In V4, pair-wise synchrony and correlations depend on the structure of the population code

Veronika Koren[1,2]*, Ariana R. Andrei[3], Ming Hu[4], Valentin Dragoi[3], Klaus Obermayer[1,2]

1 Neural Information Processing Group, Institute of Software Engineering and Theoretical Computer Science, Technische Universität Berlin, Berlin, 10587, Germany **2** Bernstein Center for Computational Neuroscience Berlin, Germany **3** Department of Neurobiology and Anatomy, University of Texas Medical School, Houston, Texas, 77030, US **4** Picower Institute for Learning and Memory, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139, US

*Correspondence: koren@ni.tu-berlin.de

Highlights

- In a match-to-sample visual task, responses of neural populations in V1 and in V4 predict the stimulus class better than chance.
- In V4, informative neurons are more strongly coupled, correlated and synchronized than uninformative neurons.
- In V4, neurons are more strongly coupled, correlated and synchronized within coding pools compared to across coding pools.
- Correlations within coding pools harm the performance of the classifier in both V1 and V4.

Abstract

In visual areas of primates, neurons activate in parallel while the animal is engaged in a behavioral task. In this study, we examine the structure of the population code while the animal performs delayed match to sample task on complex natural images. The macaque monkeys visualized two consecutive stimuli that were either the same or different, while recorded with laminar arrays across the cortical depth in cortical areas V1 and V4. We decoded correct choice behavior from neural populations of simultaneously recorded units. Utilizing decoding weights, we divide neurons in most informative and less informative, and show that most informative neurons in V4, but not in V1, are more strongly synchronized, coupled and correlated than less informative neurons. As neurons are divided in two coding pools according to their coding preference, in V4, but not in V1, spiking synchrony, coupling and correlations within the coding pool are stronger than across coding pools.

Keywords: monkey, visual cortex, choice behavior, classification, noise correlations, neural coding, population code, information

Introduction

The general understanding is that sensory neurons respond to features of sensory stimuli and transmit this information downstream. Downstream neurons read-out activity of projecting neurons and use it for generating a signal that is highly correlated with the behavioral choice of the animal ([Shadlen and Newsome, 2001]). Traditionally, stimuli in choice probability paradigms targeted a specific feature of the stimulus that a given brain area is known to respond to ([Newsome et al., 1989, Zohary et al., 1994, Britten et al., 1996, Shadlen and Newsome, 2001]). While cited studies made an important breakthrough in our understanding of the function of the cortical activity, we can further improve our understanding of the neural function by studying responses of neural ensembles to complex natural stimuli. In this study, we ask if classification of complex natural images can be read-out from the activity of neural ensembles in the early (V1) and middle-level visual cortex (V4).

The coding function of neural responses in the visual cortex is often described as a set of filters ([Rao and Ballard, 1999]), where the response of each single neuron acts as a filter of the specific feature of a stimulus (e.g., the orientation). In contrast to such rigid coding schemes, animals use sensory information in a flexible way, showing the ability to manipulate the information with temporal delays and to classify stimuli with respect to abstract rules. Is such flexibility an exclusive property of higher brain areas, or can it be observe already in sensory areas? We propose to study responses of neural ensembles to naturalistic stimuli, where the classification task does not rely on filtering properties of single neurons, but instead requires discrimination of abstract classes of stimuli. In particular, the animal visualizes two stimuli consecutively, interleaved with a delay, and decides whether they are the same ("match") or different ("non-match"). Since the identity of stimuli changes from one trial to another, correct classification cannot rely on encoding of a particular orientation.

The way neural populations encode task-relevant information could either be a by-product of the activity of single units or else have an intrinsic structure that cannot be explained by the response properties of single units ([Elsayed and Cunningham, 2017]). The organism has to generate behavioral decisions in every trial while the trial-to-trial variability of cortical neurons makes the signal conveyed by single neurons unreliable. Summing activities across neurons can, in general, overcome the variability problem, but only in case where the variability is uncorrelated across neurons ([Britten et al., 1992]). This clearly does not hold for local neural ensembles that share a big proportion of their inputs ([Cohen and Kohn, 2011, Smith and Kohn, 2008]). If neurons are correlated, there should be a non-trivial relation between activities of single neurons and the population code. Recent theoretical and modeling work has shown that the variability of neural responses can be reconciled with deterministic population signals ([Koren and Denève, 2017, Boerlin et al., 2013]). This is possible if we assume that task variables (such as the stimulus and the behavioral choice) are represented by the network as a low-dimensional signal, and the coding of such a signal is distributed across a multitude of neurons with redundant coding function. While coding of low-dimensional variables with spiking neural networks has been laid down in theory, the relation between the activity of single neurons and the signal that is effectively transmitted downstream and contributes to behavior, remains elusive in biological networks of behaving animal agents.

Over the last decade, two major insights have been made on the principles of neural coding in cortical networks. The first insight concerns the relation between the similarity of the tuning function of a pair of neurons (signal correlation) and their common variability across trials (noise correlation). If a pair of neurons has similar tuning, positive noise correlations will harm the information transfer. Vice-versa, if a pair of neurons has dissimilar tuning, positive noise correlation can enhance the quantity of the transferred information

([Averbeck et al., 2006]). In order to maximize the information transfer, neurons with similar tuning therefore should have low correlations, while neurons with dissimilar tuning should have stronger correlations. However, in the brain, tens of thousands of neurons activate during a visual discrimination task. Neural signals are effective only if they are transmitted downstream, and the activity that maximizes the information transfer might result in a neural signal that cannot be reliably read-out by a downstream neuron. The second insight on neural coding and dynamics comes from the observation of neural activity in decision-making paradigms. While many neurons activate during a visual task, in experiments, we only observe a tiny fraction of these. Surprisingly, this small subset of neurons appears to be informative about the choice behavior of animal agents. This can be explained with the existence of coding pools, where a (large) number of neurons share a similar coding function. If within the coding pool neurons are more strongly correlated than across pools, the activity of a randomly chosen neuron from the pool has a high probability of being informative about the behavioral variable. In short, if correlations within the pool are stronger than across pools, such a correlation structure allows to infer task variables from the observation of only a small number of neurons ([Shadlen, 1996, Nienborg et al., 2012]). At the first glance, these two ideas imply divergent views on the role of correlations for coding of behavioral variables, since correlations that harm the information transfer are also those that should be the strongest.

In the present work, we study the coding function of the neural population and relate it to correlations. We assume that the computation that the population is performing is discrimination of binary stimulus classes. We compute the coding function of a single neuron relatively to the activity of the network that surrounds the neuron and also takes into account inter-neuron interactions. We apply an optimal linear decoder that gives the upper bound of the transferred information by a linear classifier ([Vapnik and Vapnik, 1998]). We then extract the structure of population responses, given by the weights of decoding model, to study the relation between the structure of population responses and pair-wise dynamics. Our decoding scheme ensures that the relation between the structure and the dynamics is actually relevant for the behavior ([Panzeri et al., 2017]). As we relate the strength of decoding weights to pair-wise dynamics, we find a strong interplay between the two in the microcircuits in V4, but not in V1. We divide neurons in two mutually exclusive coding pools and show that in V4, but not in V1, correlations within pools are stronger than across pools. In general, the relation between the structure of the population model and correlations show many differences between V1 and V4 microcircuits, which might reflect different stages of information processing, from analysis of the visual scene (V1), towards convergent information about binary choice behavior (V4).

Results

Neurons in the brain read-out the activity of projecting units.

Two adult monkeys *Macaca mulatta* performed a match-to-sample task on visual stimuli. In every trial, the subject visualized two consecutive stimuli, with a delay period in between (Figure 1A). Stimuli were complex naturalistic images in black and white, lasted 300 ms each and were either identical (condition "match") or else the test stimulus was rotated with respect to the target stimulus (condition "non-match"). The angle of rotation changed randomly across trials and was adjusted on-line to have on average 70 % of correct responses with non-matching stimuli. While all stimuli fell into one of the two classes, "match" and "non-match", the identity of the stimuli changed from one trial to another. The task of the animal was to decide about the

similarity of the two stimuli by holding a bar for the choice "different" and releasing the bar for the choice "same". Laminar arrays with 16 recording channels were inserted perpendicularly to the surface of the brain, recording the multi-unit signal across the cortical depth in V1 and V4 visual areas (see methods; Figure 1B). In both areas, neural populations responded with a transient peak in firing rates, followed by tonic firing (Figure ??C).

Neurons and neural ensembles perform their coding function with their dynamics (e.g., by firing more for stimulus 1 than for stimulus 2), and reading-out neural activity with respect to a given low-dimensional variable, such as the stimulus class or the behavioral choice, is conceptualized as "decoding". In decision-making paradigms, decoding of neural activity tells us how accurately one can predict the future behavioral choice of the animal by observing the neural dynamics that precedes it. Here, we investigate decoding models that are readily interpretable in the biological setting, i.e., models that could be implemented in the brain itself. We adopt the view of a read-out neuron that can estimate the average firing rate from the projecting unit and calculate the probability of discriminating between the two stimulus classes, given the firing rate of the projecting unit in the single trial. We limit our investigation to correct choices, i.e., trials where the visualization of one of the two possible stimulus types, "match" and "non-match", has led to the congruent behavioral choice, "same" and "different". In correct trials, the representation of the stimulus is expected to be the primary source of information for determining the choice behavior. However, at least part of the information about the choice is likely to exist independently of the stimulus, possibly influencing sensory areas as the feedback activity from higher brain areas ([Gilbert and Sigman, 2007]). We therefore assume the existence of a mixed variable that contains the information about both the stimulus class as well as the behavioral choice (table 1).

Stimulus	Choice	Performance	Info. content
non-match	"different"	correct	stimulus + choice
match	"same"	correct	
non-match	"same"	incorrect	stimulus + performance
match	"same"	correct	

Table 1. Informational content of task variables.

Read-out of the activity of populations in V1 and in V4 predicts correct choices better than chance.

We decode the variable *stimulus+choice* from the activity of neural populations, using an optimal linear classifier, linear Support Vector Machine (SVM) [Vapnik and Vapnik, 1998]. We call this decoding model the *population model*, since it uses the activation of a population of simultaneously observed neurons (Figure 1A, see methods). The quality of prediction of the population model is assessed on hold out sets with cross-validated balanced accuracy, which is the probability of correctly predicting condition "match" vs. "non-match" in single trials. In case all test samples were classified correctly, we would have the balanced accuracy of 1. The balanced accuracy of 0.5 indicates that the validation data has been classified equally often correctly as incorrectly, which is chance level performance. Significance of the prediction accuracy is evaluated with a permutation test, using models with randomly permuted class labels (see methods).

