Research Articles: Behavioral/Cognitive

The effects of population tuning and trial-by-trial variability on information encoding and behavior

https://doi.org/10.1523/JNEUROSCI.0859-19.2019

Cite as: J. Neurosci 2019; 10.1523/JNEUROSCI.0859-19.2019

Received: 15 April 2019
Revised: 4 October 2019
Accepted: 11 October 2019

This Early Release article has been peer-reviewed and accepted, but has not been through the composition and copyediting processes. The final version may differ slightly in style or formatting and will contain links to any extended data.

Alerts: Sign up at www.jneurosci.org/alerts to receive customized email alerts when the fully formatted version of this article is published.

Copyright © 2019 the authors
The effects of population tuning and trial-by-trial variability on information encoding and behavior

Abbreviated: Dissecting population information and behavior

Ramon Nogueira¹²*, Nicole E. Peltier³, Akiyuki Anzai³, Gregory C. DeAngelis³, Julio Martínez-Trujillo⁴, and Rubén Moreno-Bote¹⁵

¹Center for Brain and Cognition & Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, 08002, Spain.
²Center for Theoretical Neuroscience, Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, New York, 10025, USA.
³Department of Brain and Cognitive Sciences, Center for Visual Science, University of Rochester, Rochester, New York, 14627, USA.
⁴Robarts Research Institute and Brain and Mind Institute, Schulich School of Medicine and Dentistry, Departments of Psychiatry, Physiology and Pharmacology, Western University, London, Ontario, N6A 3K7, Canada.
⁵Serra Húnter Fellow Programme, Universitat Pompeu Fabra, Barcelona, 08002, Spain.

*correspondence: rn2446@columbia.edu
Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank K.H. Britten, J.A. Movshon, W.T. Newsome, M.N. Shadlen and E. Zohary for making their data available in the ‘Neural Signal Archive’ (http://www.neuralsignal.org/), and W. Bair for maintaining this database. We thank Matthew Leavitt for useful comments on the manuscript. R.N. is supported by NSF’s NeuroNex program (DBI-1707398) and by Gatsby Charitable Foundation (GAT3419). R.M-B. is supported by BFU2017-85936-P and FLAGERA-PCIN-2015-162-C02-02 from MINECO (Spain) and the Howard Hughes Medical Institute (HHMI; ref 55008742). G.C.D., N.E.P., and A.A., as well as experiments in the DeAngelis lab, were supported by NEI grants EY013644 and EY016178. JMT was supported by grants from CIHR and NSERC.
Abstract

Identifying the features of population responses that are relevant to the amount of information encoded by neuronal populations is a crucial step toward understanding population coding. Statistical features such as tuning properties, individual and shared response variability, and global activity modulations could all affect the amount of information encoded and modulate behavioral performance. We show that two features in particular affect information: the modulation of population responses across conditions (population signal, PS) and the inverse population covariability along the modulation axis (projected precision, PP). We demonstrate that fluctuations of these two quantities are correlated with fluctuations of behavioral performance in various tasks and brain regions consistently across four monkeys (one female and one male *macaca mulatta* and two male *macaca fascicularis*). In contrast, fluctuations in mean correlations among neurons and global activity have negligible or inconsistent effects on the amount of information encoded and behavioral performance. We also show that differential correlations reduce the amount of information encoded in finite populations by reducing PP. Our results are consistent with predictions of a model that optimally decodes population responses to produce behavior.

Significance Statement

The last two or three decades of research have seen hot debates about what features of population tuning and trial-by-trial variability influence the information carried by a population of neurons, with some camps arguing, for instance, that mean pairwise correlations or global fluctuations are important while other camps report opposite results. In this study we identify the most important features of neural population responses that determine the amount of encoded information and behavioral performance by combining
analytic calculations with a novel non-parametric method that allows us to isolate the effects of different statistical features. We tested our hypothesis on four macaques, three decision-making tasks and two brain areas. The predictions of our theory were in agreement with the experimental data.

**Introduction**

Identifying the statistical features of neuronal population responses that affect the amount of encoded information and behavioral performance is critical for understanding neuronal population coding (Arandia-Romero et al., 2017; Panzeri et al., 2017). Changes in network states such as global modulations of activity (Harris and Thiele, 2011; Luczak et al., 2013; Gutnisky et al., 2017), as well as changes in correlated noise among neurons, have been shown to constrain the amount of information encoded by neuronal populations (Zohary et al., 1994; Ecker et al., 2014; Lin et al., 2015; Schölvinck et al., 2015). Indeed, it has been suggested that changes in neuronal tuning, global activity modulations, and noise correlations affect behavioral performance in certain conditions (Cohen and Newsome, 2008; Cohen and Maunsell, 2009; Mitchell et al., 2009; Gu et al., 2011; Verhoef and Maunsell, 2017; Ni et al., 2018). Additionally, theoretical studies have determined the exact pattern of noise correlations that limit the amount of encoded information for very large neuronal populations, so-called differential correlations (Moreno-Bote et al., 2014; Kanitscheider et al., 2015).

However, the aspects of neuronal responses that most directly affect the amount of encoded information for finite neuronal populations are not clear, since experimental designs often do not allow control over other statistical features that could potentially be involved. Furthermore, it is unknown whether the same features of population responses that affect the
amount of encoded information also impact behavioral performance (Arandia-Romero et al., 2017; Panzeri et al., 2017).

In a population of $N$ neurons, it is possible to define $N$ mean firing rates, $N(N+1)/2$ independent covariances, as well as features based on combinations of these quantities. What statistical features matter the most for encoding information? Do these same features also affect behavioral performance? To address these questions, we characterized the amount of encoded information and behavioral performance in three different tasks based on responses of multiple neurons recorded simultaneously in macaque monkeys. We examined neurons recorded in two different brain areas: the middle temporal area (MT), and area 8a in the lateral prefrontal cortex (LPFC). We developed a conditioned bootstrapping approach that allowed us to determine the features of neuronal population responses that influence the amount of information encoded and behavioral performance by generating fluctuations of one feature while keeping the other features constant. Using this approach, we found that the amount of information encoded in neuronal ensembles was primarily determined by two features: 1) the length of the vector joining the mean population responses in different experimental conditions (referred here as population signal, PS), and 2) the inverse population covariability projected onto the direction of the PS vector (projected precision, PP). Contrary to previous suggestions (Zohary et al., 1994; Kanitscheider et al., 2015; Lin et al., 2015; Ecker et al., 2016; Gutnisky et al., 2017), other statistical features, such as mean pairwise correlations (MPC) and global activity (GA) modulations, did not affect the amount of information encoded when PS and PP were kept constant. Strikingly, we also found that PS and PP are predictive of behavioral performance, whereas MPC and GA are not.

We show that MPC and GA are correlated with, and thus confounded with, PS and PP.

This can explain why previous studies report that changes in MPC (Zohary et al., 1994; Mitchell et al., 2009; Ecker et al., 2010; Renart et al., 2010; Gu et al., 2011; Ni et al., 2018) or
GA (Kanitscheider et al., 2015; Lin et al., 2015; Ecker et al., 2016; Gutnisky et al., 2017) modulate the amount of information encoded and behavioral performance, as those effects may have been due to concomitant changes in PS and PP, not MPC or GA per se. Finally, we link our results with previous theoretical work by showing that PP is reduced when differential correlations are added into the system. Our results are broadly consistent with predictions of a model that optimally decodes population responses to produce behavior.

Materials and Methods

1. Theoretical expression for the amount of encoded information

1.1. Theoretical decoding performance for an arbitrary linear classifier

If we assume that the activity of a neuronal population \( \mathbf{r} \) (\( N \) neurons) follows a multivariate Gaussian distribution, the covariance matrix for \( \mathbf{r} \) is stimulus-independent, the probability of presenting condition 1 is same as presenting condition 2 (\( p(C_1) = p(C_2) = 0.5 \)), and the classification is based on a linear projection of \( \mathbf{r} \) onto a scalar variable \( z = \mathbf{w}^T \mathbf{r} + \omega_0 \) (linear classifier) (see Fisher 1936), then the performance of the linear classifier can be expressed as

\[
DP = \Phi \left( \frac{1}{\sqrt{2} \omega} \right),
\]

where \( \Delta \mathbf{f} \equiv \mu_1 - \mu_2, \mu_1 = \mathbb{E}[\mathbf{r} \mid C_1], \mu_2 = \mathbb{E}[\mathbf{r} \mid C_2], \Sigma = \mathbb{E}[\mathbf{r} \mathbf{r}^T] \mid C_1] = \mathbb{E}[\mathbf{r} \mathbf{r}^T] \mid C_2], \Phi(\cdot) \) is the cumulative Gaussian function, and \( DP \) is the decoding performance. This expression gives the percentage of correct classifications (i.e. \( DP \)) that would be achieved by a linear classifier that reads out from a neuronal ensemble using an arbitrary set of weights, \( \mathbf{w} \).

1.2. Theoretical DP for the optimal linear classifier
By optimizing Eq. (2) with respect to $\omega$, we find that $\omega_{opt} \propto \Sigma^{-1}\Delta f$, which corresponds to the solution for the Linear Discriminant Analysis (LDA, Fisher 1936). We can substitute $\omega$ for $\omega_{opt}$ in Eq. (2) to find the DP for the optimal classifier:

$$DP = \Phi \left( \frac{1}{2} \sqrt{\Delta f^T \Sigma^{-1} \Delta f} \right).$$

(3)

The term inside $\Phi(\cdot)$ is known as $d' = \sqrt{\Delta f^T \Sigma^{-1} \Delta f}$ (Averbeck and Lee, 2006; Chen et al., 2006). It is a scalar quantity and therefore it remains invariant under unitary rotations of the reference frame. By rotating the original neuron-based orthogonal basis so that $\Sigma$ (and thus $\Sigma^{-1}$) becomes diagonal, we can express Eq. (3) as:

$$DP = \Phi \left( \frac{1}{2} |\Delta f| \sqrt{\sum_{i=1}^{N} \frac{\cos^2 \theta_i}{\sigma_i^2}} \right).$$

(4)

(see Moreno-Bote et al., 2014 for a similar derivation in the case of linear Fisher information). The first term, $|\Delta f|$, which we will refer to as Population Signal (PS), is the norm of the stimulus tuning vector $\Delta f$. It measures the overall modulation of the activity of the neuronal population as a function of the stimulus conditions. The second term, $\sqrt{\sum_{i=1}^{N} \frac{\cos^2 \theta_i}{\sigma_i^2}}$, which we will call Projected Precision (PP), is a function of $\theta_i$, the angle between the $i$-th eigenvector of the covariance matrix $\Sigma$ and the direction of the stimulus tuning vector $u_{\Delta f} \equiv \frac{\Delta f}{|\Delta f|}$, and $\sigma_i^2$, the $i$-th eigenvalue of the covariance matrix (see Fig. 1A,B). Thus, Projected Precision is the square root of the sum of squares of the eigenvalues of the precision matrix (inverse of the noise covariance matrix, $\Sigma$) projected onto the direction $u_{\Delta f}$. This rotation allows us to dissect the independent contributions of population tuning (first-order statistics) and trial-by-trial variability (second-order statistics) to the amount of information encoded by a neural population, which is not possible with the standard factorization of $d'$ into signal and noise components (see Discussion).
1.3. Theoretical DP for shuffled neuronal recording data

