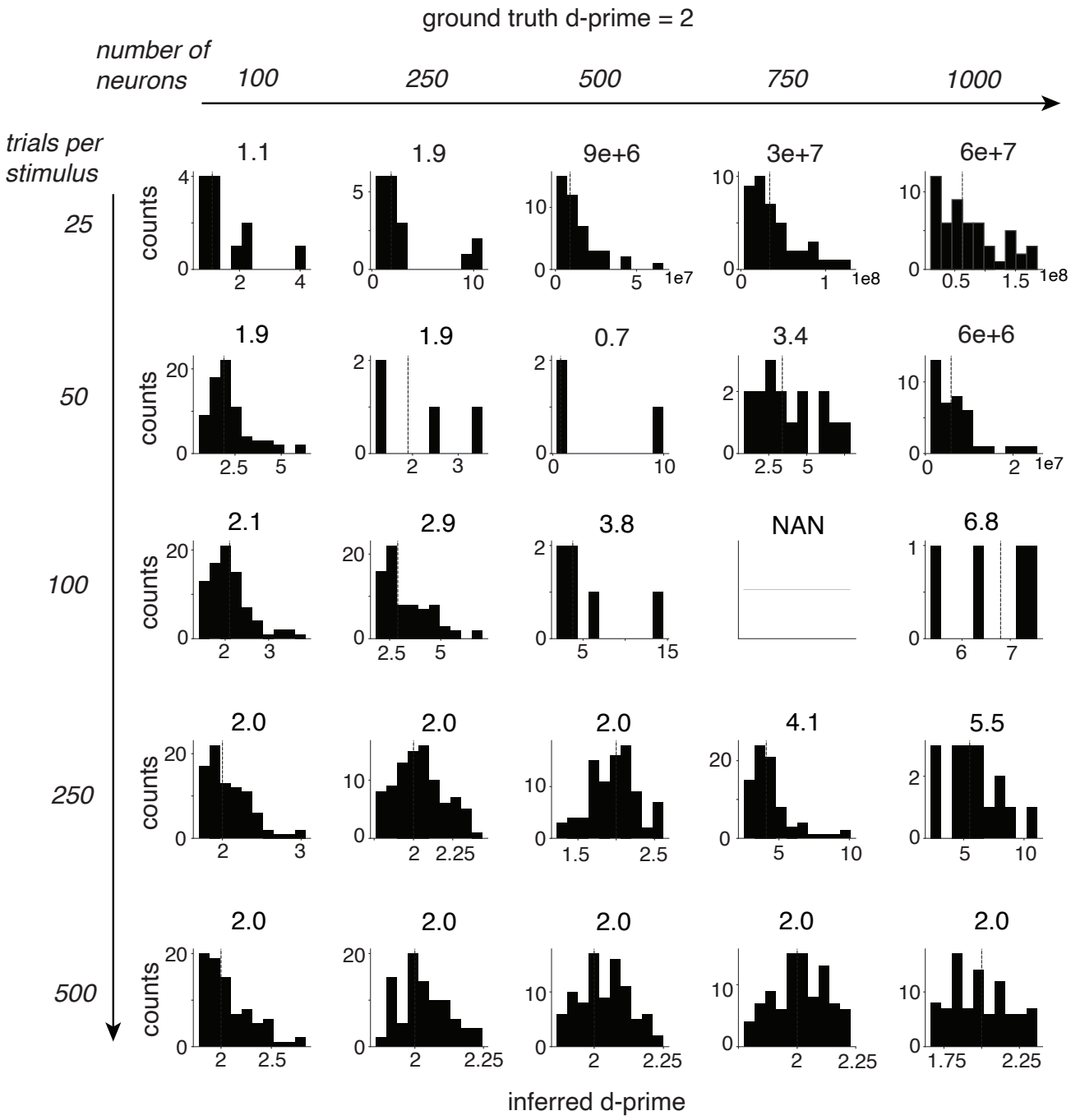

Figure S3: Histograms of estimated d-prime limit for the method from [18] based on extrapolation. The ground truth d-prime in this case is 2. For each panel, the median of the estimated d-prime values across 100 simulations was shown on the top. Note that, when there are few samples, the linear extrapolation method is highly unstable. Additionally, certain combinations of large cell count and large samples may be biased towards classifying a population as having unbounded information. This is likely due to the particular bias-corrected fisher information calculation developed used in the original report. When the estimated information is unbounded, such estimates were not shown in the plots (thus the sum of the counts in the histogram may be smaller than 100). In some case, none of the 100 simulations results in a finite estimate of the d-prime values, resulting in all NANs.