Our results show that, during the test time window, the population model performs significantly better than chance ($p < 0.001$ in both areas, permutation test, Bonferroni correction, Figure 1D). On average, the

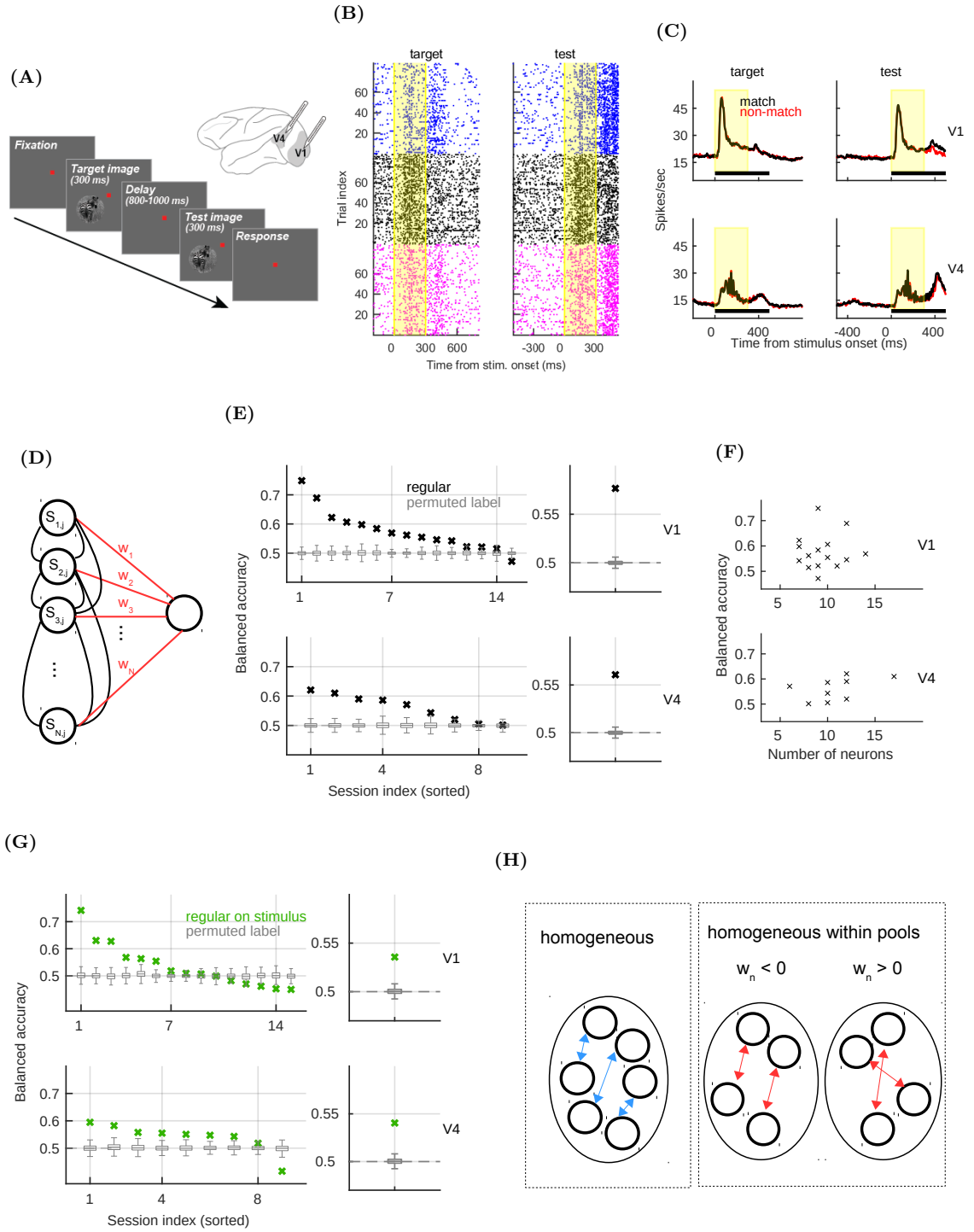

Figure 1. The Population model uses high-dimensional patterns of activation for predicting the stimulus class. A) Schema of the experimental paradigm and illustration of the macaque brain, with shaded areas indicating visual areas V1 and V4. **B)** Raster plots of three representative neurons in V4 (blue, black, magenta) in condition "non-match", during the target (left) and the test time window (right). The yellow area marks the presence of the stimulus. **C)** Population PSTH in V1 (top) and in V4 (bottom), during target (left) and test (right). We separated the activity in conditions "match" (black) and "non-match" (red). **D)** Schema of the population model. Parallel spike counts of the neural population, weighted by the weight vector, project to the read-out neuron. **E)** Left: Prediction accuracy of the population model (black) in V1 (top) and in V4 (bottom). Results have been computed in each recording sessions and have been sorted, from best to worst performance. Gray boxplots mark distribution of results with permuted class labels. Right: Session-averaged results for the regular model (black) and the model with permutations (gray). **F)** Prediction accuracy in sessions as a function of the number of neurons.

Figure 1. G) Same as in **B**, but for the classification problem with information on the stimulus class and the behavioral performance, but without the information on the behavioral choice. **H** Schema of two types of randomization of the structure of parallel spike counts. We permute the neural index across all neurons, without repetition, creating the data that is homogeneous across all neurons (left). On the right, the schema shows permutation of the neural index that is limited to neurons with the same sign of the weight (homogeneous within pools, right).

high-dimensional model predicts the choice behavior 7,7 % better than chance in V1 and 6,1 % in V4 (Figure 1E). The prediction accuracy varies considerable across recording sessions, presumably because linear arrays are inserted *de novo* in each recording session, capturing small neural ensembles that vary in their relevance for the classification task. Importantly, balanced accuracy in recording sessions is uncorrelated with the number of neurons that we observe ($p = 0.447$ in V1, $p = 0.103$ in V4, permutation test, Figure 1F). During the target time window, prediction accuracy is at chance ($BAC = 0.50$ in V1 and $BAC = 0.49$ in V4, non-significant; Figure S1A). The use of a nonlinear decoder, SVM with Radial Basis Function, did not improve the prediction accuracy with respect to the linear model (Figure S1B). Since the SVM is the optimal classifier, we conclude that the performance of the population model coincides with the amount of information that a linear read-out on parallel spike counts contain about the variable *stimulus+choice*. As we decode from conditions that differ in the stimulus class and in behavioral performance, but do not differ in the choice (e.g., with the informational content of *stimulus+performance*, table 1), we get the prediction accuracy lower than with correct choice (3,6 % above chance in V1 and 4 % above chance in V4, Figure 1G).

Prediction accuracy of the population model relies on high-dimensional patterns of activation of the neural ensemble. To get better understanding of this mechanism, we design biologically unrealistic read-out models, that serve the purpose of gaining insights into the population model. As a toy example, consider a network with 2 neurons, where neuron 1 increases its firing rate for stimulus "non-match" while neuron 2 decreases the firing rate for the same amount, and the opposite happens when stimulus "match" is shown. A read-out model that is agnostic about the identity of neurons would give chance level prediction of the stimulus, while model that weights contributions of neurons with appropriate weights can utilize the difference in these activation patterns as a source of information. To test this idea on the population model, without changing the dimensionality of the model, we create patterns of neural activity that are homogeneous across neurons. The homogeneous model is given by permuting the neural index of the vector of spike counts in every trial (Figure 1H). Notice that the homogeneous model is equivalent to an unstructured network. As we compare the performance of the homogeneous and the regular model, we see that the performance of the homogeneous model is consistently worse (Figure 2A), showing that patterns of activation of the structured network are an important source of information for discrimination.

Hand-picking the best single neuron in each recording session, on the other hand, gives performance that is higher than that of the population model (Figure 2B). Such an unrealistic read-out procedure gives, on average, prediction accuracy that is 9.6 % better than chance in both V1 and V4. If neuronal populations were uncorrelated, the SVM could simply weight strongly the most informative neuron in each session, and attribute zero weights to all other neurons. However, the SVM has to take into account inter-neuron correlations. The fact that the best single neuron is better than the population model implies that at least some correlations within the neural circuit are such as to harm the performance of the classifier. A uniform read-out, where

the prediction accuracy is computed for every neuron independently and then averaged across neurons (see methods), gives a poor prediction of *stimulus+choice*. The prediction accuracy is 2,4 % above chance in both areas (Figure 2B).

The population code has its own intrinsic structure.

The fact that the population model has higher prediction accuracy than an average single neuron suggests that the population model cannot be reduced to the sum of predictive functions of single neurons without loss of information ([Elsayed and Cunningham, 2017]). While this is expected, the question remains, how does the population model relate to activity of single neurons. The use of a linear SVM makes the comparison between two straightforward, since it allows to compute weights of the population model (see methods). The population model relies on a high-dimensional plane in the space of inputs, fully defined by the vector of weights and the offset of the plane from the origin. Weights describe the role of each neuron for the classification task, relatively to the activity of other neurons from the population and by taking into account inter-neuron interactions. An identical activity of a single neuron would results in a different decoding weight, if this neuron was surrounded by a different population. If the activity of a particular neuron has no role in discrimination, its weight will be 0. The bigger the deviation of neuron's weight from 0, the stronger the importance of the neuron for discrimination. The sign of the weight is also important to consider, since weights of the opposite sign have the opposite effect on the direction of the separating hyperplane. Computing weights of the population model for simultaneously recorded neurons and collecting results across recording sessions, we find that almost no neurons have significant weights (permutation test using weights from models trained on permuted class labels; Figure 2C, left). Pooling weights across neurons, we find that during test, positive and negative weights are balanced in V1, but unbalanced in V4, with an excess of positive weights ($p = 0.005$, permutation test; Figure 2C, right). During target, positive and negative weights are balanced in both areas (Figure S2A).

Next, we estimate the discriminatory power of single neurons as the area under the Receiver-Operating characteristic curve ($AUROC$, see methods;Figure 2D). The $AUROC$ score measures the distance between distributions of spike counts in conditions "match" and "non-match" and is the probability of predicting the condition from the observation of spike counts. Identical distributions yield $AUROC = 0.5$. Neurons with $AUROC > 0.5$ are those that fire more in condition "match" while neurons with $AUROC < 0.5$ are those that fire more in condition "non-match". We measure the $AUROC$ score for each single neuron and collect results across recording sessions. We find that a small proportion of neurons have $AUROC$ significantly different than chance (13 % in V1 and 8 % in V4, 2-tailed permutation test with 1000 permutations, Bonferroni correction; fig 2E, left). As we pool $AUROC$ scores across sessions, we find that during the test time window, there are more neurons that prefer matching than non-matching stimuli ($p < 0.001$ in V1 and in V4, permutation test on the sum of $AUROC$ scores; Figure 2E, right). The shift of the mean $AUROC$ from 0.5 is of 3% in V1 and of 4 % in V4. During the target, the distribution of AUROC scores is centered around 0.5 in both areas (Figure S2B).

Finally, we compare the strength of population weights, $|\tilde{w_n}|$ with the strength of $AUROC$ scores, $|AUROC_n - 0.5|$. We also distinguish neurons with positive weights ($\tilde{w}_n > 0$, *plus* neurons) and neurons with negative weights ($\tilde{w}_n < 0$, *minus* neurons). The strength of weights and of $AUROC$ scores is moderately (V1) to strongly (V4) positively correlated for both plus and minus neurons (Figure 2F). The strength of population weights is therefore positively correlated with the discriminatory power of single neurons. Nevertheless, the