Trial shuffling is a commonly used technique to destroy the trial-by-trial shared fluctuations among neurons while preserving their mean firing rate to different experimental conditions. To understand the effects of noise correlations on DP, it is useful to derive theoretical expressions of DP for shuffled data. Shuffling the activity of single neurons across trials for a fixed stimulus condition transforms the covariance matrix $\Sigma$ into $\Sigma_{sh}$ (Averbeck et al., 2006), which is only approximately diagonal due to finite data size effects. The theoretical expression of DP for the shuffled data is given by

$$DP = \Phi \left( \frac{1}{\sqrt{2}} \frac{\omega^T \Delta f}{\sqrt{\omega^T \Sigma_{sh} \omega}} \right).$$

(5)

The optimal classifier is therefore $\omega_{opt} \propto \Sigma_{sh}^{-1} \Delta f$, and Eq. (4) becomes

$$DP = \Phi \left( \frac{1}{\sqrt{2}} |\Delta f| \sqrt{\sum_{i=1}^{N} \frac{\cos^2 \theta_i}{\sigma_i^2}} \right).$$

(6)

In this expression, PS remains the same as in Eq. (4) while the PP for the shuffled data becomes $\sqrt{\sum_{i=1}^{N} \frac{\cos^2 \theta_i}{\sigma_i^2}}$, where $\sigma_i^2$ is the variance of the response of each neuron and $\theta_i$ is the angle between the stimulus tuning axis $\mathbf{u}_{\Delta f}$ and each vector of the neuron-based orthogonal basis defining the original $N$-dimensional space of the neuronal activity. The above equation can also be expressed as

$$DP = \Phi \left( \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{N} \frac{(\Delta f)_i^2}{\sigma_i^2}} \right),$$

(7)

where $\sqrt{\sum_{i=1}^{N} \frac{(\Delta f)_i^2}{\sigma_i^2}}$ can be understood as the square root of the sum of the signal-to-noise ratio (SNR) of all neurons in the ensemble (Seung and Sompolinsky, 1993).

1.4. Theoretical DP under suboptimal read-outs
Here, we derive theoretical expressions of DP for two sub-optimal classifiers: one blind to response variability and another blind to pairwise correlations (Pitkow et al., 2015).

The variability-blind classifier takes into account the neuronal tuning but it is blind to any elements of the covariance matrix (i.e., it considers the covariance matrix to be proportional to the identity matrix). Thus, the readout weights for this classifier are given by $\omega \propto \Delta f$. Introducing this expression into Eq. (2), we find

$$DP = \Phi \left( \frac{1}{2} \frac{|\Delta f|}{\sqrt{\sum_{l=1}^{N} \hat{\sigma}_l^2 \cos^2 \hat{\theta}_l}} \right), \quad (8)$$

where $\hat{\sigma}_l^2$ and $\hat{\theta}_l$ are as defined before. The variability-blind classifier is directly related to the ‘axis of discrimination’ framework and associated d’ measure previously proposed (Cohen and Maunsell, 2009). In this framework, the axis of discrimination becomes precisely the difference in means vector $\Delta f$, which corresponds to a sub-optimal read-out direction that ignores both the shared and individual trial-by-trial variability of the population. The variability-blind PP is defined as $(\sum_{l=1}^{N} \hat{\sigma}_l^2 \cos^2 \hat{\theta}_l)^{-1}$, which is a lower bound on the real population’s PP. As shown in Extended Data Fig. 1-1, the analytical expression for encoded information based on the variability-blind classifier performs more poorly than the theoretical DP obtained using the full covariance matrix. Moreover, the degree of suboptimality of the variability-blind classifier grows with neural ensemble size.

The correlation-blind classifier (Pitkow et al., 2015) only takes into account the signal-to-noise ratio of each neuron, and ignores pairwise correlations among the neuronal ensemble (that is, the off-diagonal terms in the covariance matrix are zero). The set of weights for this classifier, then, is given by $\omega_l \propto \frac{\Delta f_l}{\sigma_l}$, which can also be written as $\omega \propto \Sigma_{\Delta f}$. By substituting this expression into Eq. (2), we obtain
\[ DP = \Phi \left( \frac{1}{2} |\Delta f| \sqrt{\frac{\sum N \cos^2 \hat{\theta}_i}{\sigma_i^2} - \frac{\sum N \cos^2 \hat{\theta}_i}{\sigma_i^2}} \right) \], \quad (9) \]

where \( \sigma_i^2 \) and \( \theta_i \) are as defined earlier and \( \bar{\sigma}_i^2 \) and \( \bar{\theta}_i \) correspond to the \( i \)-th eigenvalue and the angle between \( \mathbf{u}_i \) and the \( i \)-th eigenvector of the matrix \( \Sigma^{-1}_{x} \Sigma \Sigma^{-1}_{x} \), respectively. The correlation blind classifier, though suboptimal for correlation-intact data, is optimal for shuffled data from which pairwise correlations have been removed. Like in the variability-blind case, PS is equivalent to that for the optimal case, while the correlation-blind PP becomes \( \frac{\sum N \cos^2 \tilde{\theta}_i}{\sqrt{\sum N \cos^2 \tilde{\theta}_i}} \), which is also a lower bound on the true PP of a population of neurons.

1.5 Relationship to information-limiting correlations (differential correlations)

There is a well-known characterization of the type of correlations that limit information in large neuronal populations, the so-called differential correlations (Moreno-Bote et al., 2014; Kanitscheider et al., 2015). The fundamental difference between the formulation of differential correlations and our approach is that the expression for differential correlations applies to correlations that limit information in very large neuronal populations (\( N \to \infty \)), while the statistical features PS and PP are guaranteed, under some assumptions, to be the only features that affect information for finite neuronal populations (\( N < \infty \)). However, as information-limiting correlations can also have an impact on the amount of encoded information for finite \( N \), we examine in detail the relationship between these two approaches.

As described above, the task of our binary classifier is to assign one of two possible labels to a multi-dimensional pattern of activity. In most experimental situations, the two discrete labels are just particular instances of a continuous variable. For example, the direction of motion is a continuous variable, but the subject may be asked to discriminate
between two particular directions (e.g., left vs right) in a motion discrimination task. Although we may measure neuronal activity only for particular values of a stimulus variable \( s \) (e.g. \( s_1 \) and \( s_2 \); \( \Delta s \equiv (s_1 - s_2) \)), the population tuning curve, \( f \), is a continuously varying function of \( s \), \( f(s) \). Differential correlations (Moreno-Bote et al., 2014) are defined as a rank-one perturbation of the covariance matrix along the \( f'(s) \) direction

\[
\Sigma = \Sigma_0 + \epsilon f f^T, \tag{10}
\]

where \( \epsilon \) is the strength of the differential correlations, \( \frac{df(s)}{ds} = f'(s) \) is the derivative of the tuning curves and \( \Sigma_0 \) is the covariance matrix without differential correlations. As we are interested in discrimination tasks for which the stimulus categories are not contiguous, we take the approximation \( f'(s) \approx \frac{\Delta f}{\Delta s} \) for both stimulus conditions (see Fig. 1C).

We can assess the impact of information-limiting correlations on the amount of encoded information in a finite population of neurons by directly inserting Eq. (10) into Eq. (2) to obtain

\[
DP = \Phi \left( \frac{1}{2} \frac{\omega^T \Delta f}{\sqrt{\omega^T \Sigma_0 \omega + \epsilon (\omega f)^2}} \right). \tag{11}
\]

In general, the optimal classifier takes the form \( \omega_{opt} \propto \Sigma^{-1} \Delta f \) (see previous section). Under \( f'(s) \approx \frac{\Delta f}{\Delta s} \), this optimal readout direction is parallel to the optimal readout direction in the absence of differential correlations, \( \omega_{opt} = \Sigma_0^{-1} \Delta f \) (Moreno-Bote et al., 2014). However, as we will see now, decoding performance is strongly affected by the presence of differential correlations. To see why, note that the amount of encoded information when there are no differential correlations (\( \epsilon = 0 \)) is \( DP = \Phi \left( \frac{1}{2} d'_0 \right) \), where \( d'_0 = \sqrt{\Delta f^T \Sigma_0^{-1} \Delta f} \). In contrast, with differential correlations, the amount of encoded information is given by

\[
DP = \Phi \left( \frac{1}{2} \frac{d'_0}{\sqrt{1 + \frac{\epsilon}{\Delta^2 d'_0}}} \right). \tag{12}
\]
Therefore, when differential correlations are introduced into the system, the original $d'_0$ is reduced by a factor $\sqrt{1 + \frac{\varepsilon}{\Delta x^2} d'^2_0}$ and thus encoded information is reduced. When the neuronal ensemble is very large ($N \to \infty$), Eq. (12) becomes

$$DP = \Phi\left(\frac{1}{2} \sqrt{\frac{\Delta x}{\varepsilon}}\right),$$

(13)

assuming that $d'_0$ grows monotonically with $N$ (this is the case if $\Sigma_0$ does not contain differential correlations). Therefore, no matter how large the neuronal population is, information is bounded by $\Phi\left(\frac{1}{2} \sqrt{\frac{\Delta x}{\varepsilon}}\right)$.

To make more explicit the contribution of information-limiting correlations, PS, and PP, we can rewrite Eq. (12) as

$$DP = \Phi\left(\frac{1}{2} PS \times \frac{PP_0}{\sqrt{1 + \frac{\bar{\varepsilon}}{PP_0^2}}}\right),$$

(14)

where $PP_0$ is the projected precision evaluated using $\Sigma_0$ instead of $\Sigma$ and $\bar{\varepsilon} = \varepsilon |f'|^2$. The larger the strength of information-limiting correlations ($\varepsilon$), the larger the decrease in encoded information in the neuronal population. In other words, when information-limiting correlations are introduced into the system, the statistical feature PS is not affected, whereas PP is reduced by a factor $\sqrt{1 + \bar{\varepsilon} \times PP_0^2}$. The framework presented in this study is complementary to the framework described in (Moreno-Bote et al., 2014) by showing that, in finite neuronal populations, the contribution of trial-by-trial variability to encoded information is not necessarily dominated by differential correlations.

2. Comparison between theoretical and cross-validated decoding performance on experimental data.

We evaluated how well decoding performance of the theoretical expression ($DP_{th}$) agrees with the decoding performance of classifiers trained and tested on experimental data ($DP_{cv}$).
We plotted DP_{th} against DP_{cv} and performed type-II regression (orthogonal linear regression) on the data. We assessed the similarity between the two measures of decoding performance by computing the percentage of the variance in DP_{cv} that can be explained by DP_{th} as 
\[
\frac{\lambda_1}{(\lambda_1 + \lambda_2)} \times 100, 
\]
where \(\lambda_1\) and \(\lambda_2\) correspond to the variance captured by the first and second principal components of the DP_{th} vs DP_{cv} plot, respectively (Extended Data Fig. 1-2).