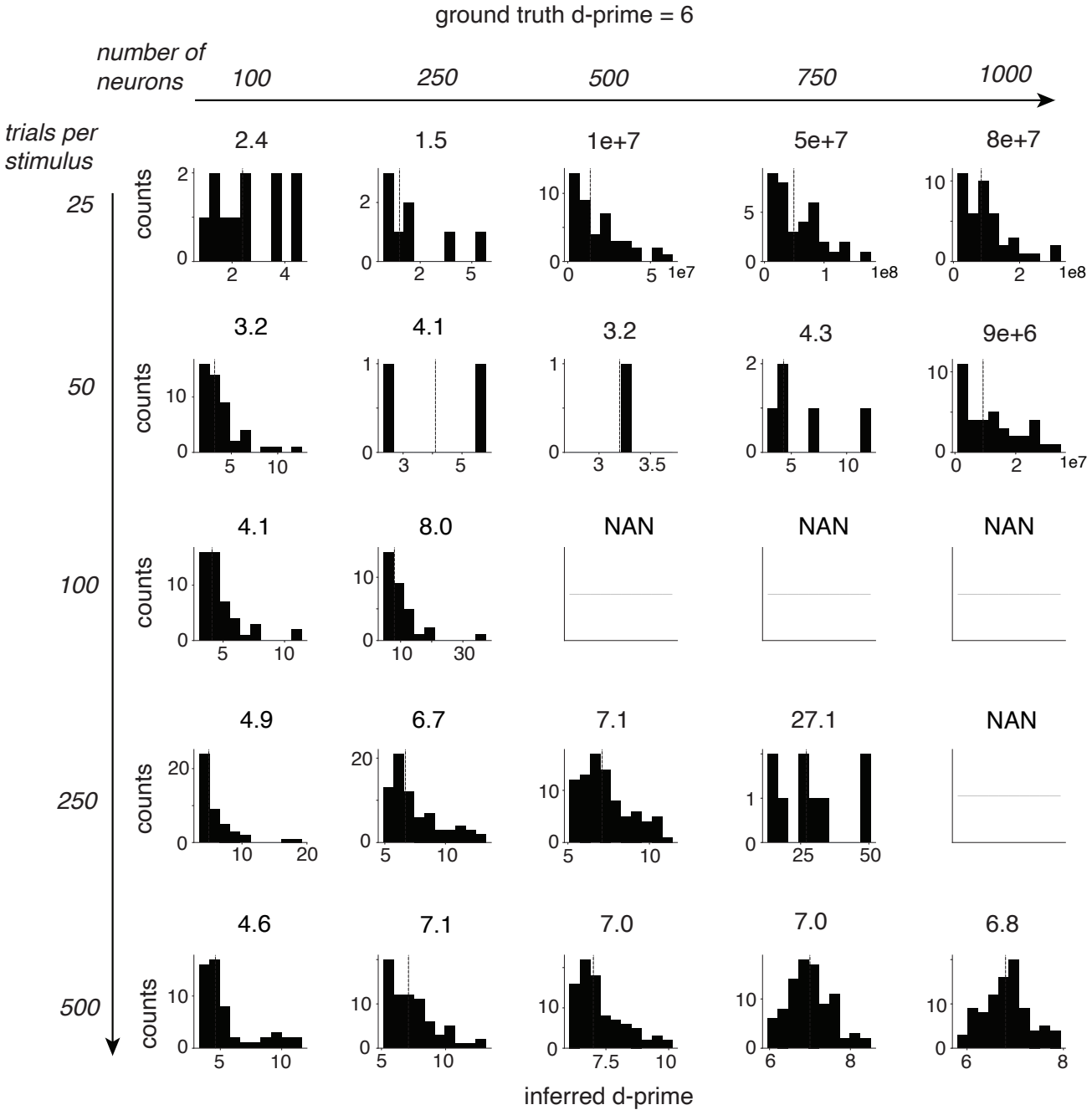

Figure S4: Histograms of estimated d-prime limit for the method from [18]. The ground truth d-prime in this case is 6. Same convention as Fig. S3.