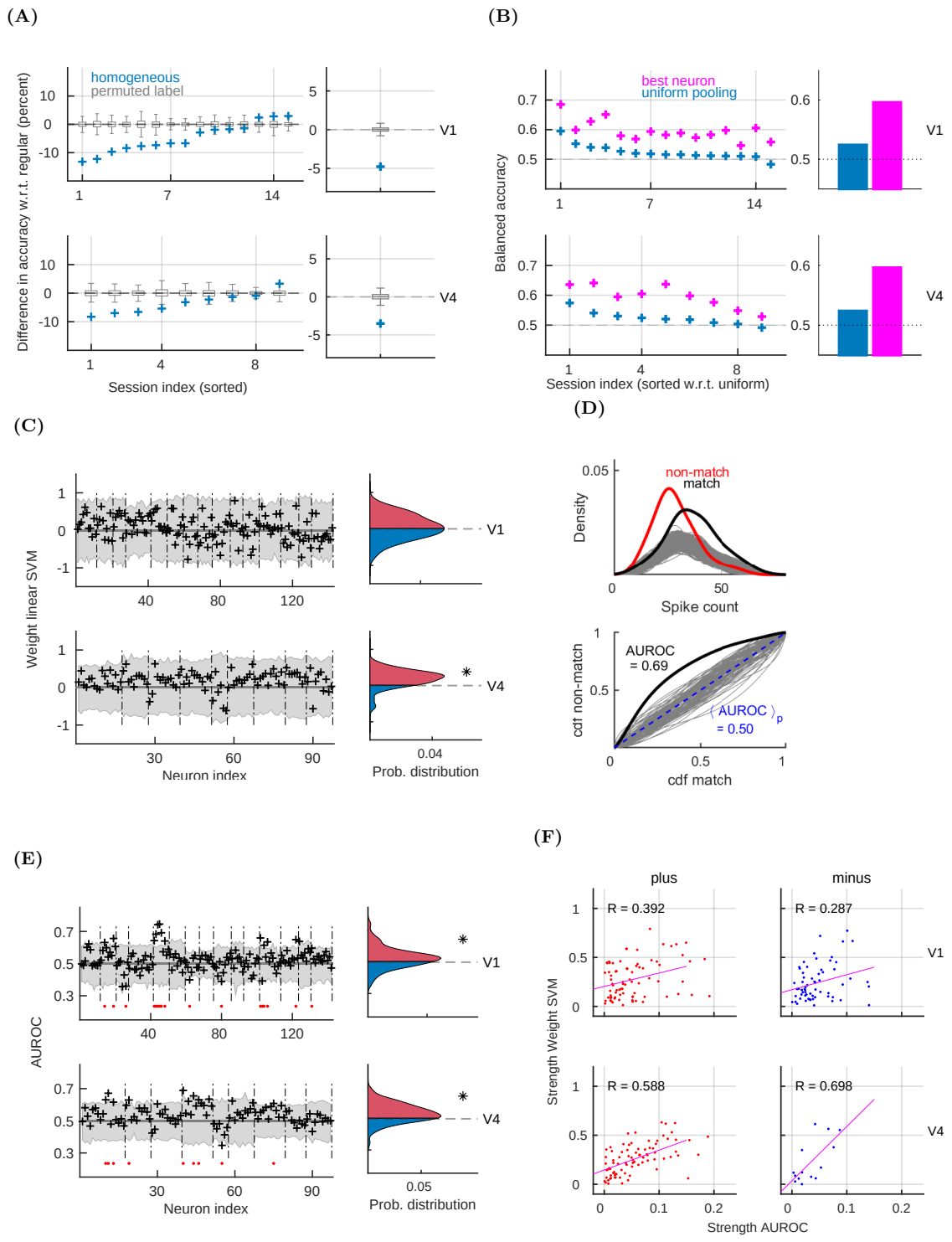

Figure 2. Decoding weights of single neurons are similar, but not identical to population decoding weights. A) Difference of the prediction accuracy of the homogeneous model w.r.t. the regular model (homogeneous-regular). Gray boxplots mark the same result for models with permuted class labels. **B)** Left: Prediction accuracy of the best neuron in each session (magenta) and of the uniform pooling (blue). Right: Session-averaged results. **C)** Left: Weights of the population model (black), collected across recording sessions in V1 (top) and in V4 (bottom). Dashed vertical lines delimit sessions. The gray area shows the distribution of weights of models with permuted class labels. Right: Distribution of weights, pooled across neurons from all recording session. Asterisk marks significant deviation from 0 (permutation test). **D)** Illustration of the ROC method. The ROC method utilizes probability distributions of spike counts in conditions "non-match" and "match" of a single neuron (top). We calculate the ROC curve by integrating the two distributions with a moving threshold and plotting the two integrals with respect to each other (bottom, black).

Figure 2. Gray traces indicate probability distributions (top) and ROC curves (bottom) for the model with randomly permuted class labels for conditions "match" and "non-match". The blue dashed line is the average across permutations. **E)** Left: Area under the ROC curve (black), with the gray area indicating the distribution of results for models with permuted class labels. Significant neurons are marked with a red dot. Right: Distribution of the Area under the ROC curve, pooled across neurons. The asterisk marks significant deviation of the distribution from 0.5 (permutation test). **F)** Scatter plot of the strength of the Area under the ROC versus the strength of weights of the population model, for neurons with positive weight (left) and negative weight (right). R marks the Pearson correlation coefficient and the magenta line is the least squares line. All results are for the test time window and the permutation test is with $N_{perm} = 1000$ permutations.

population model has an intrinsic structure that, in particular in V1, cannot be accounted for by the activity of single neurons.

In V1, prediction accuracy is stable across the trial, while in V4, it increases between the first and the second half of the trial.

In the following, we divide neurons with respect to the sign of their weight into plus and minus subpopulation and compute the population model using the information from each group individually. As we are interested in the prediction accuracy of plus neurons, we keep the activation of plus neurons intact and replace the activation of minus neurons with the activation from a randomly chosen trial. This way, the activation of minus neurons is randomly associated with the class label ("match"/"non-match"), making this group of neurons perform at chance. With this procedure, we find that plus neurons predict the stimulus better than minus neurons (Figure 3A-B). While minus neurons in V1 still predict the stimulus better than chance (for 2.8 %), minus neurons in V4 perform at chance. It seems therefore that classification in V4 relies entirely on neurons with positive weights.

During the trial, neural populations are expected to receive inputs from different sources. In particular, we would expect that neuronal activity is primarily driven by feed-forward inputs at the beginning of the trial, while local and top-down inputs might be important during the second part of the trial. We divide the time window of [0,500] ms w.r.t. the stimulus onset, utilized thus far, into first half ([0,250]ms) and second half of the trial ([250,500]ms w.r.t. stimulus onset). As we compute the prediction accuracy for each window, results show that, on average, the performance is stable in V1, but strongly increases from the first to the second half of the trial in V4 (Figure 3C). Even though the average performance is stable in V1, it differs in individual recording sessions, suggesting different underlying classification models. Indeed, as we compare weights of the model from the first and the second half, they are only moderately positively correlated in V1 (R=0.19) and almost uncorrelated in V4 (R=0.07,Figure 3D). As expected, we observe stronger correlation as we compare the weights from the first (Figure 3E) and second part (Figure S3A) to weights from the whole window. The distribution of weights between the first and second part of the trial remains unchanged (Figure 3D, upper plots). We remark that, interestingly, during target, weights between the first and second part of the trial are uncorrelated in both brain areas (Figure S3B).

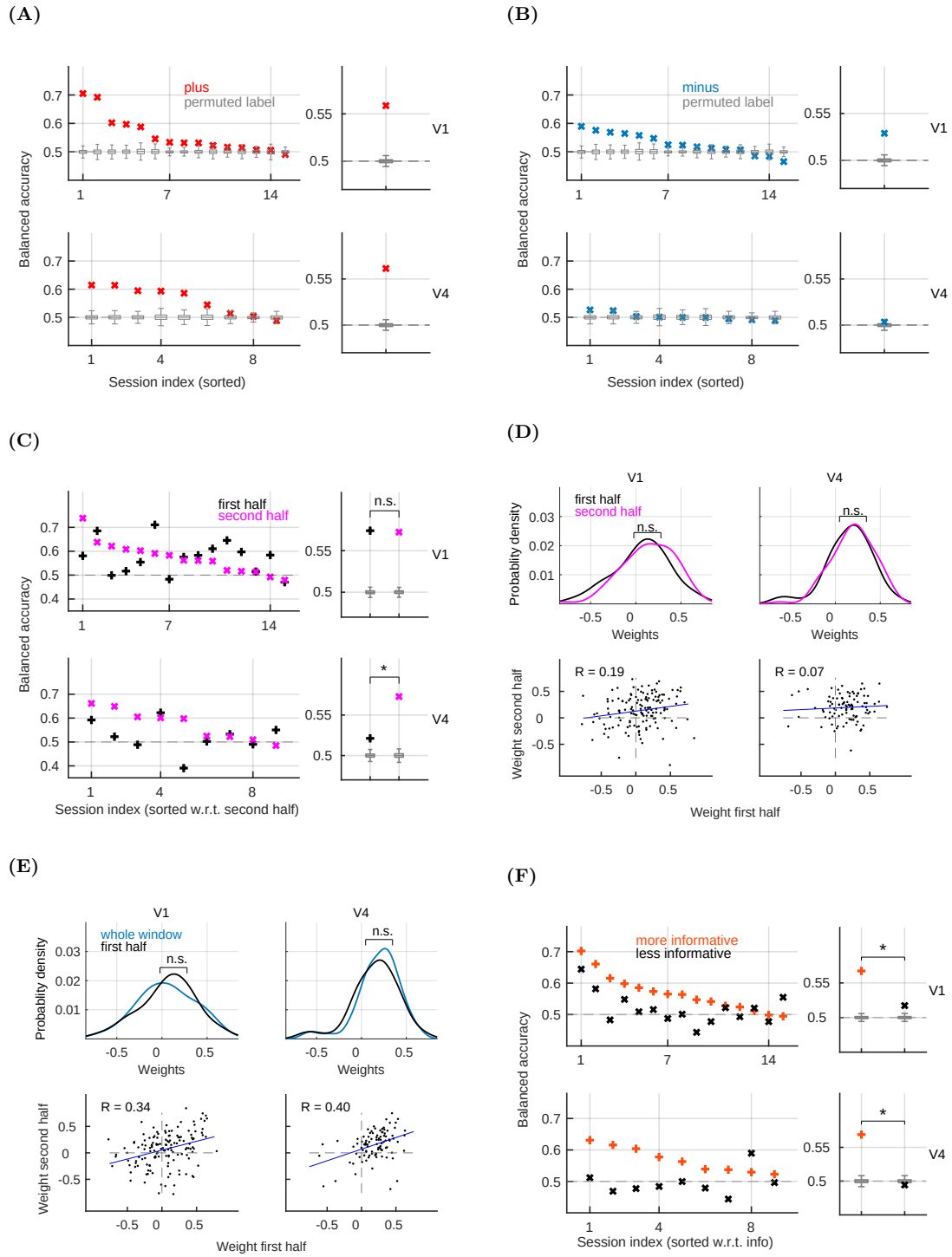

Figure 3. The population model changes between the first and the second half of the test time window. A) Left: Prediction accuracy of neurons with positive weights. We plot results in recording sessions for the true model (red), and distribution of models with permuted class labels (gray boxplots). Right: Session-averaged results. **B)** Same as in **A**, but for neurons with negative weights. **C)** Left: Prediction accuracy during the first half (black) and the second half (magenta) of the test time window. We show results in recording sessions, sorted w.r.t. the performance in the second half. Right: Session-averaged results and distribution of results for models with permuted class labels (boxplots). **D)** Top: Distribution of weights for the first (black) and the second half (magenta) of the time window. We plot results in V1 (left) and in V4 (right). Bottom: scatter plot of weights in the first vs. second half of the time window. R is a linear correlation coefficient. **E)** Same as in **E**, but comparing weights in the first half of the time window and the weights for the entire time window.

Figure 3. F) Left: Prediction accuracy for informative neurons (orange) and less informative neurons (black) in V1 (top) and in V4 (bottom). Results are sorted for the performance of more informative neurons. Right: Session-averaged results and the distribution of results for models with permuted class labels (boxplots). All results are for the test stimulus and significance is assessed with the permutation test ($N_{perm} = 1000$).

In V4, but not in V1, informative neurons are more strongly coupled, synchronized and correlated than less informative neurons.

Utilizing the strength of the weight, $|\tilde{w}_n|$, the proxy of the importance of the activity of the neuron n for classification, we divide neurons in informative and less informative. We rank the weight of each neuron within the distribution of weights resulting from models with permuted class labels, and divide neurons in informative and less informative in such a way as to keep the number of neurons in each group approximately balanced. First, we ask whether the information from neurons with strong weights is sufficient to reproduce the prediction accuracy of the population model. To test this, we keep activation of informative neurons intact, and replace the activation of less informative neurons with the activation from a randomly chosen trial. Prediction accuracy of such a model in V1 is slightly but significantly lower compared to the regular model ($p < 0.001$, permutation test), and not significantly different from the regular model in V4 (Figure 3F). As we now replace activation of informative neurons with activation from a random trial, we get chance level prediction in V4 and slightly higher than chance prediction in V1 (Figure 4B). Notice that these results depends on the threshold for splitting neurons in informative and less informative. Nevertheless, they suggest that reading-out from the subpopulation of informative neurons, the decoder can be (almost) as good as reading out from all neurons.

Next, we ask whether informative neurons differ from less informative neurons in the strength of coupling of single neurons to the population (Figure 4A). The coupling function is computed as the strength of the population activity around the spike time of a selected neuron (see methods). Examples of the coupling function are plotted on Figure 4B (left). We find that in V4, informative neurons are more strongly coupled to the rest of the population than less informative neurons ($p = 0.240$, permutation test), while no effect is observed in V1 (Figure 4B, middle and right).