In order to account for possible idiosyncrasies from choosing a LDA when evaluating DP_{cv} and possible over-fitting on DP_{cv} due to the limited number of trials in our datasets, we also computed DP_{cv} using Logistic Regression (LR) on both the test and the training sets (Extended Data Fig. 1-2). In addition to the percentage of explained variance as the goodness-of-fit metric for the linear fit (% explained variance or E.V.), we also considered the slope and intercept parameters as complementary goodness-of-fit metrics (Extended Data Fig. 1-2) and obtained very similar overall results.

We compared how well the optimal expression (Eq. 4) approximated DP_{cv} with respect to the suboptimal expressions for DP_{th} (Eqs. 8 and 9) by plotting the ratio between the percentages of explained variance (Extended Data Fig. 1-1A). We performed the same analysis after shuffling across trials the activity of each individual neuron for a given stimulus condition (Extended Data Fig. 1-1B). Finally, we compared a linear discriminant analysis (LDA), a quadratic discriminant analysis (QDA) and a logistic regression (LR). Since LDA performed the best among the three (Extended Data Fig. 1-3), we chose LDA for computing DP_{cv} in all analyses presented here. For a detailed description of how these fits were performed for each dataset, see section 5.

3. Conditioned bootstrapping method

We performed an analysis based on bootstrapping to test the effects of fluctuations in PS, PP and other statistical quantities of neuronal responses on the amount of encoded information...
(DPcv) and behavioral performance. The conditioned bootstrapping method involves two steps: (1) generating fluctuations in statistical quantities through bootstrapping, and (2) conditioning by selecting bootstrap samples that produce fluctuations in certain statistical features but not in others.

For a particular set of trials (sub-dataset, see below), bootstrap samples were generated by randomly selecting $M$ trials ($M$ being the total number of trials for this particular sub-dataset) with replacement. For each bootstrap sample we calculated the following quantities:

- Behavioral performance of the animal, denoted as $B$ (see below for the definition of behavioral performance in each task).
- Theoretical decoding performance, $D_{P_{th}}$ (Eq. 4).
- Decoding performance of a cross-validated linear classifier (LDA), $D_{P_{cv}}$.
- Population Signal, PS, and projected precision, PP.
- Mean pairwise correlation, MPC, defined as the average of all pairwise response correlations among neuron pairs for a fixed stimulus condition.
- Global activity of the neuronal population, GA, defined as the mean neuronal activity across all neurons and trials.

We generated $2 \times 10^4$ bootstrap samples when analyzing behavior (Fig. 9) and $10^3$ when analyzing $D_{P_{cv}}$ (Figs. 4,6,8), as fitting the classifier is computationally expensive. From this procedure we obtained a distribution of each quantity listed above. We will denote the obtained value of the statistical feature $i$ for the bootstrap iteration $j$ as $x_{ij}$, and the fluctuation of $x_{ij}$ with respect to the median value of the distribution across bootstrap iterations as $\delta x_{ij} = x_{ij} - \bar{x}_i$, where $\bar{x}_i$ denotes the median of the distribution for that particular feature.
We used the median rather than the mean so that the method works just as robustly for skewed distributions as for symmetrical distributions, although strongly skewed distributions were typically not observed (Fig. 3). It is important to note that bootstrap samples could include non-unique trials. When evaluating statistical features computed using the covariance matrix, if the number of unique trials is small, singular matrices could be generated. Nevertheless, the number of unique trials in all datasets is typically 10 times larger than the number of neurons used in the analysis (see next section). Therefore, the rank of the bootstrapped covariance matrix is always bound by the number of neurons rather than by the number of unique trials subsampled for each bootstrap iteration. In other words, the probability of generating covariance matrices of full rank (number of neurons) in $2 \times 10^4$ (or $10^3$) bootstrap iterations is effectively one.

In order to assess what statistical features affect the amount of encoded information and behavioral performance the most, we evaluated the dependency of $\delta D_{cv}$ and $\delta B$ on $\delta x_i$, where, as above, $x_i$ represents a particular statistical feature of the neuronal responses. Here, dependency is measured by various quantities including correlation. A dependency across bootstrap iterations between $\delta D_{cv}$ and $\delta x_i$ (referred here as $p(\delta D_{cv}, \delta x_i)$), and between $\delta B$ and $\delta x_i$ ($p(\delta B, \delta x_i)$) could result from dependencies mediated by a third quantity $\delta x_k$. For instance, Eq. (4) predicts that $D_{cv}$ depends exclusively on $PS$ and $PP$, but since $PP$ decreases as the ensemble's noise (both individual and pairwise) increases, an inverse relationship between $\delta MPC$ and $\delta PP$ is expected. Therefore, we may find a correlation between $\delta D_{cv}$ and $\delta MPC$ as a result of their dependencies on $\delta PP$. To estimate the strengths of the dependencies $p(\delta D_{cv}, \delta x_i)$ and $p(\delta B, \delta x_i)$ that are not confounded by other variables, we developed a method for minimizing the correlation due to dependencies of $\delta D_{cv}$ (or $\delta B$) and $\delta x_i$ on a third variable $\delta x_k$. Among all generated bootstrap samples, we selected those that yielded $\delta x_k \approx 0$, i.e., $x_k$ values within the $\pm 15^{th}$ percentile of its median value. Based on the
selected samples, we computed $p(\delta D_{cv}, \delta x_l|\delta x_k \approx 0)$ (or $p(\delta B, \delta x_l|\delta x_k \approx 0)$), i.e. the dependency between $\delta D_{cv}$ (or $\delta B$) and $\delta x_l$ conditioned on $\delta x_k$. This method can easily be extended to multiple conditioning variables ($\delta x_l$, $\delta x_r$, etc) as long as there are enough bootstrap samples for the analysis that satisfy the conditions. Using this technique, we computed the following conditional dependencies:

- $p(\delta D_{cv}, \delta PS|\delta PP \approx 0, \delta MPC \approx 0, \delta GA \approx 0)$ and $p(\delta B, \delta PS|\delta PP \approx 0, \delta MPC \approx 0, \delta GA \approx 0)$

- $p(\delta D_{cv}, \delta PP|\delta PS \approx 0, \delta MPC \approx 0, \delta GA \approx 0)$ and $p(\delta B, \delta PP|\delta PS \approx 0, \delta MPC \approx 0, \delta GA \approx 0)$

- $p(\delta D_{cv}, \delta MPC|\delta PS \approx 0, \delta PP \approx 0, \delta GA \approx 0)$ and $p(\delta B, \delta MPC|\delta PS \approx 0, \delta PP \approx 0, \delta GA \approx 0)$

- $p(\delta D_{cv}, \delta GA|\delta PS \approx 0, \delta PP \approx 0, \delta MPC \approx 0)$ and $p(\delta B, \delta GA|\delta PS \approx 0, \delta PP \approx 0, \delta MPC \approx 0)$

Note that bootstrap samples were selected based on conditioned features, not on tested features, and the feature under study played no role on the conditioning procedure. For instance, to determine the real dependency between $\delta D_{cv}$ and $\delta PS$, we used only samples for which $\delta PP \approx 0, \delta MPC \approx 0, \delta GA \approx 0$, but $\delta PS$ was unconstrained. To evaluate $p(\delta D_{cv}, \delta x_l|\delta x_k \approx 0, \delta x_l \approx 0, \delta x_r \approx 0)$ (Figs. 6 and 10B) and $p(\delta B, \delta x_l|\delta x_k \approx 0, \delta x_l \approx 0, \delta x_r \approx 0)$ (Figs. 9, 10C and Extended Data Fig. 9-1), we used $10^3$ and $2 \times 10^4$ bootstrap iterations, respectively. With the conditioning criterion of ±15th percentile around the median, about 2.7% of the total number of bootstrap iterations (27 for dependencies on $D_{cv}$ and 540 for B) satisfied the conditioning on all three $x_k$, $x_l$ and $x_r$ from which to compute dependencies.
The dependencies \( p(\delta DP_{cv}, \delta x_i|\delta x_k \approx 0, \delta x_i \approx 0, \delta x_r \approx 0) \) and \( p(\delta B, \delta x_i|\delta x_k \approx 0, \delta x_i \approx 0, \delta x_r \approx 0) \) were evaluated by using two metrics: (1) the Pearson correlation coefficient (Extended Data Fig. 9-1B) and (2) the percent change (Figs. 4, 6, 9, 10B,C, and Extended Data Fig. 9-1A). Percent change in \( DP_{cv} \) with respect to \( \delta x_i \) (% change \( DP_{cv}^i \)) was defined as

\[
\text{% change } DP_{cv}^i = \frac{<\delta x_i^+_{\text{DF}_{cv}^i}> - <\delta x_i^-_{\text{DF}_{cv}^i}>>}{<\delta x_i^-_{\text{DF}_{cv}^i}>}, \tag{15}
\]

where \( <\delta x_i^+_{\text{DF}_{cv}^i}> \) and \( <\delta x_i^-_{\text{DF}_{cv}^i}> \) represent the mean \( DP_{cv} \) values for the bootstrap iterations that produced \( \delta x_i \) above \( (\delta x_i^+) \) and below \( (\delta x_i^-) \) its median value \( \bar{x}_i \), respectively. Similarly, percent change in \( B \) with respect to \( \delta x_i \) (% change \( B^i \)) was defined as

\[
\text{% change } B^i = \frac{<B_{\delta x_i^+}> - <B_{\delta x_i^-}>}{<B_{\delta x_i^-}>}, \tag{16}
\]

where \( <B_{\delta x_i^+}> \) and \( <B_{\delta x_i^-}> \) represent the mean \( B \) values for the bootstrap iterations that produced \( \delta x_i \) above \( (\delta x_i^+) \) and below \( (\delta x_i^-) \) its median value \( \bar{x}_i \), respectively. Because behavior \( (B) \) was quantified as the mean reaction time in the attentional task, the sign of \( p(\delta B, \delta x_i|\delta x_k \approx 0, \delta x_i \approx 0, \delta x_r \approx 0) \) was inverted so that negative fluctuations in \( B \) correspond to positive fluctuations in performance (Figs. 8 and 9B,C and Extended Data Fig. 9-1) for monkeys 2 and 3 (attentional task with recordings in LPFC 8a). Note that a negative correlation between the statistical feature \( \delta x_i \) and % change \( DP_{cv}^i \) and/or % change \( B^i \) signifies an anti-correlation between the particular statistical feature and the amount of encoded information and/or behavioral performance.

To confirm that our results still hold for different conditioning criteria, we also computed % change \( B \) using bootstrap iterations that restricted the range of the conditioning variables \( x_k, x_i \) and \( x_r \) to be within \( \pm 10^{th} \) and \( \pm 20^{th} \) percentile of their respective median.
values (Extended Data Fig. 9-1). We observed no significant differences between the results from the different conditioning criteria.