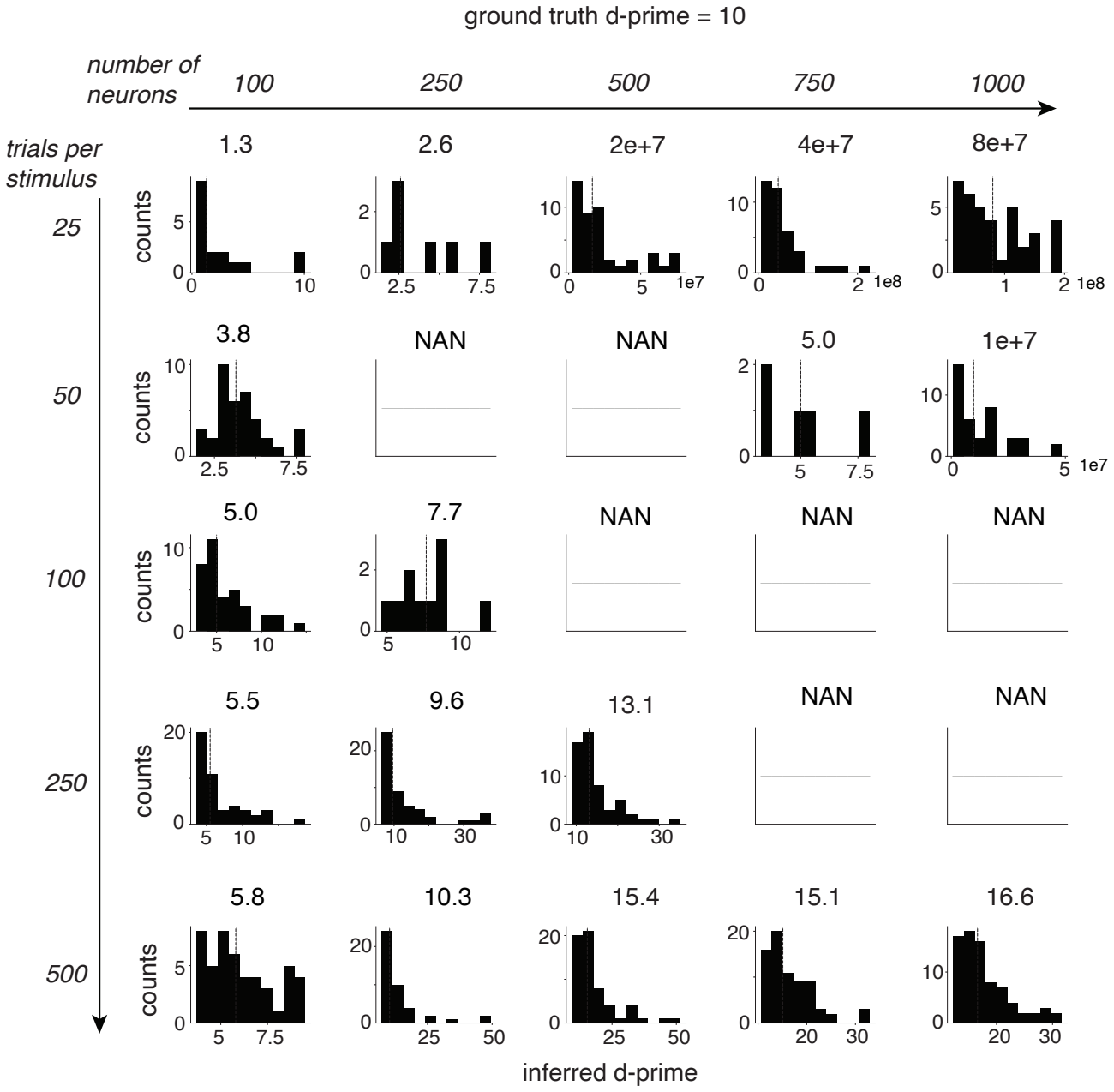

Figure S5: Histograms of estimated d-prime limit for the method from [18]. The ground truth d-prime in this case is 10. Same convention as Fig. S3.