Finally, we ask whether informative and uninformative neurons differ in synchrony of spiking and correlations of spike counts ([Bair et al., 2001], see methods). Synchrony and correlations are computed for pairs of neurons within the group of informative and less informative neurons (Figure 4C). Representative examples of the correlation function are plotted on the left of Figure 4D. As we test the height of the correlation function at zero time lag (e.g., synchrony) for informative vs. less informative pairs, we see that in V4, synchrony is stronger among informative neurons (Figure 4D, middle and right). Same result is observed as the correlation function is computed independently for the first half and the second half of the time window (Figure S4 A-B). We also measure noise correlations of binned spike trains, with spike counts in bins of length [20, 50, 75, 100] ms, and for all, correlations between informative neurons are stronger than between less informative neurons ($p < 0.001$ for bin length of 20, 50 and 75 ms and $p = 0.003$ for bin length of 100 ms;Figure 4E-F). Notice that this effect consistently takes place in V4, but not in V1. In V4, no such effect is observed during the target time window (Figure S5)

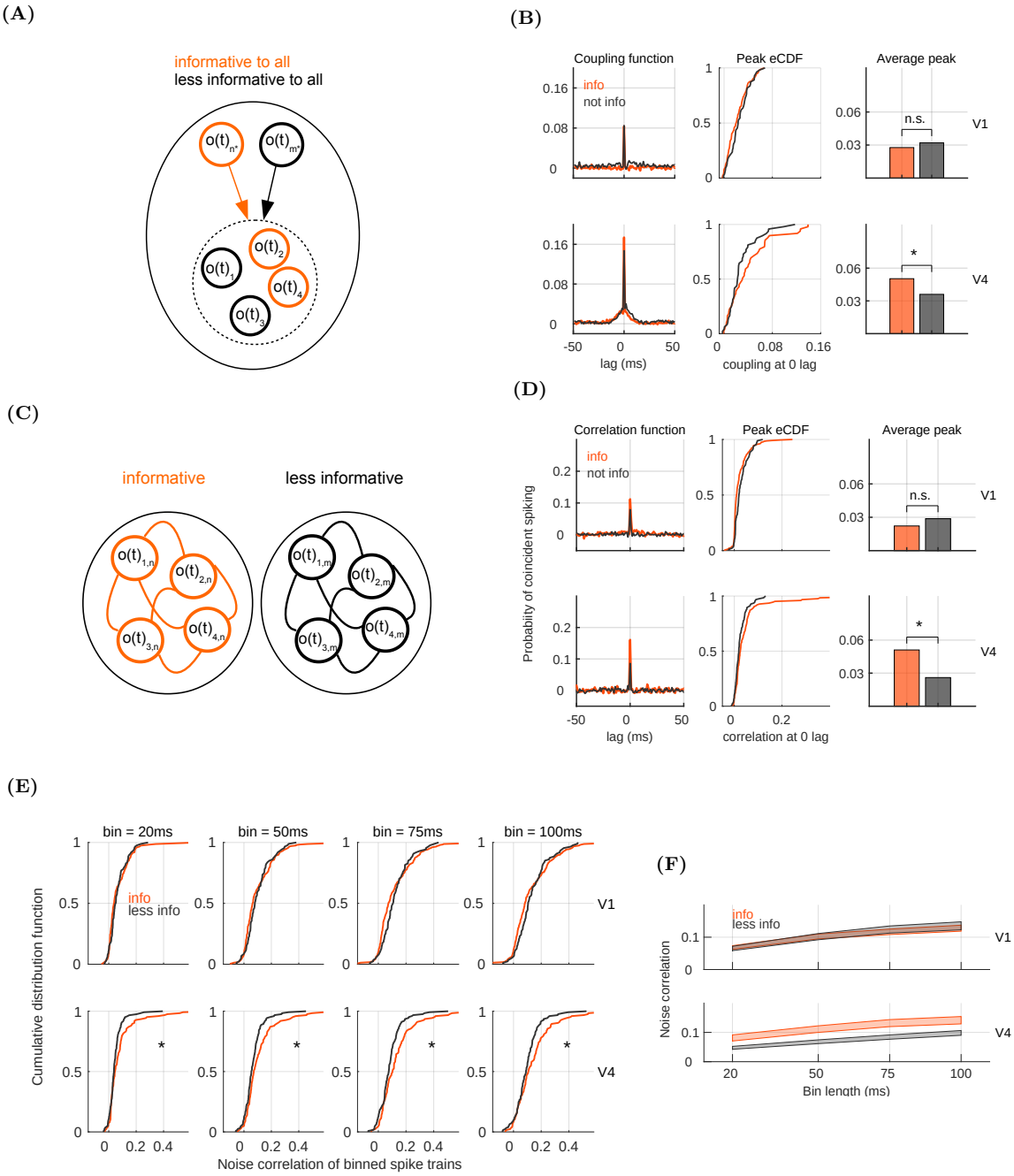

Figure 4. In V4, pair-wise synchrony and correlations among informative neurons are stronger than among less informative pairs. A) Schema of coupling of the spiking activity of informative (orange) and less informative (black) neurons to the population activity. The population activity comprises the activity of all neurons but the selected one. **B)** Left: Coupling function for an example informative (orange) and less informative (black) neuron. Middle: Empirical cumulative distribution function (CDF) of the peak of the coupling function at zero lag for informative (orange) and less informative neurons (black). Right: Average peak for the two groups of neurons. Asterisk marks significance (permutation test). **C)** Schema of pair-wise interactions within the group of informative (orange) and less informative (black) neurons. **D)** Left: Correlation function for an example pair of informative and less informative neurons. Middle: Empirical CDF of the correlation function at zero lag (synchrony). Right: Average synchrony for pairs of informative and less informative neurons. **E)** Noise correlations of spike counts for pairs of informative and less informative neurons. Noise correlations are computed for bin sizes of [20,50,75,100] ms. We plot the empirical CDF. **F)** Mean ± SEM for results in **E**. All results are for the test stimulus and $N_{perm} = 1000$.

In V4, but not in V1, coupling, correlations and spiking synchrony within the coding pool are stronger than across pools.

Besides the strength of the weight, the sign of the weight is important for the structure of population responses, since neurons with the opposite weight have the opposing effect on the angle of the classification boundary (see methods). By dividing the population in neurons with positive and negative weights, we obtain two mutually exclusive coding pools. We have shown that removing the structure of population responses across all neurons causes the population model to loose most of its predictive power (1F). We now ask, how much of the predictive power is lost if we selectively remove the structure only within coding pools (Figure 1E, right). The model that is homogeneous within coding pools performs better than chance in both areas (Figure 5A). Comparing the model with homogeneous pools to the regular model, the model with homogeneous pools is significantly worse in V1 ($p < 0.001$), but not in V4, where the difference is not significant ($p = 0.062$, permutation test). This means that in V4, the information about the sign of the weight is sufficient to reproduce the performance of the regular model, and the the amplitude od weights does not add much information to it. Notice that in V4, the size of coding pools is heavily imbalanced (with an excess of plus neurons), and that only plus neurons perform better than chance (Figure 3A-B). Permutation within the pool of plus neurons must therefore engage a disproportionately big pool of neurons. In spite of that, the prediction accuracy did not drop significantly w.r.t. the regular model. These results persist when we compute the prediction accuracy separately for the first and second half of the trial (Figure 3C). In V4, the performance of the model with homogeneous pools is similar to the regular model, low during the first half and higher during the second half (Figure 3C). In V1, meanwhile, the performance with homogeneous pools is consistently lower than the regular model in all time windows (compare Figure 5B and Figure 3C).

Next, we compare the strength of coupling of single neurons to the population, distinguishing coupling within pools and across pools. For coupling within pools, the selected neuron and the neural population are from the same coding pool, while in the second case, the selected neuron and the population are from different pools (Figure 5C). During the test time window in V4, coupling is significantly stronger within the pool compared to across pools ($p < 0.001$,Figure 5D), while there is no effect in V1 ($p = 0.654$). Same result is observed as we split the trial in the first half ($p < 0.001$ in V4, $p = 0.170$ in V1) and the second half ($p < 0.001$ in V4, $p = 0.092$ in V1; Figure S6 A-B).

Finally, we compare the strength of pair-wise interactions for pairs within and across coding pools (Figure 6A). We test the spiking synchrony (Figure 6B, left) and noise correlations on progressively longer time scales. Noise correlations are measured on spike counts in binned spike trains, for bin length of [20,50,75,100] ms and as noise correlations of the trial-to-trial variability (in the 500 ms time window, see methods). Notice that these measures assess pair-wise interactions ranging from fast time scales (1 ms) to longer time scales up to 500 ms. In V4, pair-wise synchrony is stronger for pairs within compared to across coding pools ($p = 0.001$, 6B, middle and right). Same effect is observed when we measure the correlation functions in the first half and second half of the trial (in V4, $p < 0.001$ for the first half and $p = 0.001$ in the second half; Figure S6 C-D). Noise correlations of spike counts in V4 are consistently stronger within compared to across pools ($p = 0.004$ for bin length of 20 ms, $p < 0.001$ for 50 ms, $p = 0.02$ for 75 ms, and $p = 0.032$ for 100 ms; Figure 6C-D). Consistently, correlations of trial-to-trial variability within pools are stronger than across pools (Figure 6E). This is again only observed in V4 ($p < 0.001$), but not in V1 ($p = 0.731$). In summary, during test in V4, pair-wise interactions on different

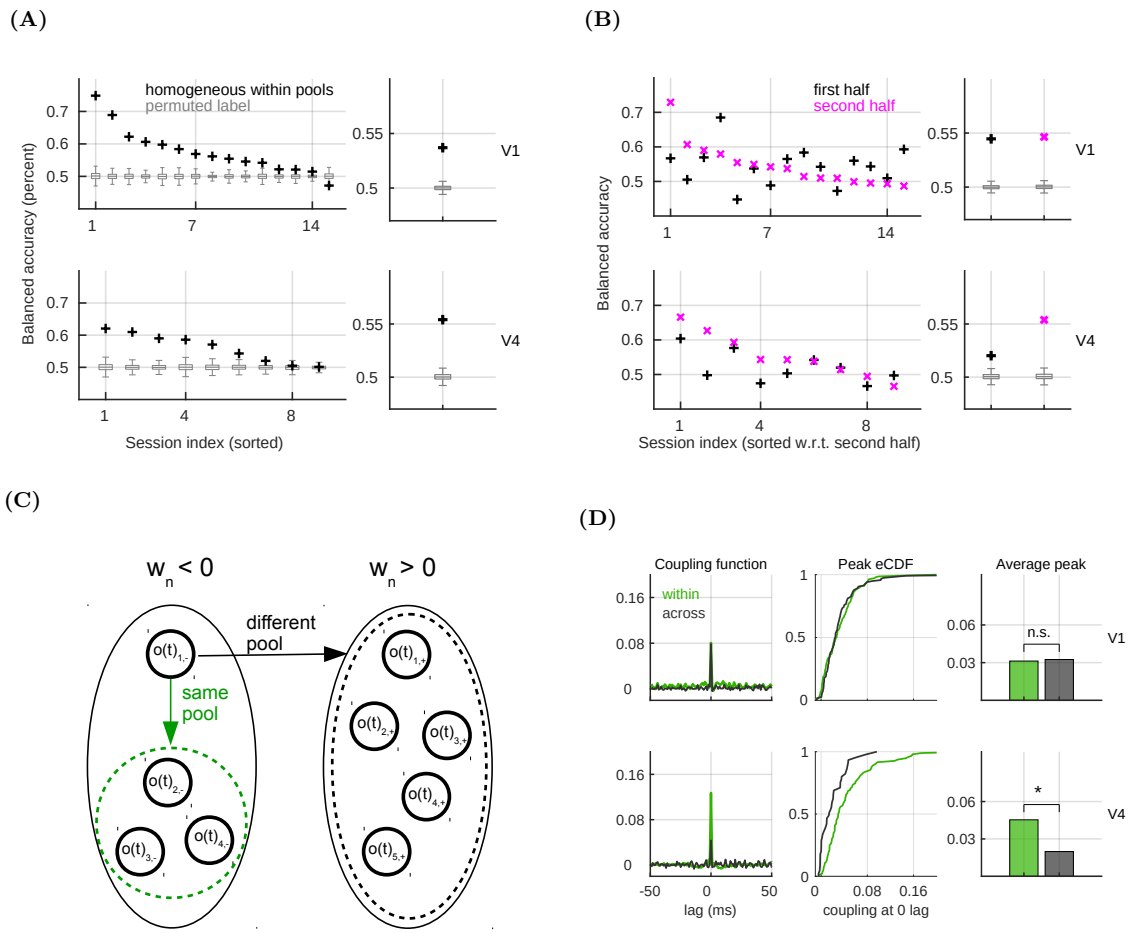

Figure 5. In V4, coupling within pools is stronger than across pools. A) Prediction accuracy of the model that is homogeneous across neurons in the same coding pool. The homogeneity of pools is achieved by permuting the neural index among neurons with the same sign of the weight. We use the entire time window ([0,500] ms w.r.t. stimulus onset). **B)** Same as in **A**, but using the first half (black) and the second half of the time window (magenta). **C)** Schema of coupling within pool (green arrow) and across pools (black arrow). The dashed line delimits the group of neurons, used to compute the population activity. **D)** Left: The coupling function for an example neuron, for coupling within the pool (green) and across pools (black). The coupling function is plotted as a function of the time lag and is computed in the entire time window ([0,500] ms w.r.t. the stimulus onset). Middle: Empirical CDF for coupling at 0 lag, distinguishing coupling within and across pools. Right: Average coupling at 0 lag, with asterisk marking significance (permutation test). All results are for the test stimulus. Permutation tests utilize $N_{perm} = 1000$ permutations.

time scales all reflect the functional separation of neurons into coding pools. This effect is absent in V1, where pair-wise interactions within and across pools are of similar magnitude.