In Fig. 4, the dependencies $p(\delta DP_{cv}, \delta MPC)$ and $p(\delta DP_{cv}, \delta GA)$ were evaluated using % change $DP_{cv}^i$ (Eq. 15) without conditioning on the rest of statistical features of the neuronal responses. For each monkey and ensemble size, the reported dependencies corresponded to the median value across independent sub-datasets (see detailed description of each experimental dataset in section 5). Statistical significance was calculated with a two-sided Wilcoxon signed-rank test, with which we tested whether the median of the distribution of independently obtained values was significantly greater or less than zero. Likewise, in Fig. 5, the dependencies $p(\delta PS, \delta MPC)$, $p(\delta PS, \delta GA)$, $p(\delta PP, \delta MPC)$ and $p(\delta PP, \delta GA)$ were evaluated with the Pearson correlation between the bootstrap fluctuations of these statistical features. For each ensemble size, the reported dependencies corresponded to the median values across independent sub-datasets and monkeys and statistical significance was calculated with a two-sided Wilcoxon signed-rank test as before. For Fig. 8 and Extended Data 10-1G, the dependency $p(\delta B, \delta DP_{cv})$ was evaluated by computing the Pearson correlation coefficient between $\delta DP_{cv}$ and $\delta B$. For each monkey and ensemble size, the median value across independent sub-datasets was reported and statistical significance was calculated with a two-sided Wilcoxon signed-rank test as for Figs. 4 and 5. Error bars correspond to the 25th-75th percentile of a distribution of medians obtained by sampling with replacement from the distribution of independent values (bootstrap error bars; 1000 iterations). Error bars are only provided for visual guidance.

4. Neural population model

We built a neural population model that captures the basic statistical properties of the neural responses that we observed experimentally. The model includes limited-range correlations,
differential correlations, and realistic values for tuning sensitivity, firing rates and Fano factors. The model also generated choices based on optimal read out of information. We evaluated the simulation results by repeating all of the analyses described in the previous sections, including describing the relationships between PS, PP, MPC, GA with both encoded information and simulated behavioral performance.

4.1. Generative model

4.1.1 Population activity model

Our neural population model consisted of \( N = 1000 \) neurons. Each neuron's firing rate was modeled as a function of stimulus parameter \( s \) (such as motion direction) and stimulus strength, \( \nu \) (representing, for instance, motion coherence, which controls task difficulty). We considered two stimuli \( s_1 = +1 \) and \( s_2 = -1 \) (analogous to the two directions of motion around the discrimination boundary) at three stimulus strength levels, \( \nu = \{0.16, 0.32, 0.48\} \).

We defined mean firing rate \( \mu \) of neuron \( k \) as a linear function of stimulus parameters (i.e., tuning curve) as

\[
\mu_k(s, \nu) = m_{1k} s \nu + m_{0k},
\]

where \( m_{1k} \) and \( m_{0k} \) are the slope (sensitivity) and the baseline (spontaneous) firing rate of neuron \( k \), respectively. The slope parameters \( m_{1k} \) were drawn randomly from a normal distribution centered at the origin with a standard deviation of \( \sigma_{m_1} = 1.3 \) and \( m_{0k} \) was drawn from a gamma distribution with shape and scale parameters set at 20 and 1, respectively. The parameters of the distributions were chosen to approximate empirical distributions. In what follows, we will use \( \mu_k(s, \nu) \) and \( f_k(s, \nu) \) (defined earlier in Section 1) synonymously. We will use the terms spike count, firing rate and neuronal activity during a stimulus presentation interchangeably as the stimulus duration was set to 1 sec in our simulations.

Responses of neurons to identical stimuli vary from trial to trial, and the trial-by-trial variabilities are partially shared among neurons (noise correlation) (Cohen and Kohn, 2011;
We incorporated noise correlations into our model in the form of limited-range pairwise correlations between neurons $k$ and $l$ (Kanitscheider et al., 2015; Kohn et al., 2016) as

$$
\rho_{kl}^{gen} = \left(A \exp \left(-\frac{|m_k - m_l|}{\tau}\right) + C\right) (1 - \delta_{kl}) + \delta_{kl}, \quad (18)
$$

where $\delta_{kl}$ is the Kronecker delta ($A = 0.1$, $C = 0$ and $\tau = 1$ are used in our model).

Then, we defined a generative covariance matrix as

$$
\Sigma_{kl}^{gen}(s, v) = \sigma_k(s, v)\sigma_l(s, v)\rho_{kl}^{gen}, \quad (19)
$$

where $\sigma_k^2(s, v)$ represents the trial-by-trial variance of neuron $k$. We used $\sigma_k^2(s, v) = \eta \mu_k(s, v)$ ($\eta = 0.5$) so that Fano Factors of the model neurons are within the range of values typically found in experiments (Shadlen and Newsome, 1998; Arandia-Romero et al., 2016; Nogueira et al., 2018). Note that $\Sigma_{kl}^{gen}(s, v)$ is not yet the full model’s covariance matrix, $\Sigma_{kl}(s, v)$, which includes differential correlations and other components as derived in the next section.

Based on $\mu(s, v)$ and $\Sigma_{kl}^{gen}(s, v)$, we generated neuronal activity as follows. First, we chose $s$ and $v$ pseudo-randomly for each trial of 1 sec duration such that there were 100 trials per stimulus condition (600 trials in total per simulated recording session). Then, we generated a preliminary response vector $r'_j$ for each trial $j$ as

$$
r'_j = \mu(s_j, v_j) + Mz_j + \mu'\delta s_j, \quad (20)
$$

where $\Sigma_{kl}^{gen} = MM^T$, $z_j$ represents an $N$-dimensional vector whose elements are drawn from a zero-mean, unit-variance Gaussian distribution, $\mu'$ denotes the derivative of the population tuning curve $(f)$, which equals $m_1$ in our model, and $\delta s_j$ indicates a common sensory noise term drawn from a Gaussian distribution with zero mean and variance $\sigma_\delta^2 = \varepsilon$. The last term $\mu'\delta s_j$ introduces differential correlations to the population activity (Moreno-Bote et al., 2014; Kanitscheider et al., 2015). The generated population activity pattern $n_j$ was calculated by
applying multiplicative and additive global modulations to the vector $r'_j$ before putting it through a Poisson process that generated spike counts for each model neuron, i.e.

$$n_j \sim \text{Poisson} \left( g_j r'_j + g_j \right).$$  \hfill (21)

Here, $g_j$ is the global modulation factor drawn from a gamma distribution with scale and shape parameter $\theta_g = 10^{-5}$ and $k_g = 10^5$ respectively (such that $< g_j > = 1$ and $\sigma^2_g = 10^{-5}$). The global modulation incorporates the short and long timescale fluctuations in population activity often seen in vivo recordings (Ecker et al., 2014; Goris et al., 2014; Lin et al., 2015; Arandia-Romero et al., 2016). We generated a total of 10 simulated recording sessions that reproduced experimentally-realistic distributions of Fano factors and pairwise correlations (Extended Data Fig. 10-1C,D).

**4.1.2 Behavioral decision model**

We also modeled trial-by-trial behavioral choices using an optimal linear classifier that reads out the activity of the simulated population. For a given set of trials, simulated behavioral performance was calculated as the percentage of correct choices, just like the behavioral performance in the task involved monkeys 1 and 4. For modelling behavior, the optimal classifier was derived from an analytical expression (Eqs. 22 and 23 below) rather than from data fitting since the simulated data had a limited number of trials (200 trials per stimulus intensity) and may result in overfitting. By introducing Eq. (17) into the analytical expression for the optimal linear classifier derived in Section 1, we obtain

$$\mathbf{w}_{\text{opt}} = 2 \Sigma^{-1} \mathbf{m}_t,$$  \hfill (22)

where $\Sigma$ is the covariance matrix for our population model, which is given by

$$\Sigma_{kl} = (\sigma^2_g + 1) \left( \Sigma^{\text{gen}}_{kl} + \varepsilon \mu_k \mu_l \right) + \sigma^2_g \left( \mu_k + 1 \right) \left( \mu_l + 1 \right) + \left( \mu_k + 1 \right) \delta_{kl}. \hfill (23)$$

For each trial $j$, a decision variable was chosen as

$$d_j = \mathbf{w}_{\text{opt}}^T \mathbf{n}_j + w_0,$$  \hfill (24)
where $\omega_0 = -2\nu m_0^T \Sigma^{-1} m_0$. The behavioral choice $c$ for trial $j$ is, then, given by the sign of the decision variable $d$, i.e. $c_j = \text{sign}(d_j)$. Behavioral performance was evaluated by the number of correct classifications over the total number of trials.

4.2. Information encoded by the model

In this section, we derive a theoretical expression for decoding performance ($DP_{th}$) of our model. In our model, $f' = \frac{at}{A_0}$, and therefore $d'$ is given by Eq. (12). When $N \to \infty$

\[ d' = \frac{2\nu}{\sqrt{\nu}}, \quad (25) \]

and therefore

\[ DP = \Phi \left( \frac{\nu}{\sqrt{\nu}} \right). \quad (26) \]

Equation (26) suggests that, when the input signal is noisy on a trial-by-trial basis, $DP$ (the amount of information that can be extracted) of a very large neural ensemble does not saturate at perfect performance. Instead, it will approach an asymptote below one determined by the signal to noise ratio of the input signal $\frac{\nu}{\sqrt{\nu}}$ (see Extended Data Fig. 10-1F).

4.3. Analysis of model simulation data

We performed the analyses described in Sections 2 and 3 on our simulated neuronal and behavioral data. First, we examined how well our theoretical decoding performance ($DP_{th}$) approximated the cross-validated decoding performance ($DP_{cv}$) of a linear classifier (LDA) evaluated on the simulated data, in the same way as we did for the in vivo recording data. For a particular ensemble size, we randomly selected a neuronal ensemble from a recording session and stimulus intensity and computed $DP_{th}$ and $DP_{cv}$. For each ensemble size, stimulus intensity and simulated recording session, we repeated this process 20 times. We then fitted a
line to the relationship between $DP_{\text{th}}$ and $DP_{\text{cv}}$, and computed the percentage of variance in $DP_{\text{cv}}$ that was explained by $DP_{\text{th}}$, as described in Section 2 (Extended Data Fig. 10-1E).