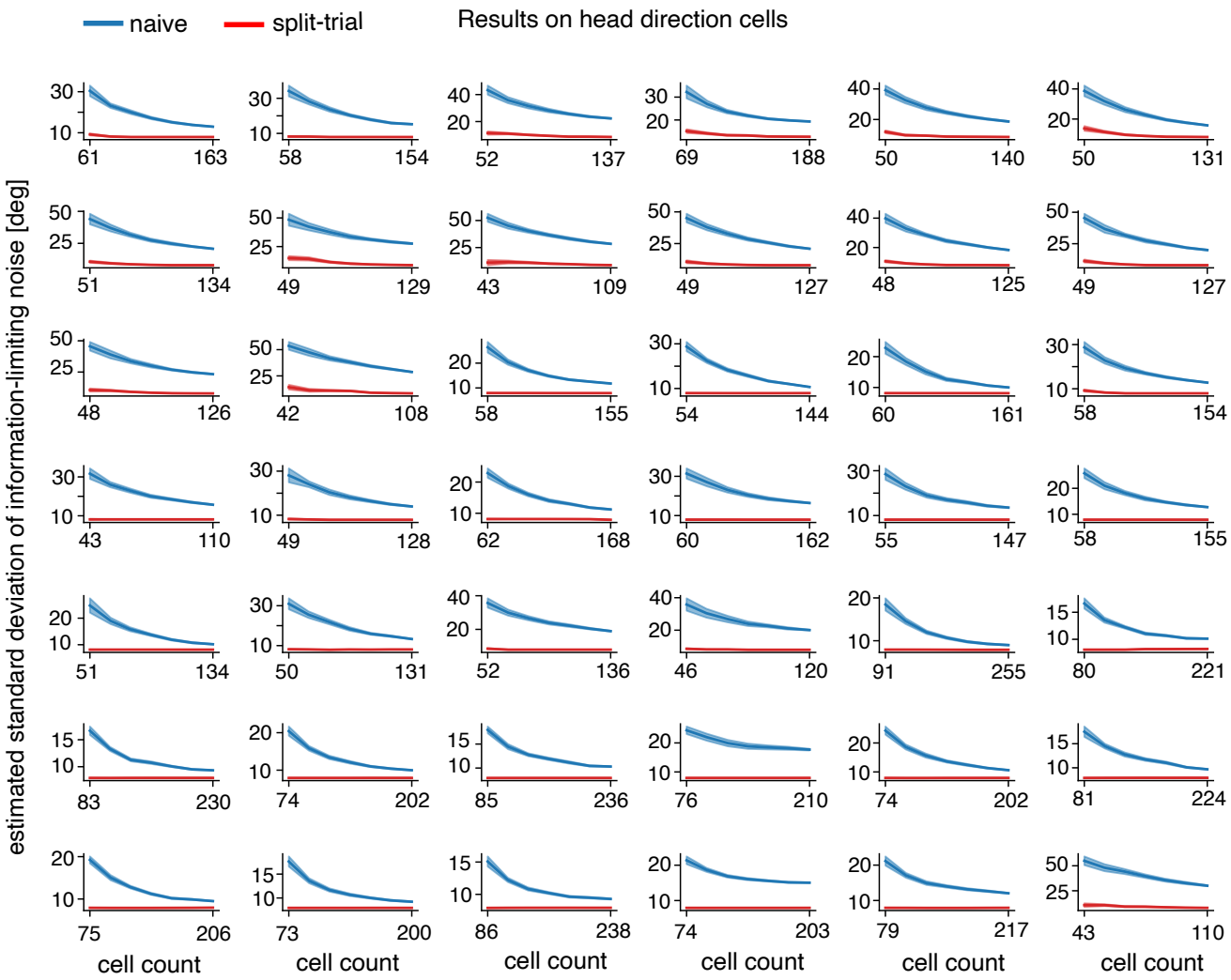

Figure S6: Individual session estimates for the information-limiting σ from [27] for both split and naive approaches
Plots were started from intermediate population sizes for visualization purposes.