Interestingly, during the target time window, correlations of trial-to-trial variability are stronger within compared to across pools not only in V4 but also in V1 ($p < 0.001$ in both areas, Figure 6F). During the first half of the test time window, we also observe that correlations within pools are stronger than across pools in both areas ($p < 0.001$ in both, Figure S6E) and that during the second half of the trial, this effect vanishes in V1 while it persists in V4 ($p = 0.067$ in V1, $p < 0.001$ in V4; Figure S6F).

Correlations within pools are harmful for the decoder while correlations across pools are not.

We measure the effect of correlations on the classification performance by removing the correlation structure between (a specific group of) neurons from the same recording session and testing whether such a perturbation changes significantly the classification performance with respect to the regular model. The removal of the correlation structure is achieved by permuting the order of trials, independently for each neuron within the group of interest. We perturb the data in two ways: 1.) by removing correlations across the two coding pools but keeping correlations within the coding pools intact and 2.) by removing all correlations (Figure 7A). Notice that these two types of correlations are mutually exclusive and comprise all correlations in the network. Removal of correlations across the coding pools does not affect the classification performance (Figure 7B), however, removing all correlations increases the classification performance in both brain areas ($p < 0.001$ in V1 and V4; Figure 7C). The increase in performance occurs mostly during the second half of the trial (Figure 7D), in particular in V4, where the increase is limited to the second half. Testing the effect of the perturbation of the type 1) in the first and second half of the trial, we see no significant effect in V4, and a small increase in performance during the second half in V1 (Figure 7E). As we compare weights of models with removed noise correlations with weights of the regular model, we see that removing noise correlations causes shifting of the distribution towards negative weights (Figure 7F).

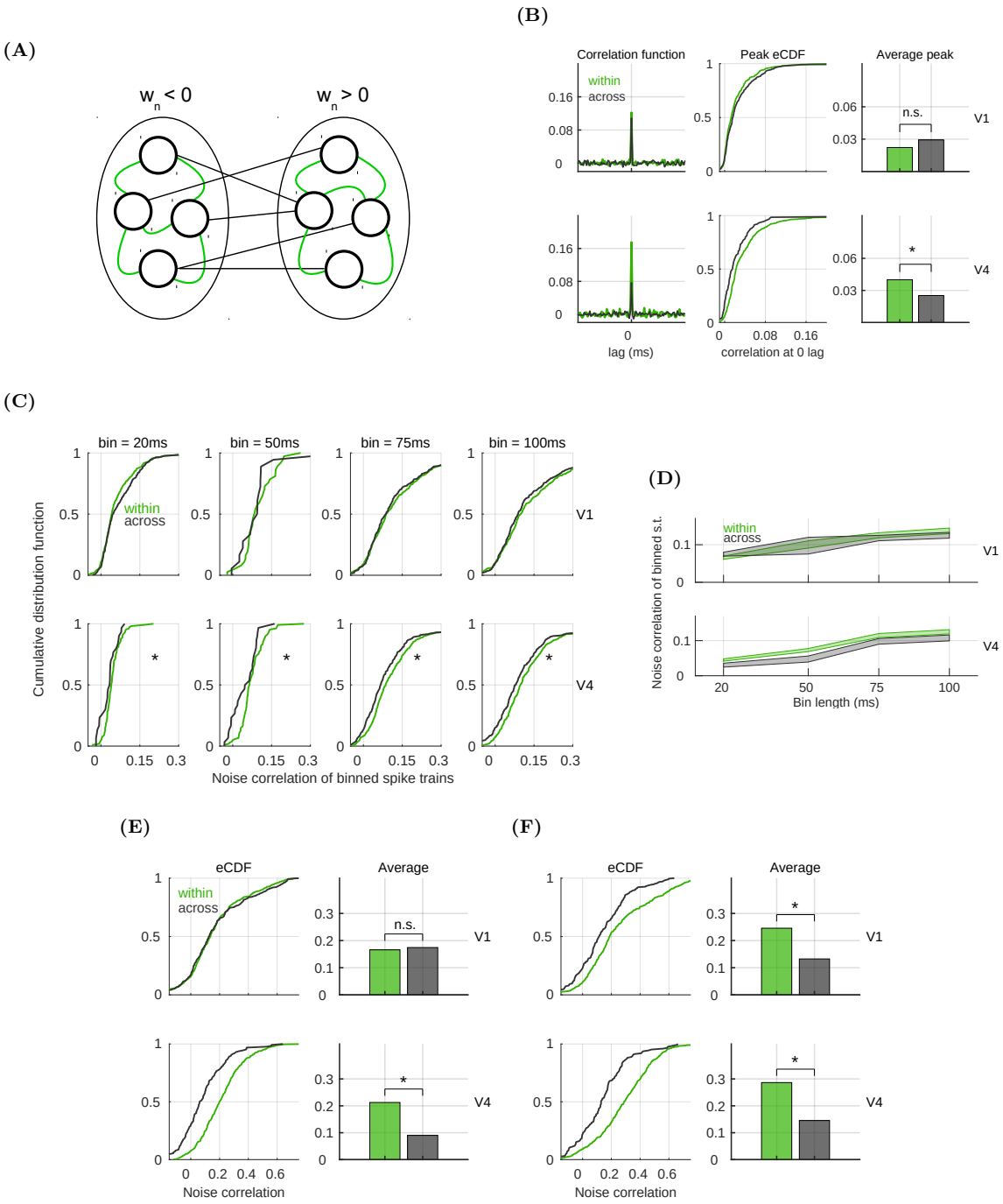

Figure 6. In V4, pair-wise synchrony and correlations within coding pools are stronger than across pools. A) Schema of two coding pools, one with negative (left) and the other with positive (right) weights. Green links are for pair-wise interactions within the pool and black are for interactions across pools. **B)** Left: Correlation function for two representative pairs of neurons, one pair from the same coding pool (green), and one pair where neurons are from different pools (black). Middle: Empirical CDF for the correlation function at zero lag (synchrony). Right: Average synchrony within and across coding pools. For all, we use the entire time window ([0,500] ms w.r.t. the stimulus onset). **C)** Noise correlations of binned spike trains, for bins of [20,50,75,100]ms, distinguishing pairs within (green) and across pools (black). We plot the empirical CDF of correlation coefficients, collected across recording sessions, **D)** Same as in **C**, but showing the mean ± SEM of the noise correlations, plotted as a function of the bin length. **E)** Noise correlations of trial-to-trial variability for pairs of neurons within the coding pool (green) and across pools (black). We plot the empirical CDF (left) and the average correlation coefficient (right). For the latter, asterisk marks significance (permutation test). **F)** Same as in **E**, but for the target time window. For all plots, we used the entire time window ([0,500] ms w.r.t. the stimulus onset) and $N_{perm} = 1000$ permutations.

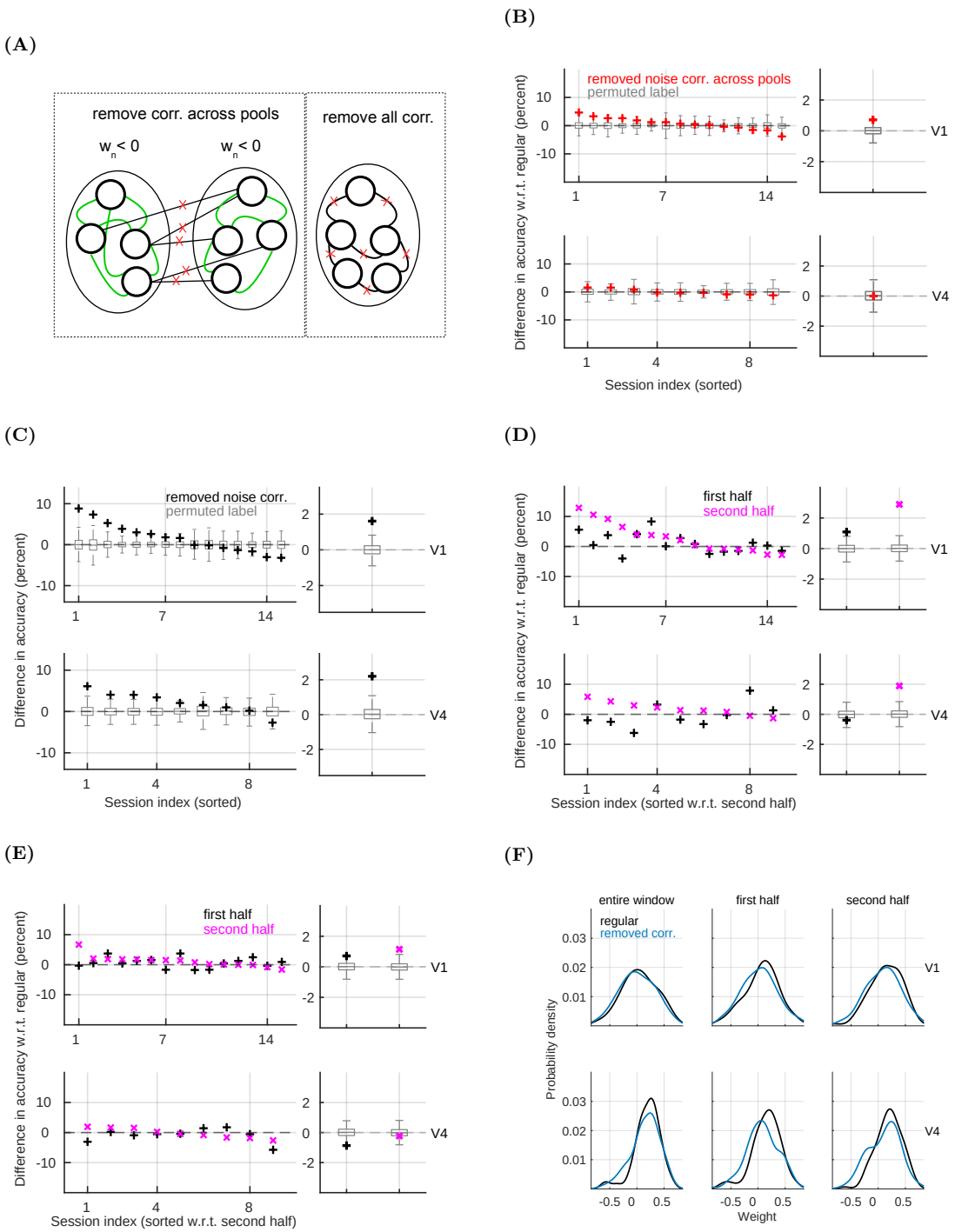

Figure 7. Removing correlations within coding pools increases the predictive power of the population model. A) Schema of selectively removing correlations across coding pools, while correlations within pools are kept intact (left), and of removing correlations across all neurons (right). **B)** Difference in prediction accuracy between models with removed correlations across pools and the regular model. We show results in recording sessions (left) and session-averages (right). The entire time window has been utilized. **C)** Same as in **B**, but for the removal of all correlations. **D)** Same as in **C**, but for the first (black) and second (magenta) half of the time window. **E)** Same as in **B**, but for the first (black) and second (magenta) half of the time window. For plots **B** to **E**, the testing is done with the permutation test, using $N_{perm} = 1000$ permutations. **F)** Distribution of weights for the regular model (black) and the model with removed noise correlations across all neurons (blue). We plot results for the entire window (left), first half of the window (middle) and second half (right).