We also conducted the conditioned bootstrapping analyses on the simulated data using the method described in Section 3. We examined the effects of bootstrap fluctuations in statistical features of the population activity on the amount of encoded information and behavioral performance. For each bootstrap iteration, we selected an ensemble of a given size (2, 4, 6, 8 and 10 units) and calculated the following quantities: theoretical decoding performance ($DP_{\text{th}}$), cross-validated decoding performance of a linear classifier ($DP_{\text{cv}}$), population signal (PS), projected precision (PP), mean pairwise correlations (MPC), global activity (GA), and behavioral performance (B). Both $DP_{\text{th}}$ and $DP_{\text{cv}}$ were evaluated for inferring the stimulus (either $s_1$ or $s_2$) from the activity pattern of the population model. This process was repeated 20 times for each of the recording sessions and stimulus intensities. The dependencies of $DP_{\text{cv}}$ and $B$ on different features of the neural activity were quantified by % change in $DP_{\text{cv}}$ and $B$ as defined in Eqs. (15) and (16), as well as by Pearson correlation coefficients. The results were averaged over repetitions. For each ensemble size, we thus obtained 30 independent samples (10 surrogate recording sessions × 3 stimulus intensities) and the median was reported. Statistical significance for the dependency between the variables $DP_{\text{cv}}$ (or $B$) and $\delta x_i$ in terms of $p(\delta DP_{\text{cv}}, \delta x_i|\delta x_k \approx 0, \delta x_l \approx 0, \delta x_r \approx 0)$ (or $p(\delta B, \delta x_i|\delta x_k \approx 0, \delta x_l \approx 0, \delta x_r \approx 0)$) for each ensemble size was calculated with a two-sided Wilcoxon signed-rank test, with which we tested whether the median of the distribution of 30 independently obtained values was significantly greater or less than zero. It is important to note that $p(\delta B, \delta x_i|\delta x_k \approx 0, \delta x_l \approx 0, \delta x_r \approx 0)$ as % change of surrogate behavior (Fig. 10C; Eq. 16) can be very small for relatively large surrogate ensemble sizes ($N \sim 100$) while it may exhibit strong Pearson correlation at the same time. In our analysis, however, we prefer using % change of behavior because of its interpretability and robustness to outliers in...
the distribution of bootstrap fluctuations. The same bias could potentially arise when evaluating \( p(\delta D_{cv}, \delta x_t | \delta x_r = 0, \delta x_l = 0, \delta x_r = 0) \) as % change of \( D_{cv} \). Nevertheless, in this case, this effect is typically masked by the strong dependency between the amount of encoded information and PS/PP, a consequence of using the same neuronal population to compute all the relevant quantities.

### 5. Experimental Design and Statistical Analysis

Our theory was tested on three different datasets obtained from four monkeys, each of which involved simultaneously recorded units (from 2 to \(~50\) units). One dataset involved pairs of middle temporal (MT) neurons from a monkey (monkey 1) performing a coarse direction discrimination task (Zohary et al., 1994). The second dataset involved lateral prefrontal cortex (LPFC, area 8a) neurons recorded from two monkeys (monkeys 2 and 3) that were trained to perform a spatial attention task (Tremblay et al., 2015). The final dataset involved recordings from groups of MT neurons in a monkey (monkey 4) trained to perform a fine direction discrimination task. The details for each experiment are described below.

#### 5.1 Coarse direction discrimination task with recordings in area MT

This dataset has been previously described (Zohary et al., 1994), and is freely available at the Neural Signal Archive (http://www.neuralsignal.org, nsa2004.2). Data from one female adult macaque monkey (Macaca mulatta) are included in this dataset. The animal was trained to perform a coarse direction discrimination task (Britten et al., 1992), and pairs of single neurons were recorded in area MT. The monkey performed a coarse direction discrimination task, in which a noisy random-dot motion stimulus moved in one of two opposite directions. In each trial, a stimulus was presented for 2 seconds. The presented direction of motion was either the preferred or null direction of the recorded neurons. If the preferred directions of the...
two neurons differed substantially, the axis of discrimination was set to the preferred-null axis of the best-responding neuron. The null direction was defined as the direction opposite to the preferred direction. The strength of the motion signal was manipulated by controlling the fraction of dots that moved coherently from one frame to the next (motion coherence), while the remaining ‘noise’ dots were randomly replotted on each video update. When motion coherence was 0%, all dots moved randomly, and there was no coherent direction in the net motion. When motion coherence was 100%, all dots moved coherently in either the preferred or null direction for the pair of neurons. A range of coherences between 0 and 100% was used to adjust task difficulty and measure neural and behavioral sensitivity. In each trial, the monkey was presented randomly with either the preferred or null direction of motion at a particular coherence. The monkey’s task was to report the direction of the net motion by making a saccade to one of two choice targets. The psychometric function averaged across sessions is shown in Fig. 7A.

Each trial started with the appearance of a fixation point. After 2 seconds of stimulus presentation, the random dot pattern and the fixation point disappeared, and two choice targets appeared on the screen, corresponding to the two possible directions of motions (preferred or null) (see Fig. 2A). The monkey made a saccadic eye movement to one of the targets to report its perceived direction of motion. Data from this experiment (Zohary et al., 1994) consisted of 41 recording sessions in which pairs of single-units were simultaneously recorded in area MT. Note that only sessions with random-dot motion stimuli generated from a variable seed were used.

Since stimulus strength was controlled by the motion coherence parameter, we subdivided each recording session according to the coherence presented in each trial. Trials belonging to the 0% coherence condition were discarded because stimulus identity and correct behavioral performance were not defined for this condition. In order to evaluate how
accurate our analytical approximation was for the amount of information encoded by the neural population, we plotted $\text{DP}_n$ against $\text{DP}_c$ for each recorded pair of neurons and coherence condition and fitted the data points with a type-II linear regression. The percentage of variance explained by the first principal component of the data represented the goodness-of-fit (see section 2).

For the analysis involving behavior, we also discarded trials with coherences that elicited a mean behavioral performance greater than 98% correct to avoid ceiling effects (25.6%, 51.2% and 99.9% coherence were eliminated by this criterion). This was done because the bootstrap distributions for behavioral performance ($B$) would have very low variability for these high coherences, and therefore calculating $p(\delta B, \delta x^r | \delta x, \approx 0, \delta x^l \approx 0)$ may be highly biased or even undefined. The criterion for discarding easy trials (98% percent correct) corresponds to eliminating conditions that are more than two standard deviations away from the mean of a cumulative Gaussian fit, since $\Phi(2\sigma) \approx 0.98$. After splitting recording sessions by coherence and discarding 0% and high coherences as described above, we obtained 187 independent sub-datasets, each of which corresponded to a set of trials for a particular coherence level and recording session. The mean number of trials per sub-dataset was 75, ranging from 30 to 231 trials. For the analysis, we used a trial-by-trial population activity vector whose entries corresponded to the spike count from the entire 2 second stimulus duration for each neuron. All statistical features (PS, PP, GA and MPC) as well as DP (and $\text{DP}_b$) were calculated using this time window.

From each sub-dataset, we generated bootstrap distributions for the following quantities: theoretical decoding performance ($\text{DP}_b$), cross-validated (trained and tested on different sets of trials) decoding performance of a linear classifier ($\text{DP}_c$), population signal (PS), projected precision (PP), mean pairwise correlation (MPC), global activity (GA), and behavioral performance ($B$). Behavioral performance was defined as the fraction of correct
choices of the monkey on each sub-dataset. Both DP_{th} and DP_{cv} were calculated by inferring
the motion direction (preferred or null) presented to the monkey on a trial-by-trial basis from
the simultaneously recorded activity of a neuronal pair.

Significance for the dependency \( p(\delta D_{P_{cv}}, \delta x_i | \delta x_k \approx 0, \delta x_l \approx 0, \delta x_r \approx 0) \) (or
\( p(\delta B, \delta x_i | \delta x_k \approx 0, \delta x_l \approx 0, \delta x_r \approx 0) \)) for the whole experiment (187 independent sub-
datasets) was calculated with a two-sided Wilcoxon signed-rank test, with which we tested
whether the median of the distribution of 187 independently obtained values was significantly
greater or less than zero.

5.2. Spatial attention task with recordings in LPFC (prearcuate area 8a)

This dataset is described in detail in (Tremblay et al., 2015). Two male monkeys
(Macaca fascicularis), both 6 years old (Monkey “F”, 5.8 Kg; Monkey “JL”, 7.5 Kg),
contributed to this dataset. In each monkey, a 96-channel “Utah” multielectrode array was
chronically implanted in the left caudal lateral prefrontal cortex. The multielectrode array was
inserted on the prearcuate convexity posterior to the caudal end of the principal sulcus and
anterior to the arcuate sulcus, a region cytoarchitectonically known as area 8a (Bullock et al.,
2019). The monkeys were trained to covertly sustain attention to one of four Gabor stimuli
(target) presented on a screen while ignoring the other three Gabor stimuli (distractors) (Fig.
2B). At the beginning of each trial, a cue indicated which of the four Gabor stimuli was the
target (cue period, 363 ms). After the cue period, all four Gabor stimuli appeared on the
screen, which marked the start of the attentional period. The attentional period ended after a
variable delay (585-1755 ms) when one or two Gabor stimuli changed orientation by 90°.
Only correct “Target” trials were used in the analysis reported here (see (Tremblay et al.,
2015)).
The mean number of correct “Target” trials per recording session was 207 (range: 172 to 224) for monkey “JL” and 221 trials (range: 198 to 246) for monkey “F”. The mean number of simultaneously recorded units was 56 (range: 53 to 61) for monkey “JL” and 54 (range: 44 to 66) for monkey “F”. Neuronal recordings included both single and multiunits. The mean percentage of single units over the total number of simultaneously recorded units was 44% and 40% for monkeys “JL” and “F”, respectively. Because units with very low firing rates precluded reliable statistical analysis, we excluded units with firing rates below 1 Hz for all subsequent analyses. After this exclusion, the mean number of simultaneously recorded units was 51 (range: 50 to 53) for monkey “JL” and 50 (range: 42 to 59) for monkey “F”. The shortest attentional period used was 585 ms; therefore, we defined a fixed attentional time window of 585 ms starting at the end of the cue period. In this way, the firing rate of all units was calculated over the same time window. All statistical features (PS, PP, GA and MPC) as well as DP (and DPth) were calculated using this time window.

To maximize the statistical power of our analysis, we created a larger number of independent sub-datasets as follows. For each recording session, we randomly selected 21, 10, 7, 5, and 4 non-overlapping ensembles of size 2, 4, 6, 8, and 10, respectively, and repeated this process 5 times. Since the smallest number of simultaneously recorded units for both monkeys was 42 units, we chose values that maximized the number of non-overlapping ensembles for each size.

In order to evaluate how accurate our analytical approximation was for the amount of information encoded by the neural population for a particular ensemble size, we plotted DPth against DPcv for each sub-dataset and fitted the data points with a type-II linear regression. The percentage of variance explained by the first principal component of the data represented the goodness-of-fit of our approximation (see section 2).
For each sub-dataset and bootstrap iteration, we calculated the following quantities: $\text{DP}_{\text{th}}, \text{DP}_{\text{cv}}, \text{PS}, \text{PP}, \text{MPC}, \text{GA}$, and behavioral performance (B). Behavioral performance was quantified as the mean reaction time (RT) across trials (either the original sub-dataset or a bootstrap iteration) (see Fig. 7B,C for RT distributions from monkeys 2 and 3, respectively).

Both $\text{DP}_{\text{th}}$ and $\text{DP}_{\text{cv}}$ were calculated by inferring the monkey's location of attention on a trial-by-trial basis from the simultaneously recorded activity of an ensemble. For decoding purposes, we considered two binary classification tasks. In the first task, the decoder classified the locus of attention as being on the right or the left side of the screen. In the second task, the decoder classified the locus of attention as being on the upper half or the lower half of the screen. The reported $\text{DP}_{\text{cv}}$ (and $\text{DP}_{\text{th}}$) values are averages over the two decoding tasks.