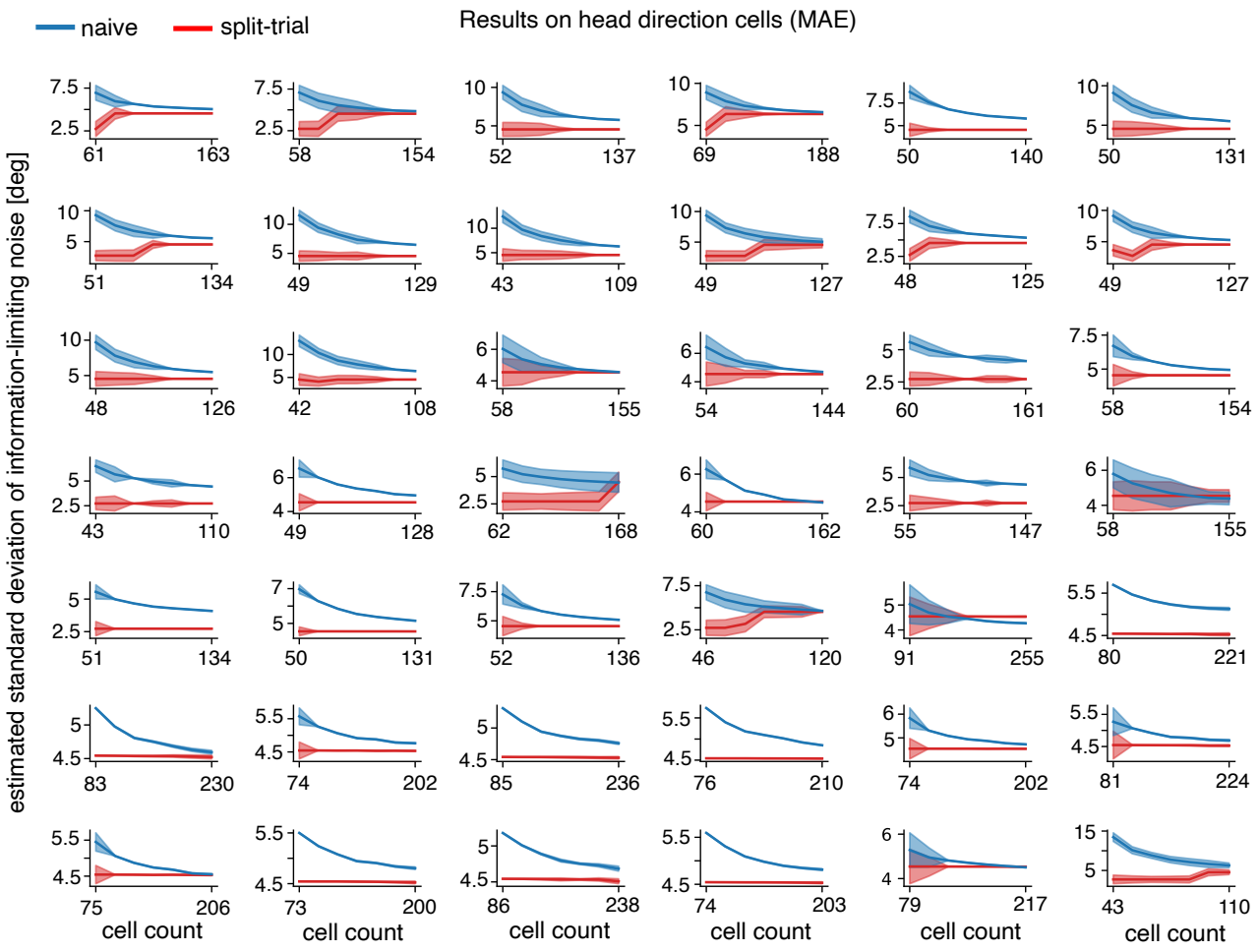

Figure S7: Individual session estimates for the median absolute error (MAE) of information-limiting niose from [27]. Red: results based on split-trial analysis based on the deconvolution approach. Blue: results based on direct decoding using a ZIG Bayesian decoder[34]. Plots were started from intermediate population sizes for visualization purposes.