Discussion

We addressed the structure of parallel spike counts and the role of noise correlations for discrimination of binary stimulus classes from the activity of neural ensembles in V1 and V4 of the macaque. We showed that the population model predicts the stimulus class with accuracy that is better than chance in both brain areas. The population model utilizes high-dimensional patterns of activation of single neurons as the most important source of information for discrimination. The strength of population weights is positively correlated with discriminatory capacity of single neurons, however, the population model cannot be fully accounted for by aggregating responses of single neurons, because of correlations. In V4, but not in V1, the structure of the population model is clearly reflected in pair-wise dynamics on short and middle-length time scales. We show that the discrimination capacity of the population model relies on a subset of most informative cells, that are in V4 also more strongly coupled, synchronized and correlated than less informative neurons. As the population is split in two decoding pools according to the sign of the population weight, we show that in V4, but not in V1, pairs of neurons within the coding pool are more strongly coupled, synchronized and correlated than across pools. The effect of the coding function on the correlation structure consistently appears on time scales going from 1 to 500 ms. Finally, we show that correlations within the pool are harmful for the decoder, while correlation across pools are not.

While the choice probability has already been reported in V4 ([Shiozaki et al., 2012]) as well as in V1 ([Nienborg and Cumming, 2014]), previous studies used artificial stimuli that specifically target filtering properties of observed neurons. While artificial stimuli might potentially maximize the choice probability in observed neurons, they are also likely to affect inter-neuron correlations. We argue that our study adds an important piece of information to existing literature on choice probabilities by introducing naturalistic stimuli, and a delayed-to-sample visual task. The present behavioral task requires to memorize the template image (the target stimulus) and compare it to a delayed sample (the test stimulus), and is expected to activate not only multiple brain areas across the visual ventral stream, but also the prefrontal cortex ([Gold and Shadlen, 2007]). Moreover, stimuli are complex natural images that change from one trial to another, which prevents the discrimination to rely on specific orientation of the test stimulus. In such a setting, it is not trivial to predict the stimulus class in early (V1) and middle (V4) visual cortex with probability that is better than chance.

Our results suggest that the predictability of the stimulus class could be due to the local structure of neural ensembles and/or to top-down projections from higher brain areas, such as IT or prefrontal cortex. The classification of matching and non-matching stimuli could be computed downstream, and the choice-related signal would then be projected back to V1 and V4 to reinforce the internal evidence for the choice. All other inputs being equal, the weights of the population model might be capturing a difference in top-down projections for matching and non-matching stimuli. Local separation into coding pools is additionally supported by the correlation structure, where neurons within the coding pool are more strongly correlated than neurons across pools. Present results suggest that having stronger correlation within compared to across pools might be the default correlation structure in the present experimental task, that vanishes from V1 responses towards the end of the test stimulus. Since correlations within pools are harmful for the prediction accuracy, correlations during test in V1 might be actively suppressed in order to improve the information transfer of the choice-related signal. Since correlations are either neutral or harmful for the quantity of decoded information, the benefit of high-dimensionality in terms of the information transfer is in the ability of forming high-dimensional patterns

of activity, and not in the correlation structure. However, the fact that informative neurons in V4 have stronger correlations than less informative neurons suggests that correlated activity of informative neurons might be useful for the read-out, perhaps in the sense of making it more reliable.

Neural microcircuits under investigation present a mixture of correlations, some of which are harmful for the decoder, and others that are not ([Moreno-Bote et al., 2014]). The reason for this might be to perform reasonably well in discrimination tasks, given the constraint of the noisy environment in which neurons operate ([Shadlen and Newsome, 1998]). We argue that in the present experimental task, the activity of neural ensembles has two main functions, one is to collect the information about the stimuli and classify them in "match" and "non-match", and the other is to calculate a binary variable than can guide the choice behavior. While the neural activity in V1 might be optimized in such a way as to maximize the transfer of the information about the visual scene, the quantity of transferred information might not need to be optimized in downstream areas, where the reliable transmission through correlated channels might be a priority. The repertoire of the choice behavior in the present setting is reduced to a binary variable (the choice "same" and "different"), which only contains 1 bit of information. As correlations with the pool are stronger than across pools, their effect is to amplify the signal within the pool, which might help a reliable transmission of the signal.

If population weights have been learned through a form of Hebbian synaptic plasticity ([Abbott and Nelson, 2000]), the formation of coding pools can be explained by the choice-specific top-down input. On a coarser spatial and temporal scale, it has been demonstrated that patterns of spontaneous activity reflect the functional architecture of the cortex ([Arieli et al. 1996, Tsodyks et al., 1999, Fox and Raichle, 2007]), including the representation of abstract information, such as rules or goals ([Waskom and Wagner, 2017]). Finding the same organization in V4 with the present classification task opens up the possibility that the correlation structure actively supports processing of stimuli in categorization tasks on a much smaller spatial scale.

Prediction accuracy of the population model is the best prediction that can be obtained from a linear classification model in the present classification task, setting the higher bound on the information about the choice that can be extracted from spike counts of observed neural ensembles in V1 and in V4. We showed that the use of a nonlinear SVM with radial basis function kernel does not improve the prediction accuracy. We also argue that the relatively low prediction accuracy of the population model might be importantly amplified as the coding function is exploited dynamically. If population weights are applied on spike trains, weighting the contribution of spikes with respect to the low-dimensional population signal, the difference between the signals in conditions "match" and "non-match" can accumulate over time. In a related work, we show that a leaky integration of weighted spike trains in conditions "match" and "non-match" results in signals that are significantly different and diverge over the time course of the trial ([Koren et al., 2019]).

It has to be pointed out that while our population model relies on a linear classifier, this linearity should not be understood as the linearity between responses of single neurons and the population. On the contrary, we show that single neurons cannot account for the structure and performance of the population model. In the biological setting, population weights have a straightforward interpretation as synaptic weights between a population of projecting neurons and a read-out neuron. The sign of the weight would correspond to strengthening (positive weight) or weakening (negative weight) of the synaptic weight between the projecting neuron and the read-out neuron with respect to the baseline. In biological networks, the contributions of projecting neurons sum at the synapse of the read-out neuron and form the synaptic input. What the linearity of the population model does imply is that synaptic inputs sum linearly as they create the synaptic current at the read-out neuron, a well-

established experimental fact, and common use in modeling ([Dayan and Abbott, 2001]). While the transfer function of the read-out neuron might not be linear due to the spike-triggered adaptation, voltage-gated currents, etc., this can be easily considered while modeling present results, but is beyond the scope of the present work.

An interesting venue for future research could be the comparison of the orientation tuning of single neurons and their decoding weights. Theoretical studies predict the relation between the similarity of tuning curves and noise correlations ([Averbeck et al., 2006, Moreno-Bote et al., 2014]). In this study, we have related noise correlations and similarity of decoding weights, obtaining similar relation as predicted for tuning curves. In the present experimental setting, computation of the choice-related signal might not directly depend on the orientation tuning of V1 and V4 neurons, and it remains unclear how orientation tuning and decoding weights are related. Future work could also address the variability of weights and how they change over the trial, topics only coarsely studied in the present work. Finally, the separation of feed-forward, local and feed-back inputs could provide major insights into how neural ensembles in the visual cortex compute choice-related signals.

STAR Methods

Lead contact and materials availability

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Veronika Koren (koren@ni.tu-berlin.de). This study did not generate new unique reagents. Part of these methods and the totality of the dataset have been used in a published work ([Koren et al., 2019]).

Experimental model and subject details

Ethics statement. All experiments performed in this study were conducted in accordance with protocols approved by The Animal Welfare Committee (AWC) and the Institutional Animal Care and Use Committee (IACUC) for McGovern Medical School at The University of Texas Health Science Center at Houston (UTHealth), and met or exceeded the standards proposed by the National Institutes of Health's Guide for the Care and Use of Laboratory Animals.

Animal subjects Two male rhesus macaques (Macaca mulatta; M1, 7 years old, 15kg; M2, 11 years old, 13kg) were used in this study. Subjects were housed individually (after failed attempts to pair house) in cages sized 73 x 69 x 31 or 73 x 34.5 x31 inches, in close proximity to monkeys in adjacent cages, allowing for visual, olfactory and auditory contact. Toys were given in rotation, along with various puzzles, movies and radio programming as environmental enrichments. Monkeys were fed a standard monkey biscuit diet (LabDiet), that was supplemented daily with a variety of fruits and vegetables. Subjects had been previously trained to perform visual discrimination task, and each implanted with a titanium head post device and two 19mm recording chambers (Crist Instruments) over V1 and V4. All surgeries were performed aseptically, under general anesthesia maintained and monitored by the veterinary staff from the Center for Laboratory Animal Medicine and Care (CLAMC), with appropriate analgesics as directed by the specialized non-human primate veterinarian at CLAMC. During the study the animals had unrestricted access to fluid, except on days when behavioral tasks were performed. These days, animals had unlimited access to fluid during the behavioral task, receiving fluid for each correctly completed trial. Following the behavioral task, animals were returned to their home

cage and were given additional access to fluid. The minimal daily fluid allotment was 50ml/kg (monkeys were weighed weekly), though monkeys could drink more through their participation in the task. During the study, the animals' health and welfare was monitored daily by the veterinarians and the animal facility staff at CLAMC and the lab's scientists, all specialized with working with non-human primates.

Method details

Experimental procedure The trial started after 300 ms of successful fixation within the fixation area and consisted in visualizing the target and the test stimuli, with a delay period in between. Stimuli were shown for 300 ms each, while delay period had a random duration between 800 and 1000 ms, in order to prevent the expectation of the timing of the test stimulus. The target and the test stimuli, naturalistic images in black and white, were either identical (condition "match") or else the test stimulus was rotated with respect to the target stimulus (condition "non-match"). The task of the animal was to decide about the similarity/difference of the target and the test stimuli by holding a bar for "different" and releasing the bar for "same". The subject had to respond within 200 and 1200 ms from the offset of the test stimulus, otherwise the trial was discarded. The difficulty of the task was calibrated on-line, in such away as to have, on average, 70 percent of correct responses on non-matching stimuli. The difference in orientation of the test stimulus ranged between 3 and 10 degrees.

Recordings In every recording session, two laminar electrodes with 16 recording channels each were inserted in V1 and V4 areas, measuring the multi-unit signal with 16 recording channels (0.1 mm spacing between adjacent contacts). Electrodes were inserted perpendicularly to the cortical surface, and calibrated such that neurons from the two areas had overlapping receptive fields. Since recording sessions were performed on different days, it is extremely unlikely that the electrode captured same neurons in different recording sessions. We therefore treated spike-sorted units from different recording sessions as distinct cells. The analysis included all cells that responded to the stimulus with a 4-fold increase of the firing rate with respect to the baseline. Since we are interested in decoding population activities, we used recording sessions with at least 6 simultaneously recorded units. This amounts to 142 neurons in V1, collected in 15 recording sessions, and to 97 neurons in V4, collected in 9 recording sessions. The average, minimum and maximum number of trials per session are reported in the table 2. Besides mentioned criteria, we did not exclude any part of the data. All the analysis is done for each session individually, and results are then collected across sessions and across the two animal subjects.

area	condition	average trials	min. trials	max. trials
V1	"non-match"	108	42	230
	"match"	118	51	290
V4	"non-match"	73	39	104
	"match"	96	78	123

Table 2. Statistics of trial number per condition.