For a given sub-dataset, the reported dependency relationship $p(\delta \text{DP}_{\text{cv}}, \delta x_i | \delta x_k \approx 0, \delta x_l \approx 0, \delta x_r \approx 0)$ (or $p(\delta B, \delta x_l | \delta x_k \approx 0, \delta x_l \approx 0, \delta x_r \approx 0)$) was evaluated as the mean across the 5 iterations and the two classification tasks (vertical and horizontal). For each ensemble size and monkey, the reported dependency $p(\delta \text{DP}_{\text{cv}}, \delta x_i | \delta x_k \approx 0, \delta x_l \approx 0, \delta x_r \approx 0)$ (or $p(\delta B, \delta x_l | \delta x_k \approx 0, \delta x_l \approx 0, \delta x_r \approx 0)$) was the median across recording sessions and non-overlapping ensembles of units (independent sub-datasets). We obtained 84, 40, 28, 20 and 16 (the number of non-overlapping ensembles $\times$ 4 recording sessions per monkey) independent values for ensemble sizes of 2, 4, 6, 8, and 10 units, respectively, and assessed the median and tested significance. Significance was calculated with a two-sided Wilcoxon signed-rank test of whether the median of the distribution of independently obtained values for each ensemble size and monkey was significantly above or below zero.

5.3. Fine direction discrimination task with recordings in MT
This dataset was obtained specifically for this analysis, as part of a larger series of ongoing studies of how MT neurons represent local motion in the presence of background optic flow. One male adult macaque monkey (*Macaca mulatta*) was used in this experiment. The animal was surgically implanted with a circular head holding device, a scleral coil for measuring eye movements, and a recording grid (Gu et al., 2006, 2008). The animal was trained to perform a fine direction discrimination task (described below) with water as a reward for correct performance. Eye movements were measured and controlled at all times. Neuronal activity in MT/V5 was recorded with 24-channel linear electrode arrays (V-probes, Plexon Inc). Spike waveforms were acquired by a Blackrock Cerebus system. All experimental procedures conformed to National Institutes of Health guidelines and were approved by the University Committee on Animal Resources at the University of Rochester.

In each recording session, a V-probe was inserted into MT/V5 and was allowed to settle for ~30 minutes. We then performed standard tests (DeAngelis and Uka, 2003) to map receptive fields and to measure the direction tuning of the recorded units. These measurements were used to determine the location and the size of the stimulus aperture such that it was larger than the receptive fields of the units under study. Note, however, that stimulus motion was not tailored to the preferred directions and speeds of the recorded neurons; a fixed set of stimulus velocities was used across sessions for the fine discrimination task. A total of 75 units were recorded across 3 sessions. The mean number of simultaneously recorded units was 25, ranging from 24 to 27. Most recordings included multi-unit activity as well as 2-4 well-isolated single units. Because units with very low firing rates precluded reliable statistical analysis, we excluded units with firing rates below 1 Hz for all subsequent analyses. Despite the exclusion, the mean number of simultaneously recorded units remained at 25 (ranging from 24 to 27).
The monkey was trained to perform a fine direction discrimination task. A pattern of random dots, presented within a circular aperture, moved upward in the visual field with either a rightward or leftward component. The animal’s task was to report whether the perceived motion was up-right or up-left by making a saccade to one of two choice targets. The motion stimulus was presented at 100% coherence in one visual hemi-field, and was localized and sized according to the receptive fields of the recorded units. Stimulus motion within the aperture followed a Gaussian velocity profile with a standard deviation of 333 ms and a duration of 2 s. Once the monkey fixated for 200 ms, the motion stimulus appeared and began to move. The stimulus was presented stereoscopically at zero disparity, such that motion appeared in the plane of the display. The experimental protocol involved 7 directions of motion relative to vertical: -12º, -6º, -3º, 0º, 3º, 6º and 12º, where negative and positive values correspond to leftward and rightward motion, respectively. After stimulus presentation was completed, two saccade targets appeared, 5 deg to the right and left of the fixation target, and the monkey reported his perceived direction of motion by making a saccade to one of the targets (Fig. 2C). Because stimuli were presented at 100% coherence, task difficulty was controlled by the direction of motion with respect to vertical (see Fig. 7D). The mean number of trials per session was 742 (range: 735 to 756). In each recording session, the same number of trials were presented for each direction of motion.

Since direction of motion controlled the difficulty of the task, this parameter was considered to be the stimulus strength. To control for stimulus strength, we divided data from each recording session according to motion direction. The task of our decoder is to correctly classify each trial as having rightward or leftward motion direction; therefore, trials from each recording session were split into 4 independent sub-datasets (±12º, ±6º, ±3º and 0º stimulus directions relative to vertical). As described above for the coarse discrimination task, we subsequently discarded stimulus conditions that were either ambiguous (0º motion
direction; correct behavioral performance is not defined) or too easy (±12° directions; mean behavioral performance ≥ 0.98). The analysis was thus performed on 6 independent sub-datasets (3 recording sessions × 2 absolute motion directions). To quantify neuronal activity, we computed firing rates in a time window that included ±1 standard deviation around the peak of the Gaussian velocity profile of the stimulus (666 ms window width). All statistical features (PS, PP, GA and MPC) as well as DP (and DPnh) were calculated using this time window.

To increase statistical power, we created a larger number of independent sub-datasets by sampling randomly non-overlapping ensembles of particular sizes. Namely, we considered ensemble sizes of 2, 4, 6, 8, and 10, which yielded 12, 6, 4, 3, and 2 non-overlapping ensembles of units, respectively. We repeated this sampling procedure 20 times. Since the smallest number of simultaneously recorded units was 24, we chose values that maximized the number of non-overlapping ensembles for each size.

We evaluated how accurate DPnh was with respect to DPcv by following the same procedure as for monkeys 2 and 3 (see previous section). For each sub-dataset and bootstrap iteration, we calculated the following quantities: DPnh, DPcv, PS, PP, MPC, GA, and behavioral performance (B). For this task, behavioral performance was defined as the fraction of correct choices for that particular sub-dataset. Both DPnh and DPcv were calculated by inferring whether motion direction was leftward or rightward of vertical for each trial based on the neuronal activity pattern. For a particular randomly constructed ensemble, the reported dependency relationship \( p(\delta DP_{cv}, \delta x_l | \delta x_k \approx 0, \delta x_l \approx 0, \delta x_r \approx 0) \) (or \( p(\delta B, \delta x_l | \delta x_k \approx 0, \delta x_l \approx 0, \delta x_r \approx 0) \)) was evaluated as the mean across the 20 iterations. For each ensemble size, the reported dependency \( p(\delta DP_{cv}, \delta x_l | \delta x_k \approx 0, \delta x_l \approx 0, \delta x_r \approx 0) \) (or \( p(\delta B, \delta x_l | \delta x_k \approx 0, \delta x_l \approx 0, \delta x_r \approx 0) \)) was the median across recording sessions, stimulus strengths, and non-overlapping ensembles of units. We used 72, 36, 24, 18 and 12 (the number of non-
overlapping ensembles × 3 independent recording sessions × 2 stimulus strengths) independent values to assess the median and test its significance for ensemble sizes of 2, 4, 6, 8, and 10 units, respectively. Significance was calculated with a two-sided Wilcoxon signed-rank test for whether the median of the distribution of independent values for each ensemble size was significantly greater or less than zero.

Data and software availability

The datasets generated in this study and the code used for their analysis are available from the corresponding author upon reasonable request.

Results

We start by showing that fluctuations of MPC and GA influence the amount of information encoded in population activity, consistent with some previous studies. We then demonstrate that these effects of MPC and GA are eliminated when PS and PP are controlled for, whereas isolated fluctuations of PS and PP strongly predict the amount of information encoded in the population. Finally, we demonstrate that fluctuations of PS and PP are correlated with behavioral performance, and we compare this finding with predictions of an optimal decoding model.

Mean pairwise correlations and global activity correlate with the amount of encoded information
MPC and GA have been thought to modulate the amount of information encoded in neuronal populations (Zohary et al., 1994; Renart et al., 2010; Ecker et al., 2011, 2016; Kanitscheider et al., 2015; Lin et al., 2015; Gutnisky et al., 2017). We tested the correlation between these statistical features and the amount of information in three different datasets consisting of simultaneously recorded units (2 to ~50 single-units or multiunits) in four monkeys, two brain areas, and three tasks: i) pairs of middle temporal (MT) neurons recorded while the animal performed a coarse motion discrimination task (Zohary et al., 1994) (monkey 1, Fig. 2A), ii) lateral prefrontal cortex (LPFC, area 8a) neurons recorded with a Utah array while two animals performed an attentional task (Tremblay et al., 2015) (monkeys 2 and 3, Fig 2B), and iii) MT neurons recorded with a 24-channel linear array while the animal performed a fine motion discrimination task (monkey 4, Fig. 2C). Behavioral performance in these tasks was defined as the percentage of correctly reported directions of motion (monkeys 1 and 4) or as the mean reaction time when correctly detecting a change in orientation of the attended Gabor patch (monkeys 2 and 3). The amount of encoded information for each dataset was quantified as the cross-validated decoding performance (DP$_{cv}$) of a linear classifier that reads out the activity of the recorded neuronal population to predict $i$) which of two opposite directions of motion was presented in the coarse motion task (monkey 1), $ii$) which Gabor patch was cued in the attention task (monkeys 2 and 3), or $iii$) whether the stimulus motion was right or left of vertical in the fine direction discrimination task (monkey 4). In each case, the classifier was trained on the activity of neuronal ensembles recorded in area MT for the motion tasks or LPFC (area 8a) for the attention task (see Methods for details).

To evaluate whether the amount of encoded information (DP$_{cv}$) was modulated by MPC and GA, we used a non-parametric method to produce fluctuations of MPC and GA generated by bootstrapping trials from each neural recording dataset (Fig. 3A,B). We then examined how fluctuations of MPC and GA affected DP$_{cv}$ (Fig 4; see Methods). We found
that an increase in MPC tended to produce a decrease in DPcv, therefore reducing the amount of encoded information. This was largely consistent across tasks and animals. For instance, for monkey 1 (Fig. 4A), an increase of MPC significantly reduced the amount of encoded information by -0.05% (Wilcoxon signed-rank test, \( P = 2.6 \times 10^{-3} \); see Methods). Qualitatively similar results were found for monkeys 2-4 (Fig. 4B-D). In contrast, positive bootstrap fluctuations of GA tended to increase the amount of encoded information for some animals, particularly those from which we made PFC recordings. For monkey 2 (Fig. 4B), bootstrap fluctuations of GA produced significant positive changes in DPcv by 0.91% (ensemble size 2, Wilcoxon signed-rank test, \( P = 6.9 \times 10^{-15} \); see Methods). For monkey 3 (Fig. 4C), results were qualitatively similar to those obtained for monkey 2, but for monkeys 1 and 4 no significant effect was observed. Some of the size differences in the effects of MPC and GA on decoding performance may be due to differences in behavioral tasks and/or brain areas (MT vs LPFC 8a). In particular, the population of LPFC 8a neurons was somewhat unbalanced with respect to spatial preference (~70% contralateral preference), whereas direction selectivity in populations of MT neurons is generally quite balanced. This might explain the overall stronger effects of MPC and GA fluctuations on encoded information in the PFC animals (monkeys 2 and 3), because those fluctuations have a stronger chance to spontaneously align with the individual, but unbalanced, sensory preferences.