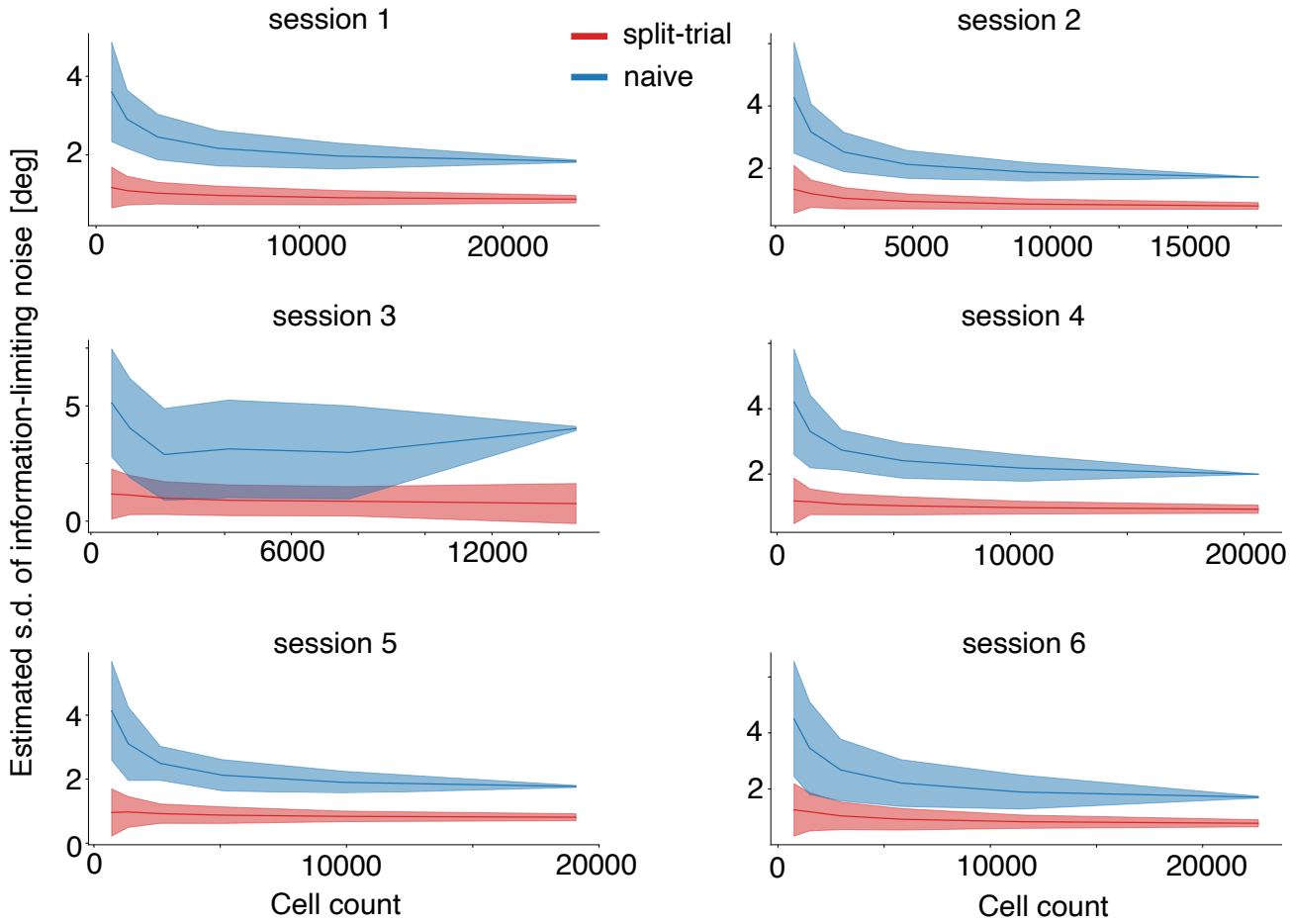

Figure S8: Results on the inferred standard deviation of information-limiting noise from the uniformly sampled orientation condition. Results from 6 individual sessions were shown. Red: estimates from split-trial analysis. Blue: estimates from the direct (naive) decoding analysis. We observed that the naive estimate of σ for session 3 increases slightly after $\sim$ 10,000 cells. The odd level of variability in this session made display challenging so was excluded from the visualization of Fig. 6c. Error bars are standard deviation over all iterations.