Quantification and statistical analysis

Classification and statistical analysis were done with Matlab, Mathworks, version R2017b. All methods were applied to the test and target time windows independently. If not mentioned otherwise, we used the window

[0,500] ms with respect to the onset of the corresponding stimulus. For part of the analysis, we used, in addition the the whole window, also the first and second half of the window ([0,250]ms and [250,500]ms w.r.t. the stimulus onset), as mentioned in the main text and in corresponding figures.

Classification and extraction of decoding weights

The spike train of a single neuron is a binary vector of zeros and ones,

$$o_{n,j}(t_k) = \begin{cases} 1, & \text{if neuron } n \text{ in trial } j \text{ spikes during the } k\text{-th millisecond} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $n = 1, ..., N$ is the neural index, $j = 1, .., J$ is the trial index and $k = 1, ..., K$ is the time index with step of 1 millisecond. The 1 millisecond bin is the temporal resolution of the spike train.

Area under the ROC curve for univariate analysis. The discriminability of conditions "match" and "non-match" from spike counts of single neurons was evaluated using the receiver operating characteristic (ROC) analysis ([Gold and Shadlen, 2007, Britten et al., 1996, Newsome et al., 1989, Shadlen and Newsome, 2001] [Zohary et al., 1994]). The ROC is a non-parametric method that evaluates the distance between distributions of spike counts in conditions "match" and "non-match", collected in multiple trials. The spike count of the neuron n in trial j is defined as the sum of spikes in the desired time window: $s_{n,j} = \sum_{k=1}^{K} o_{n,j}(t_k)$.

Spike counts are collected across trials and define two probability distributions, corresponding to conditions "match", $P_M(s)$, and "non-match", $P_{NM}(s)$, for $s \in \{0, \Delta s, .., s_{max}\}$, where s_{max} is the maximal spike count across all recorded activities, $\Delta s = 2s_{max}/(nstep - 1)$, $nstep = 200$. Probability distributions are convolved with a normalized Gaussian kernel,

$$\begin{aligned} \tilde{P}_{NM}(s) &:= \sum_{x \in X} P_{NM}(s - x)u(x), \\ \tilde{P}_M(s) &:= \sum_{x \in X} P_M(s - x)u(x), \end{aligned} \tag{2}$$

with $u(x) = \left(\sum_{x \in X} \exp(-\frac{x^2}{2{\sigma_u}^2}\right)^{-1} \exp(-\frac{x^2}{2{\sigma_u}^2})$. The width of the kernel σ_u is defined as follows: first, an estimate for the optimal width of each distribution is computed using Silverman's rule, $\sigma^*_{M,NM} = \sigma_{M,NM}(\frac{4}{3J_{M,NM}})^{\frac{1}{5}}$, where σ_M (σ_{NM}) is the standard deviation of the distribution of spike counts in condition "match" ("non-match") and J_M (J_{NM}) is the number of trials in condition "match" ("non-match"). Then, we set $\sigma_u = \frac{\sigma^*_M + \sigma^*_{NM}}{2}$.

We define empirical cumulative distribution functions, $D_M(s)$ and $D_{NM}(s)$, as follows:

$$D_{M,NM}(s) := \sum_{s'=0}^{s} \tilde{P}_{M,NM}(s')\Delta s \tag{3}$$

Note that the cumulative distribution functions are bounded, $0 \leq D_{M,NM}(s) \leq 1$, and the area under the ROC curve is the following sum:

$$AUROC = \frac{1}{2} \sum_{s \in S} (D_{NM}(s+1) + D_{NM}(s))(D_M(s+1) - D_M(s)) \tag{4}$$

Two identical distributions give the $AUROC$ of 0.5, at chance prediction of conditions "match" and "non-match" from the observation of spike counts. In order to estimate if the prediction is significantly better than chance, we have to take into account the variance of the data. This is done with the permutation test. We pool the input statistics ($s_{n,j}$) from the two conditions and randomly assign labels "match" and "non-match" to every trial. The pooled distribution is then split into two subsets, with the number of samples in subsets corresponding to the number of original samples in conditions "match" and "non-match". Using these two random subsamples, the area under the curve is then computed by eq. 2 - 4. This procedure is repeated $nperm = 1000$ times, and results in a distribution of areas under the ROC curve, $AUROC_p,\ p = 1, ..., n_{perm}$. We compute the fraction of $AUROC_p$ that are larger (p_L) and smaller (p_S) than the true $AUROC$,

$$\begin{aligned} p_L &= \frac{1}{n_{perm}} \sum_{p=1}^{n_{perm}} \mathbb{1}(AUROC_p > AUROC) \\ p_S &= \frac{1}{n_{perm}} \sum_{p=1}^{n_{perm}} \mathbb{1}(AUROC_p < AUROC) \end{aligned} \tag{5}$$

where $\mathbb{1}(x > y)$ is the following logical function:

$$\mathbb{1}(x > y) = \begin{cases} 1 & \text{if } x > y \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

The $AUROC$ is considered to be significantly smaller (larger) than 0.5 if the following is true: $p_L \leq {}^{0.05}/_{N_{tot}}$ ($p_S \geq {}^{0.05}/_{N_{tot}}$), where N_{tot} is the total number of neurons in a given brain area.

Finally, we are interested in the distribution of $AUROC$ across neurons. We pool $AUROC$ across neurons from all recording sessions and from the two animal subjects and get the following distribution: $AUROC_n, n = 1, ..., N_{tot}$. We compute the mean of the distribution, $\langle AUROC \rangle_n$, and test its significance by comparing it with the distribution of averages from the permutation procedure, $\langle AUROC_p \rangle_n,\ p = 1, ..., n_{perm}$. The p-value is then computed by eq. 5-6, and corrected for multiple testing with Bonferroni correction.

Linear Support Vector Machine for multivariate analysis. The second read-out method consists in discriminating conditions "match" and "non-match" on high-dimensional activity profiles with linear Support vector machine (SVM). We use z-scored spike counts,

$$\tilde{s}_{n,j} = \frac{s_{n,j} - \langle s_{n,j} \rangle_j}{\sqrt{\mathrm{Var}_j\ (s_{n,j})}}. \tag{7}$$

where $\langle s_{n,j} \rangle_j$ is the empirical mean and $\mathrm{Var}_j\ (s_{n,j})$ is the empirical variance across trials. One sample for the classifier is an N-dimensional activity vector of N simultaneously recorded neurons in trial j,

$$\tilde{\mathbf{s}}_j = [\tilde{s}_{1,j}, \tilde{s}_{2,j}, ..., \tilde{s}_{N,j}]. \tag{8}$$

The optimization problem of the linear SVM, in its primal form, is expressed with a Lagrangian [Vapnik and Vapnik, 1998],

$$L_p = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{j=1}^{J} \lambda_j [y_j(\mathbf{w}^T\tilde{\mathbf{s}}_j + b) - 1], \tag{9}$$

where $\mathbf{w}$ is $(N, 1)$-dimensional vector of weights and N is the number of simultaneously active neurons, b is the offset of the separating hyperplane from the origin, λ_j is the Lagrange multiplier ($\lambda_j \leq 0 \; \forall j$) and y_j is the class label in trial j, with $y_j \in \{-1, 1\}$.

The model is trained on 80 % of trials, and tested on the remaining 20 %. We perform Monte-Carlo cross-validation for random splits in the training and test set with 100 iterations. The classification performance on the test set is computed with balanced accuracy,

$$BAC = \frac{TP}{2(TP + FN)} + \frac{TN}{2(TN + FP)}, \tag{10}$$

where TP, FN, TN and FP stand for the number of true positive, false negative, true negative and false positive classifications on the test set, respectively. We call a true positive the correct classification of condition "match", a true negative the correct classification of condition "non-match", etc. Reported balanced accuracy is averaged across cross-validations. The regularization parameter of the SVM is estimated with 10-fold cross-validation on the training data, and chosen from the following range: $C \in \{0.0012, 0.00135, 0.0015, 0.002, 0.005, 0.01, 0.05, 0.1, 0.5\}$.

If the classifier performs at chance, its balanced accuracy is 0.5. The significance of the deviation of balanced accuracy from 0.5 is determined with permutation test. We train models with randomly permuted class labels for "match" and "non-match". Iterating the permutation procedure gives a distribution of classification performances, $BAC_p, \; p = 1, ..., n_{perm}$, with $n_{perm} = 1000$. The p-value of the permutation test is the probability of models with permutation of having higher balanced accuracy than the true model.

$$p_{svm} = \frac{1}{n_{perm}} \sum_{p=1}^{n_{perm}} \mathbb{1}(BAC_p > BAC) \tag{11}$$

If $p_{svm} < \frac{0.05}{N_{test}}$, the balanced accuracy is considered to be significantly higher than chance, where $N_{test} = 4$ stands for the Bonferroni correction for multiple testing.

In order to compare the predictive power of the high-dimensional model with single neurons, we also compute the SVM on the single neuron data. The classification procedure is, in this case, done for each neuron independently. Balanced accuracy is then averaged across neurons from the same recording session.

Extraction of decoding weights. The minimization of the Lagrangian (eq. 9) allows to compute the vector of weights,

$$\mathbf{w} = \sum_{q=1}^{Q} \lambda_q y_q \mathbf{s}_q, \tag{12}$$

where Q is the number of support vectors [Vapnik and Vapnik, 1998]. The range of the weight vector depends on the regularization parameter of the SVM (C parameter) and is therefore different from one session to another. Since we combine results from different recording sessions, we normalize the weight vector with L^2 norm,

$$\tilde{\mathbf{w}} = \frac{\mathbf{w}}{||\mathbf{w}||} \tag{13}$$

with $||\mathbf{w}|| = \sqrt{w_1^2 + ... + w_N^2}$. The vector of decoding weights, $\tilde{\mathbf{w}} = [\tilde{w}_1, \tilde{w}_2, ..., \tilde{w}_N]$, associates the activity of each neuron with it's role for classification in the N-dimensional space of inputs.

The distance of weights is defined as the Euclidean distance of weights between pairs of neurons from the

same recording session,

$$d_{n,m} = |\tilde{w}_n - \tilde{w}_m| \tag{14}$$

with $n, m = 1, ..., N$. Results are then pooled across recording sessions.

Coupling of single neurons with the population

We measure the strength of the coupling of single neurons with the population using a measure of spike-triggered population activity, defined similarly as in methods of Part 1 (see also [Koren and Denève, 2017, Luczak et al., 2013]). We consider the spike train of neuron n^* in trial j. The spike train can be expressed as a sum of Kronecker delta functions,

$$o_{n^*,j}(t_k) = \sum_{i=1}^{\alpha} \delta_{t_i}(t_k) \tag{15}$$

where t_i is the spike time of the $i-th$ spike and α is the number of spikes of neuron n in trial j. To compute the population activity in trial j, we average spike trains across neurons, excluding the neuron n^*:

$$\beta_j(t_k) = \frac{1}{N-1} \sum_{\substack{n=1 \\ n \neq n^*}}^{N} o_{n,j}(t_k). \tag{16}$$

For each spike of neuron n^*, we measure the population activity as a function of the time lag before and after the spike:

$$\gamma_j(\tau) = \frac{1}{\alpha K} \sum_{i=1}^{\alpha} \sum_{k=1}^{K} \beta_j(t_k - \tau) \delta_{t_i}(t_k), \tag{17}$$

where the lag is $\tau = -\tau_{max}, -\tau_{max} + 1, ..., \tau_{max}$. Note that $\tau = 0$ corresponds to coupling at the time of the spike. Coupling function $\gamma_j(\tau)$ depends on the stimulus, and we make it independent by subtracting the trial-invariant coupling. Trial-invariant coupling is computed as coupling where the spike train and the population activity are not from the same trial, e.g., we use $o_{n^*,j'}(t_k)$ and $\beta_j(t_k)$, with $j' \neq j$. The trial index for the former is determined with random permutation of the order of trials. The permutation is repeated $nperm = 100$ times and the coupling function is averaged across repetitions, resulting in $\gamma_j^{inv}(\tau)$. Finally, we obtain the trial-specific (or "noise") coupling function by subtracting the trial-invariant from the original coupling function, and averaging across trials,

$$\gamma^{noise}(\tau) = \frac{1}{J} \sum_{j=1}^{J} (\gamma_j(\tau) - \gamma_j^{inv}(\tau)). \tag{18}$$

The coupling coefficient is defined as the coupling function at zero time lag, $\Gamma = \gamma^{noise}(0)$.