Although the observed correlations between MPC, GA, and the amount of encoded information are generally consistent with those from previous experimental and theoretical studies (Zohary et al., 1994; Renart et al., 2010; Ecker et al., 2011, 2016; Kanitscheider et al., 2015; Lin et al., 2015; Gutnisky et al., 2017), the effects are rather small and somewhat inconsistent. It is important to note that the results reported in Fig. 4 are obtained by correlating MPC and GA with DPcv using all bootstrap samples. Thus, the correlation between MPC and GA fluctuations with DPcv could be mostly due to other statistical features.
that co-fluctuate with MPC or GA. Indeed, as shown below, MPC and GA are confounded with PP and PS, and do not exhibit significant correlations with DPcv once PP and PS are held constant.

Population signal (PS) and projected precision (PP) determine the amount of information encoded in neuronal population responses

In this section, we identify the statistical features of population tuning and trial-by-trial variability that affect the amount of encoded information (as measured by DPcv), and we test our predictions on the neural data described above.

Under some common assumptions (see Methods), an analytical expression for the theoretical decoding performance (DPth) of a linear classifier can be derived (Averbeck and Lee, 2006). This expression takes the form $\text{DP}_{\text{th}} = \Phi(d'/2)$, where $\Phi(x)$ is the cumulative normal function and $d' = \sqrt{\Delta f^T \Sigma^{-1} \Delta f}$ is the signal-to-noise ratio generalized for a population of neurons (Averbeck and Lee, 2006; Chen et al., 2006; Gutnisky et al., 2017). The term $\Delta f$ is the vector joining the means of the population responses in the two stimulus conditions and $\Sigma$ is the stimulus-invariant noise covariance matrix of the neuronal population. To understand the roles of population tuning and trial-by-trial variability in determining the amount of encoded information, it is useful to re-write this equation by rotating the original $N$-dimensional neural response space along the eigenvectors of the covariance matrix as follows,

$$\text{DP}_{\text{th}} = \Phi\left(\frac{1}{2} |\Delta f| \sqrt{\sum_{i=1}^{N} \cos^2 \theta_i \hat{\sigma}_i^2}\right),$$  

where $\theta_i$ represents the angle between the $i$-th eigenvector of the covariance matrix and the unitary direction defined by the stimulus vector $\Delta f$, and $\hat{\sigma}_i^2$ denotes the $i$-th eigenvalue of the
covariance matrix (see Fig. 1). Equation (1) reveals that the amount of information encoded by a neural population can be divided into two independent components: the first-order contribution from the population tuning (population signal, $PS = |\Delta f|$) and the second-order contribution from the trial-by-trial variability (projected precision, $PP = \sqrt{\sum_{l=1}^{N} \frac{\cos^2 \theta_l}{\sigma_{l}^2}}$). It is important to remark that $d'$ is often expressed as the signal to noise ratio of the linear classifier $\left( d' = \frac{\Delta \mu_z}{\sigma_z} \right)$, where under optimality, the signal is given by $\Delta \mu_z = \Delta f^T \Sigma^{-1} \Delta f$, and the noise is given by $\sigma_z = \sqrt{\Delta f^T \Sigma^{-1} \Delta f}$. While this $d'$ formulation aims to differentiate signal from noise with respect to the classifier’s decision variable (Averbeck and Lee, 2006) (see Fig. 1), it does not separate the contributions of firing rate modulation (PS) and trial-by-trial variability (PP) of the population to the amount of encoded information, since $\Delta \mu_z$ and $\sigma_z$ both contain terms associated with PS and PP.

There are two salient points to make regarding the relationships between PS/PP and other measures of neuronal population information. First, it is important to emphasize that PP is not simply the projection of correlated noise (covariance matrix) onto the vector joining the means of the population responses in the two stimulus conditions, $\Delta f$ (see Methods for additional details). For instance, it is possible to have a finite projection of correlated noise onto $\Delta f$ but PP can nevertheless be very large. This can happen if there is an eigenvector of the covariance matrix that is not orthogonal to $\Delta f$ with a very small eigenvalue, indicating the presence of highly correlated sets of neurons. Second, previous studies have identified the exact pattern of noise correlations that limit information for very large neuronal populations, the so-called differential correlations (Moreno-Bote et al., 2014; Kanitscheider et al., 2015). Differential correlations add neuronal activity fluctuations along $f'$ (parallel to the discrimination direction, $\Delta f$) and they are known to be the only type of correlations that can limit information for very large neuronal populations ($N \to \infty$). We show (see Methods) that
adding differential correlations into the system reduces PP but does not affect PS. The
reduction of PP by differential correlations is $\sqrt{1 + \bar{\epsilon} \times PP_0^2}$, where $\bar{\epsilon} = |f'|^2$ and $PP_0$ is the
projected precision evaluated on the original covariance matrix without information-limiting
correlations. The variable $\epsilon$ directly measures the strength of differential correlations. Thus,
differential correlations can have a strong effect on PP, but they are not the only factor in
determining PP for finite neuronal populations.

We find that Eq. (1) provides a very good approximation to the empirical measure of
the amount of encoded information, as it accounts for more than 96% of the variance in the
cross-validated decoding performance of a linear classifier ($DP_{cv}$) for all datasets (Extended
Data Fig. 1-2). In addition, Eq. (1) is a better approximation to $DP_{cv}$ than analytical
expressions that simplify the covariance structure by removing the off-diagonal terms
(Averbeck and Lee, 2006; Averbeck et al., 2006; Pitkow et al., 2015) or calculations that
assume a covariance proportional to the identity matrix (“axis of discrimination” method,
Cohen and Maunsell 2009; see Methods, Extended Data Fig. 1-1). Finally, linear
discriminant analysis (LDA) produced better matches to $DP_{cv}$ across different ensemble sizes
and monkeys than logistic regression (LR) and quadratic discriminant analysis (QDA) did,
justifying our use of LDA to evaluate the amount of information encoded by neural
populations ($DP_{cv}$; Extended Data Fig. 1-3).

Having identified analytically the two features of population tuning and trial-by-trial
variability that should determine the amount of encoded information (i.e., PS and PP), we
now test the central prediction that information depends exclusively on PS and PP and does
not depend on other statistical features, such as MPC and GA, unless those features are
correlated with PS and PP. Indeed, we found that bootstrap fluctuations of MPC, GA, PS and
PP showed substantial correlations among them (Fig. 5; see Methods). For instance, for
monkey 1, the correlations between fluctuations of PS and GA, PP and MPC, and PP and GA
were all highly significant (ensemble size 2, $\rho_{PS,GA} = 0.18$, $P = 2.4 \times 10^{-24}$; $\rho_{PP,MP} = -0.16$, $P = 1.1 \times 10^{-11}$; $\rho_{PP,GA} = -0.091$, $P = 1.1 \times 10^{-12}$), whereas the correlation between fluctuations of PS and MPC was not statistically significant (ensemble size 2, $\rho_{PS,MP} = 0.019$, Wilcoxon signed-rank test, $P = 0.15$; see Methods). For monkeys 2, 3 and 4, results were qualitatively similar, although for monkeys 2 and 3 the correlation $\rho_{PP,MP}$ was weaker. Therefore, we suspected that the relationships between MPC (and GA) and the amount of encoded information (see Fig. 4) would be reduced, if not eliminated, once the modulations in MPC (GA) were made independent of PS, PP and GA (MPC) by selecting bootstrap samples with constant PS, PP and GA (MPC) values.

We tested these predictions by using a conditioned bootstrapping method to evaluate the effect of fluctuations in one feature on the amount of encoded information while the values of other features are kept constant (Fig. 3C; Methods). For example, to determine the effects of MPC on $DP_{cv}$ independent of PS, PP, and GA, we generated bootstrap samples from the original dataset and selected those bootstrap iterations that produced PS, PP and GA values within a narrow range around their medians. Then, for the selected data, we evaluated the % change in $DP_{cv}$ introduced by the bootstrap fluctuations in MPC. Any dependence of $DP_{cv}$ on MPC, then, would not be explainable by the dependencies of $DP_{cv}$ on PS, PP, or GA. This method can be used to test the effects of any statistical feature on the amount of encoded information by fixing other features to their representative values, thus isolating the individual effects. There are other possible approaches to isolating the effects of different statistical features on the amount of encoded information ($DP_{cv}$), such as a model-dependent analysis based on generalized linear models (GLMs). However, due to the linearity assumptions underlying GLMs and the potential collinearity between several statistical features, it is preferable to use a model-independent approach based on conditioned bootstrapping (Methods).
Applying this conditioned bootstrapping approach to our datasets, we found that bootstrap fluctuations of PS and PP greatly affected the amount of encoded information, whereas fluctuations of MPC and GA produced negligible changes in \( \text{DP}_{\text{cv}} \) across different tasks, animals, and neuronal ensemble sizes (Fig. 6). As anticipated, when fluctuations of MPC and GA were conditioned on PS and PP, the effects reported in Fig. 4 largely vanished (Fig. 6, right column). For monkey 1 (Fig. 6A), fluctuations of PS and PP produced significant positive changes in the amount of encoded information by 5.43\% (Wilcoxon signed-rank test, \( P = 3.2 \times 10^{-37} \); see Methods) and 2.75\% (\( P = 1.3 \times 10^{-35} \)), respectively. In contrast, fluctuations of MPC and GA had no significant effects on \( \text{DP}_{\text{cv}} \) (\( P = 0.33 \) and \( P = 0.16 \), respectively). For monkeys 2-4 (Fig. 6B-D), results were qualitatively similar to those obtained for monkey 1 (see Extended Data Fig. 6-1). We also compared the difference in % change in \( \text{DP}_{\text{cv}} \) for all pairs of statistical features, and found again that PS and PP had the most significant effects on the amount of encoded information (Extended Data Fig. 6-2). The observed dependency of \( \text{DP}_{\text{cv}} \) on PP is generally weaker than that on PS. This could be explained at least partially by the fact that PP is a second order statistic and therefore is expected to be noisier than the first order statistic, PS, when estimated from limited experimental data.

Population signal and projected precision are also the strongest predictors of behavioral performance

We have shown that PS and PP are the major statistical features of population activity that affect the amount of encoded information in neuronal populations. But do these features also have an impact on behavioral performance? We reasoned that if downstream neurons that contribute to behavioral choices can extract most of the information encoded by a neuronal
population, then behavioral performance should depend on PS and PP while it is largely independent of MPC and GA. However, since there could be many layers of non-linear computations between the information encoding stage and the final behavioral choice, finding such a relationship is not guaranteed a priori.