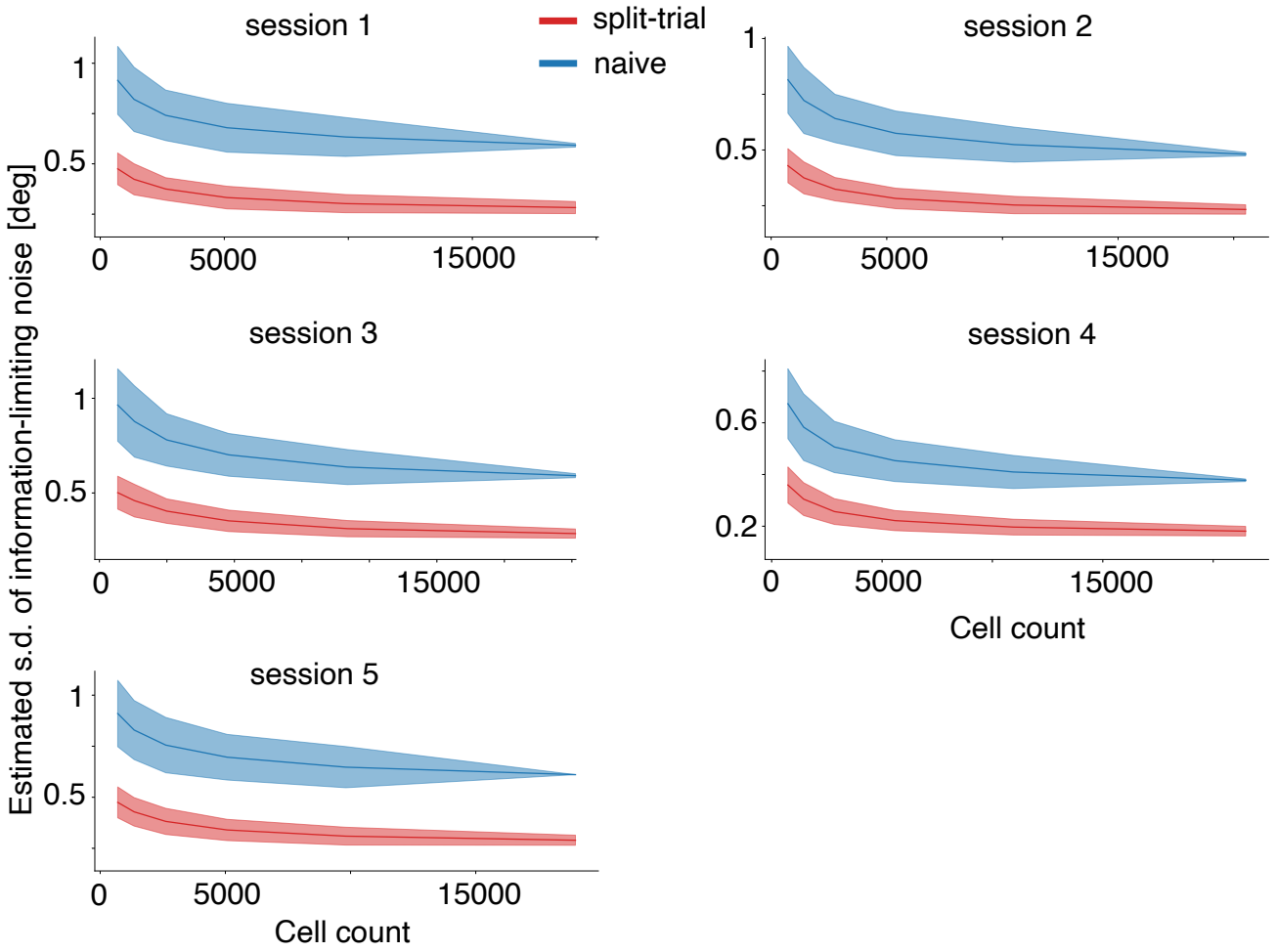

Figure S9: Results on the inferred standard deviation of information-limiting noise from the densely sampled orientation condition. Red: estimates from split-trial analysis. Blue: estimates from the direct (naive) decoding analysis.

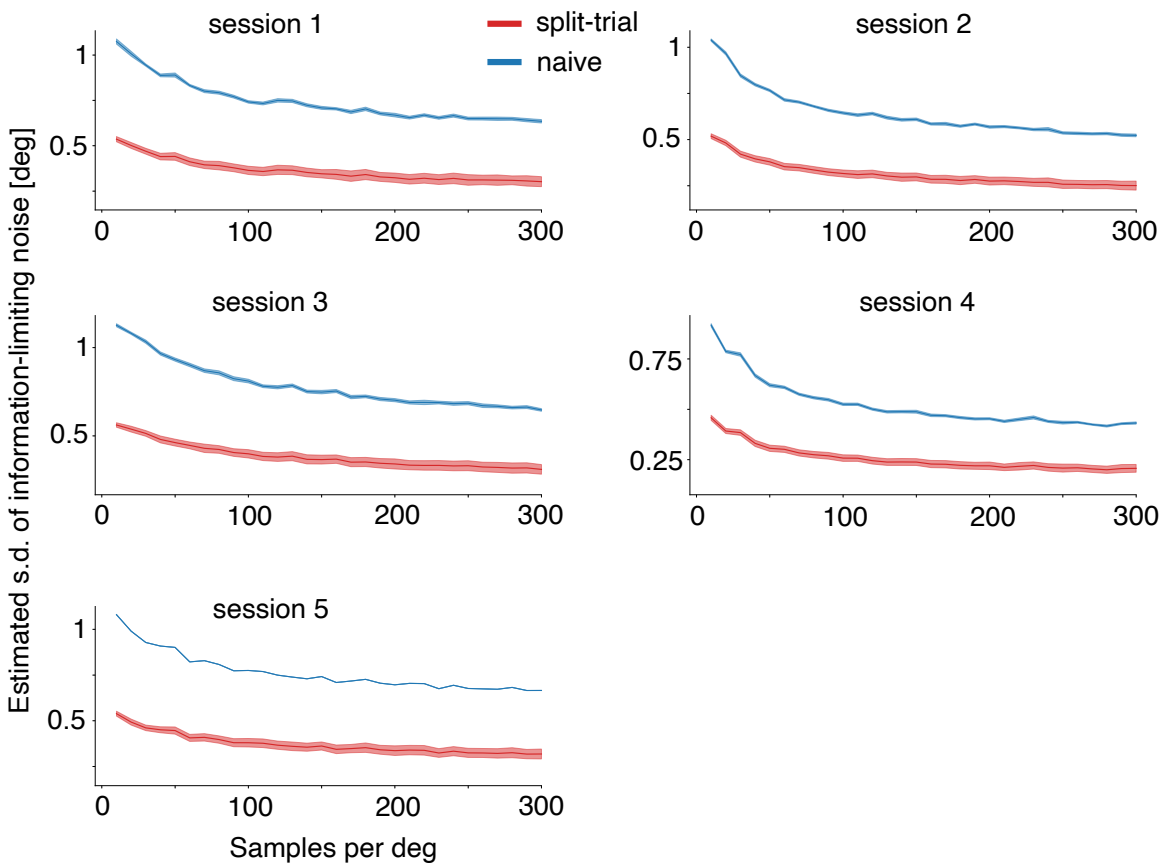

Figure S10: Results on the inferred standard deviation of information-limiting noise from the densely sampled orientation condition when the dataset was downsampled. Data from the dense discrimination sessions was downsampled to a range of different stimulus densities, up to 300, which is about the full dataset density. Notably, the split-trial estimates are stable and no longer appear to decrease after ∼150 samples/deg, suggesting that additional experimental trials would not affect the estimate.