Measures of pair-wise interactions

Pair-wise measures are computed for pairs of neurons from the same recording session and are then pooled across sessions.

Correlation of trial-to-trial variability. We use z-scored spike counts, $\tilde{s}_{n,j}$. For each neuron, we define a vector of activation across trials, $\tilde{\mathbf{s}}_n = [\tilde{s}_{n,1}, \tilde{s}_{n,2}, ..., \tilde{s}_{n,J}]$. The correlation coefficient between activation of a

pair of neurons n and m is computed as the Pearson correlation coefficient,

$$\begin{aligned} r_{n,m}^{trial} &= \frac{\mathbb{E}[\tilde{\mathbf{s}}_n, \tilde{\mathbf{s}}_m] - \mathbb{E}[\tilde{\mathbf{s}}_n]\mathbb{E}[\tilde{\mathbf{s}}_m]}{\sqrt{\mathrm{Var}(\tilde{\mathbf{s}}_n)\mathrm{Var}(\tilde{\mathbf{s}}_m)}} \\ &= \mathbb{E}[\tilde{\mathbf{s}}_n, \tilde{\mathbf{s}}_m] \end{aligned} \tag{19}$$

where the second equality is due to z-scoring [Bair et al., 2001].

Correlation of spike counts with binned spike trains. We bin spike trains in bins of length $L = [20, 50, 75, 100]$ ms and count spikes in bins. Spike counts are z-scored for each bin independently,

$$\tilde{sc}_{n,j}(b) = \frac{sc_{n,j}(b) - \langle sc_{n,j}(b) \rangle_j}{\sqrt{\mathrm{Var}_j\ (sc_{n,j}(b))}}, \tag{20}$$

where b is the index of the bin, $b = 1, 2, ..., B$, and the mean and variance are computed across trials. We define the vector of activation that concatenates binned spike trains across trials, $\tilde{\mathbf{sc}}_n = [\tilde{sc}_{n,1}(b_1), \tilde{sc}_{n,1}(b_2), ..., \tilde{sc}_{n,1}(B), \tilde{sc}_{n,2}(b_1), \tilde{sc}_{n,2}(b_2)..., \tilde{sc}_{n,J}(B)]$. After this procedure is repeated for all neurons, we compute Pearson's correlation coefficient between z-scored spike counts for each pair of neurons n and m,

$$r_{n,m}^{binned} = \frac{\mathbb{E}[\tilde{\mathbf{sc}}_n, \tilde{\mathbf{sc}}_m] - \mathbb{E}[\tilde{\mathbf{sc}}_n]\mathbb{E}[\tilde{\mathbf{sc}}_m]}{\sqrt{Var(\tilde{\mathbf{sc}}_n)Var(\tilde{\mathbf{sc}}_m)}} \tag{21}$$

Synchrony. Now, let $n, m \in \{1, ..., N\}$ be fixed, and we define the spike trains of neurons n and m in trial j by $f_j(t_k) := o_{n,j}(t_k)$ and $g_j(t_k) := o_{m,j}(t_k)$. The cross-correlation function measures the co-occurence of spikes,

$$ccg_j(\tau) = \frac{1}{2K} \sum_{\tau=-K}^{K} f_j(t_k) g_j(t_k + \tau) \Delta t_k. \tag{22}$$

We also measure autocorrelation functions for neurons n and m. The autocorrelation function for neuron n is the following:

$$acg_{1,j}(\tau) = \frac{1}{2K} \sum_{\tau=-K}^{K} f_j(t_k) f_j(t_k + \tau) \Delta t_k. \tag{23}$$

The cross-correlation function depends on the number of spikes of neurons n and m. To make it independent on the number of spikes, we normalize it with autocorrelations at zero time lag,

$$ccg_j^{norm}(\tau) = \frac{ccg_j(\tau)}{\sqrt{acg_{1,j}(0) acg_{2,j}(0)}}. \tag{24}$$

It is expected that an important part of the correlation is generated by the common input to the two neurons [Bair et al., 2001]. The correlation due to the common input is captured by the trial-invariant correlation function. The latter is computed from spike trains that are not from the same trial,

$$ccg_{j,p}^{inv}(\tau) = \frac{1}{2K} \sum_{\tau=-K}^{K} f_{j'}(t_k) g_j(t_k + \tau) \Delta t_k,\ j' \neq j, \tag{25}$$

where p is the permutation index and j' is defined with random permutation of the order of trials for the neuron n, while the neuron m keeps the correct order of trials. The permutation of the order of trials is iterated

$nperm = 100$-times, and we average the resulting functions across permutations, to get $ccg_j^{inv}(\tau)$. We capture the trial-specific correlation (i.e., the "noise" correlation) by subtracting the trial-invariant correlation from the overall correlation, and average across trials:

$$ccg^{noise}(\tau) = \frac{1}{J}\sum_{j=1}^{J}(ccg_j^{norm}(\tau) - ccg_j^{inv}(\tau)), \tag{26}$$

where J is the number of trials. The coefficient of synchrony is defined as the trial-specific cross-correlation function at zero time lag, $C = ccg^{noise}(0)$.

Data and code availability

The dataset and the code generated during this study are freely available here: https://github.com/VeronikaKoren/struct_pop_cellreports. All results and figures reported in the published article have been generated from the dataset and code in the repository. The code is published for documentation purposes and requires adjustments in order to run, such as editing the path for importing files, etc. In case of problems or questions, contact the Lead Contact.

Supplemental information

S1: Prediction accuracy during the target time window. Prediction accuracy during the test time window with a non-linear SVM.

S2: Decoding weights of the population model and of single neurons during target.

S3: Weights of the population model during the first and second half of the target time window.

S4: Synchrony within the group of informative and less informative neurons during the first and the second part of the test time window.

S5: Noise correlations of binned spike trains within the group of informative and less informative neurons during target.

S6: Coupling and correlations within and across coding pools during the first and second half of the test time window.

References

[Abbott and Nelson, 2000] Abbott, L. F., & Nelson, S. B. (2000). Synaptic plasticity: taming the beast. Nature neuroscience, 3(11s), 1178.

[Arieli et al. 1996] Arieli, A., Sterkin, A., Grinvald, A., & Aertsen, A. D. (1996). Dynamics of ongoing activity: explanation of the large variability in evoked cortical responses. Science, 273(5283), 1868-1871.

[Averbeck et al., 2006] Averbeck, B. B., Latham, P. E., & Pouget, A. (2006). Neural correlations, population coding and computation. Nature reviews neuroscience, 7(5), 358-366.

[Bair et al., 2001] Bair, W., Zohary, E., & Newsome, W. T. (2001). Correlated firing in macaque visual area MT: time scales and relationship to behavior. Journal of Neuroscience, 21(5), 1676-1697.

[Boerlin et al., 2013] Boerlin, M., Machens, C. K., & Denève, S. (2013). Predictive coding of dynamical variables in balanced spiking networks. PLoS computational biology, 9(11), e1003258.

[Britten et al., 1992] Britten KH, Shadlen MN, Newsome WT, Movshon JA (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance, J. Neurosci;12:4745.

[Britten et al., 1996] Britten KH, Newsome WT, Shadlen MN, Celebrini S, Movshon JA, (1996) A relationship between behavioral choice and the visual responses of neurons in macaque MT. Visual Neurosci., 13(1), 87-100.

[Cohen and Kohn, 2011] Cohen, M. R., & Kohn, A. (2011). Measuring and interpreting neuronal correlations. Nature neuroscience, 14(7), 811.

[Dayan and Abbott, 2001] Dayan, P., & Abbott, L. F. (2001). Theoretical neuroscience (Vol. 806). Cambridge, MA: MIT Press.

[Elsayed and Cunningham, 2017] Elsayed, G. F., & Cunningham, J. P. (2017). Structure in neural population recordings: an expected byproduct of simpler phenomena?. Nature neuroscience, 20(9), 1310.

[Fox and Raichle, 2007] Fox, M. D., & Raichle, M. E. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nature reviews neuroscience, 8(9), 700.

[Gilbert and Sigman, 2007] Gilbert, C. D., & Sigman, M. (2007). Brain states: top-down influences in sensory processing. Neuron, 54(5), 677-696.

[Gold and Shadlen, 2007] Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. Annual review of neuroscience, 30.

[Hansen et al., 2012] Hansen, B. J., Chelaru, M. I., & Dragoi, V. (2012). Correlated variability in laminar cortical circuits. Neuron, 76(3), 590-602.

[Koren and Denève, 2017] Koren, V., & Denève, S. (2017). Computational Account of Spontaneous Activity as a Signature of Predictive Coding. PLoS computational biology, 13(1), e1005355.

[Koren et al., 2019] Koren V, Andrei AR, Hu M, Dragoi V, Obermayer K (2019) Reading-out task variables as a low-dimensional reconstruction of neural spike trains in single trials. PLoS ONE 14(10): e0222649. https://doi.org/10.1371/journal.pone.0222649

[Luczak et al., 2013] Luczak, A., Bartho, P., & Harris, K. D. (2013). Gating of sensory input by spontaneous cortical activity. Journal of Neuroscience, 33(4), 1684-1695.

[Moreno-Bote et al., 2014] Moreno-Bote, R., Beck, J., Kanitscheider, I., Pitkow, X., Latham, P., & Pouget, A. (2014). Information-limiting correlations. Nature neuroscience, 17(10), 1410.

[Nienborg et al., 2012] Nienborg HR, Cohen MR, Cumming BG. (2012) Decision-related activity in sensory neurons: Correlations among neurons and with behavior, Annu. Rev. Neurosci., 35, 463–83.

[Nienborg and Cumming, 2014] Nienborg H, Cumming BG. (2014) Decision-related activity in sensory neurons may depend on the columnar architecture of cerebral cortex, J Neurosci.; 34(10):3579-85

[Newsome et al., 1989] Newsome WT, Britten KH, Movshon JA, (1989), Neuronal correlates of a perceptual decision. Nature, 341:52-54.

[Panzeri et al., 2017] Panzeri, S., Harvey, C. D., Piasini, E., Latham, P. E., & Fellin, T. (2017). Cracking the neural code for sensory perception by combining statistics, intervention, and behavior. Neuron, 93(3), 491-507.

[Rao and Ballard, 1999] Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature neuroscience, 2(1), 79.

[Shadlen, 1996] Shadlen MN. (1996). A computational analysis of the relation between neuronal and behavioral responses to visual motion, J Neurosci., 1486-1510

[Shadlen and Newsome, 1998] Shadlen, M. N., & Newsome, W. T. (1998). The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. Journal of neuroscience, 18(10), 3870-3896.

[Shadlen and Newsome, 2001] Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. Journal of neurophysiology, 86(4), 1916-1936.

[Shiozaki et al., 2012] Shiozaki, H. M., Tanabe, S., Doi, T., Fujita, I. (2012). Neural activity in cortical area V4 underlies fine disparity discrimination. The Journal of Neuroscience, 32(11), 3830-3841.

[Smith and Kohn, 2008] Smith MA, Kohn A. (2008), Spatial and temporal scales of neuronal correlation in primary visual cortex. J Neurosci., 28(48), 12591-603.

[Tsodyks et al., 1999] Tsodyks, M., Kenet, T., Grinvald, A., & Arieli, A. (1999). Linking spontaneous activity of single cortical neurons and the underlying functional architecture. Science, 286(5446), 1943-1946.

[Vapnik and Vapnik, 1998] Vapnik, V. N., & Vapnik, V. (1998). Statistical learning theory (Vol. 1). New York: Wiley.

[Waskom and Wagner, 2017] Waskom, M. L., & Wagner, A. D. (2017). Distributed representation of context by intrinsic subnetworks in prefrontal cortex. Proceedings of the National Academy of Sciences, 114(8), 2030-2035.

[Zohary et al., 1994] Zohary E., Shadlen MN., Newsome WT. (1994)., Correlated neuronal discharge rate and its implications for psychophysical performance, Nature 1994;370:140.