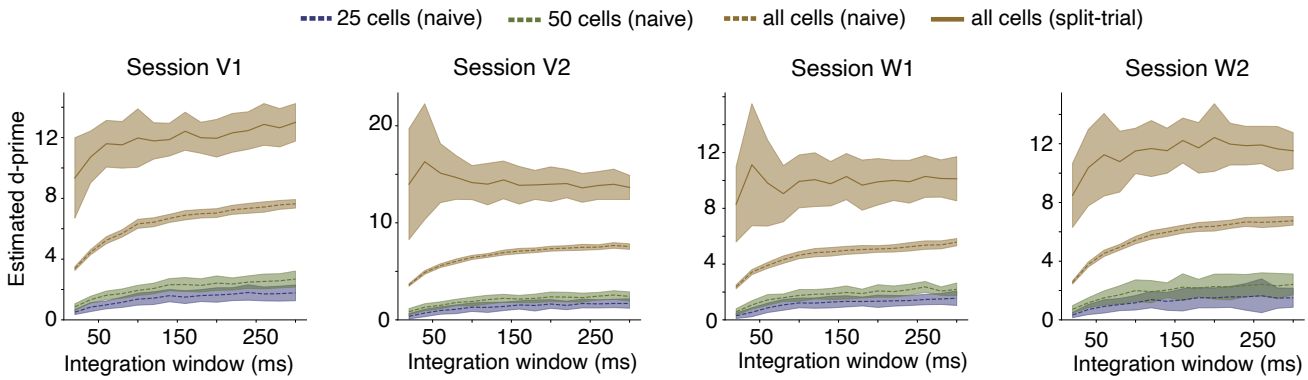

Figure S11: Results on how d-prime values scale with the increasing integration windows. Results from the split-trial analysis shown as solid line, showing relatively little change due to temporal integration. The four sessions have 770, 828, 509, 529 neurons, respectively.

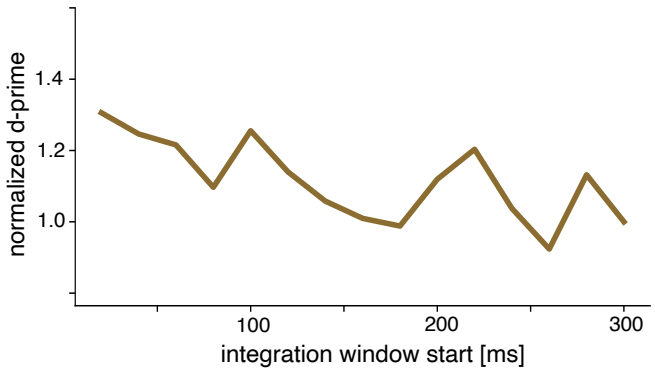

Figure S12: Average normalized d-prime values across normalized sessions (n=4) when windows are 20ms wide and discrete steps are taken through the 300ms window. D-prime values were inferred using split-trial analysis based on all neurons from individual sessions. The results show a trend that indicates information-limiting noise increases as reward delivery time approaches.

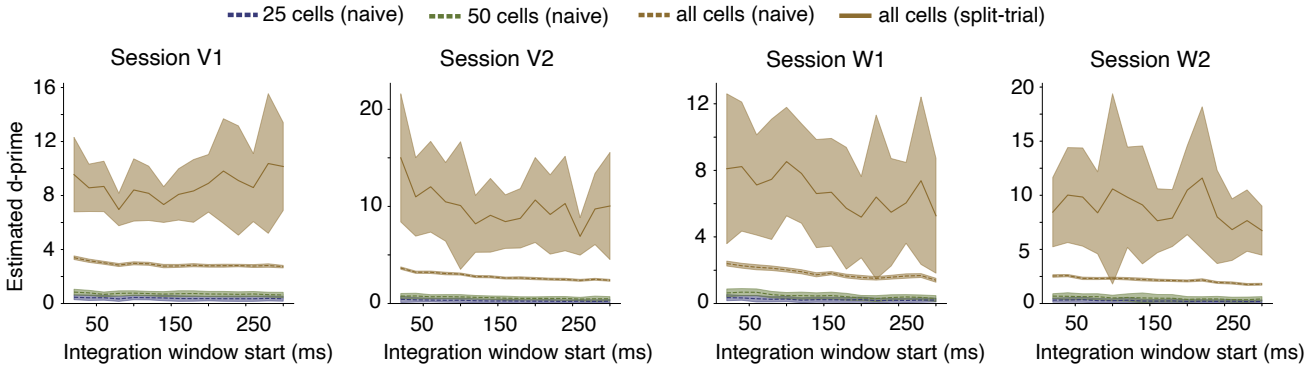

Figure S13: Non-normalized version of discrete windows individual sessions from the dataset from [20]. Despite early windows having higher values than later windows, they are typically lower than the (final) values recovered from the integrated calculations.