Behavioral performance was measured as the percentage of correctly reported directions of motion (monkeys 1 and 4) (Fig. 7A,D) or as the mean reaction time (monkeys 2 and 3) (Fig. 7B,C). We first confirmed that modulations in the amount of encoded information (DPcv) over different bootstrap samples were significantly correlated with the corresponding changes in behavioral performance (Fig. 8). Although the reported correlations are weak, they are nevertheless consistent with the predictions of a model that reads out neuronal population activity optimally to produce behavioral choices (see next section). We then performed the conditioned bootstrap analysis on behavioral performance separately for each stimulus strength or task difficulty in order to control for trivial dependencies (see Methods). We found that bootstrap fluctuations of PS and PP predicted significant modulations of behavioral performance across different datasets and ensemble sizes, whereas changes in behavioral performance produced by fluctuations in MPC and GA were either weak or inconsistent across different animals and ensemble sizes (Fig. 9). For example, bootstrap fluctuations of PS or PP, while keeping all other statistical features fixed, predicted significant changes in behavioral performance for monkey 1 (Fig. 9A; behavioral change predicted by PS alone: 1.59%, Wilcoxon signed-rank test, $P = 3.6 \times 10^{-14}$; PP alone: 0.53%, $P = 1.4 \times 10^{-3}$). However, the average change in behavioral performance predicted by bootstrap fluctuations of either MPC or GA alone were very small and not statistically significant for this animal (MPC: 0.06%, Wilcoxon signed-rank test, $P = 0.33$; GA: -0.03%, $P = 0.44$). We obtained qualitatively similar results for other monkeys and tasks (Fig. 9B-D; see Extended Data Fig. 9-2), except that bootstrap fluctuations of GA produced significant
fluctuations in behavioral performance for monkey 2, but not monkey 3, in the attentional
task. We also analyzed the data using Pearson’s correlation coefficient in place of % change
in DPcv and obtained qualitatively similar results (Extended Data Fig. 9-1). These results
were also robust across different conditioning thresholds (Extended Data Fig. 9-1; see
Methods) and for all pairs of statistical features (Extended Data Fig. 9-3).

In summary, we find that PS and PP are the primary statistical features of neural
activity that exhibit modulatory effects on behavioral performance, while MPC and GA show
little or no consistent effects. These results are in line with those described earlier for the
amount of encoded information (Fig. 6).

An experimentally-constrained model accounts for the findings

To determine whether our experimental findings (Fig. 9) are consistent with an optimal
readout of a neuronal population much larger than the observed ensembles, we developed a
model constrained by the observed statistical properties of the neural responses (Fig. 10A,
Extended Data Fig. 10-1A and Methods). The model simulates a number of trials (M trials in
total) in which a stimulus (s = {+1, −1}) drives a large population consisting of N = 1000
neurons. Each neuron has a linear tuning curve with a slope (m) drawn from a normal
distribution. Neuronal correlations combined limited-range dependencies (Kanitscheider et
al., 2015; Kohn et al., 2016), differential correlations (Moreno-Bote et al., 2014), and
multiplicative and additive gains (Arandia-Romero et al., 2016). These correlations were
generated though a sequence of steps as follows. For each trial j, a vector r′j (size N = 1000
neurons), was drawn from a multivariate Gaussian with mean μ and covariance matrix Σgen.
The response vector r′j was then corrupted with sensory noise, δsj, which introduces
differential correlations and limits the amount of information encoded in large populations of

neurons (Moreno-Bote et al., 2014). Multiplicative and additive gains were then applied (Arandia-Romero et al., 2016) \((g_j)\) and a final Poisson step converted each response rate into an observed spike count \(n_j\) (Kanitscheider et al., 2015) (see Methods). Fano factors (FF) that roughly matched typical values found experimentally were used (Shadlen and Newsome, 1998; Arandia-Romero et al., 2016; Nogueira et al., 2018). Finally, based on the neuronal activity generated by the network, we generated surrogate choices \((c_j)\) by optimally decoding the whole population of 1000 simulated neurons \((c_j = \text{sign}(\omega_{opt}^T n_j + \omega_0))\) on a trial-by-trial basis. The presence of differential correlations ensures that the amount of encoded information could not scale indefinitely with the number of neurons.

To compare to our experimental data, we randomly selected small ensembles (up to 10 model neurons) out of the full population (consisting of 1000 neurons) to test how fluctuations in PS and PP of the small ensembles correlate with encoded information and simulated behavior of the full network (Extended Data Fig. 10-1). We repeated the analyses performed on the experimental data (Figs. 6 and 9). Consistent with the experimental data, PS and PP computed from these small ensembles of model neurons are the only features that correlated significantly with \(D_{psv}\) and surrogate behavioral performance, while MPC and GA did not (Fig. 10B,C). We also found that changes in the model’s simulated behavioral performance associated with bootstrap fluctuations of PS and PP were quite small (Fig. 10C; ensemble size 2; surrogate behavioral change predicted by PS alone: 0.32%, Wilcoxon signed-rank test, \(P = 2.12 \times 10^{-6}\); PP alone: 0.028%, \(P = 0.049\)) and roughly similar in magnitude to the experimental results (see Fig. 9). As expected, when neuronal ensembles of up to 100 neurons were sampled from the full population, fluctuations of PS and PP had a larger effect on surrogate behavioral performance (ensemble size 100; surrogate behavioral change predicted by PS alone: 0.90%, Wilcoxon signed-rank test, \(P = 1.73 \times 10^{-6}\); PP
alone: 0.21%, \( P = 4.38 \times 10^{-3} \). Finally, correlations between fluctuations of \( D_{P_v} \) and simulated behavioral performance were weak (Extended Data Fig. 10-1G) and comparable to those observed in the experimental data (Fig. 7). In summary, the weak correlations that we observed experimentally between PS/PP and behavioral performance are broadly consistent with the weak dependencies between neuronal activity and choice observed using single-neuron analyses (Britten et al., 1996; Uka and DeAngelis, 2004; Nienborg and Cumming, 2009; Haefner et al., 2013), and are also consistent with our hypothesis that behavior is generated by an optimal read out of a much larger population of neurons.

**Discussion**

Identifying the statistical features of neural population responses that determine the amount of encoded information and predict behavioral performance is essential for understanding the link between neuronal activity and behavior. Based on data collected from two brain areas and three behavioral tasks, we identified two critical features: the length of the vector joining the mean responses of the population of neurons across two conditions to be distinguished (population signal, PS), and the inverse of the population covariability projected onto the direction of that vector (projected precision, PP). We found that changes in PS and PP are significantly correlated with the amount of encoded information and behavioral performance, but MPC and GA are not, once PS and PP are held fixed. Our experimental results are consistent with predictions of a neuronal population model with realistic neuronal tuning, variability, and correlations that is read-out optimally to generate behavioral choices.

Although previous studies have examined how input signals are represented in the average activity of neurons (Hubel and Wiesel, 1959; Mountcastle et al., 1967; Renart and van Rossum, 2012), described the types of neuronal correlations that can benefit or harm
encoded information (Zohary et al., 1994; Abbott and Dayan, 1999; Averbeck et al., 2006; Ecker et al., 2011; Moreno-Bote et al., 2014; Verhoef and Maunsell, 2017), and cautioned about directly linking MPC with information (Abbott and Dayan, 1999; Averbeck et al., 2006; Moreno-Bote et al., 2014), they have not clearly separated the primary features of neuronal population activity that constrain the amount of encoded information and affect behavioral performance. As shown above, the traditional interpretation of neuronal population sensitivity, $d'$, considers the signal and noise components of the classifier’s decision variable. However, contrary to this classic interpretation, our framework separates the contributions of population tuning (first order statistic) and trial-by-trial variability (second order statistic) of the neuronal responses to the amount of encoded information. Separating these two contributions of $d'$ is important for several reasons. First, as we have shown, the effects of statistical features such as MPC and GA on the amount of encoded information and behavioral performance can be understood in terms of the contributions by the first (PS) and second-order (PP) statistical components. This suggests that PS and PP are fundamental statistical features that underlie many other statistical features that affect the amount of information encoded and behavioral performance. Second, when addressing how learning affects neuronal coding, for example, it is important to understand not just how $d'$ changes during learning but also how learning alters the tuning curves of individual neurons (i.e., PS) and the coordination among neurons along relevant task axis (i.e., PP). Therefore, distinguishing the first and second-order contributions may help us understand circuit-level models of learning.

Previous studies have also identified the exact pattern of noise correlations that limit information for very large neuronal populations (Moreno-Bote et al., 2014; Kanitscheider et al., 2015), but as we have shown, the precision of the neuronal code for finite neuronal populations depends only partially on differential correlations, which act by reducing PP.
Thus, one important contribution of our study is identifying PS and PP as the critical features of neuronal activity that modulate the amount of encoded information and predict behavioral performance for finite populations sizes.

Another contribution of this study involves demonstrating that other features of neuronal activity, namely MPC and GA, have little or no impact on the amount of encoded information and behavioral performance once PS and PP are controlled for (Zohary et al., 1994; Shadlen and Newsome, 1998; McAdams and Maunsell, 1999; Kanitscheider et al., 2015; Lin et al., 2015; Arandia-Romero et al., 2016; Ecker et al., 2016; Gutnisky et al., 2017; Verhoef and Maunsell, 2017). We developed a novel approach for studying correlations among multiple variables while eliminating the effects of other variables (conditioned bootstrapping method). This method uses bootstrapping to generate continuous distributions of multiple statistical features so that fluctuations of some features can be computed for a subset of bootstrap samples that yielded no fluctuations in other features. This method allowed us to isolate the effects of a particular feature on both the amount of encoded information and behavioral performance. Thus, our approach clearly differs from the traditional trial shuffling method that destroys all dependencies in the data to examine the effects of correlations (Romo et al., 2003; Kohn and Smith, 2005; Tremblay et al., 2015; Leavitt et al., 2017). It also differs from maximum entropy models that generate surrogate data while fixing the first and second moments of the generated distribution to desired values (Elsayed and Cunningham, 2017).

Previous studies have suggested that MPC and GA affect behavioral performance (Cohen and Newsome, 2008; Cohen and Maunsell, 2009; Mitchell et al., 2009; Gu et al., 2011; Ni et al., 2018). On the surface, these studies appear to be at odds with our main finding that only PS and PP should affect the amount of encoded information or behavioral performance. Indeed, we find that bootstrap fluctuations of MPC and GA have little effect on
the amount of encoded information or behavioral performance when PS and PP are kept constant. Our results suggest that the previous studies found effects of MPC and GA on behavioral performance because these variables are themselves correlated with PS and PP (see Fig. 5). In fact, there may be many other statistical features of neuronal responses that seemingly influence the amount of encoded information and behavioral performance, but such relationships could be explained as a byproduct of correlations of these statistical features with PS and PP.

Although theoretical research has proposed that fluctuations of GA can modulate the amount of encoded information (Kanitscheider et al., 2015; Lin et al., 2015; Ecker et al., 2016), a recent study based on large neuronal populations in monkey primary visual cortex found that the amount of information does not vary with large fluctuations in GA (Arandia-Romero et al., 2016). Another recent study found a modest, but significant, increase in decoding performance when population activity decreased (Gutnisky et al., 2017), which appears to be inconsistent with our theoretical prediction. Again, however, some of the previous reports of positive effects of GA may have arisen from correlations between fluctuations in GA and PS. Indeed, we have found a highly significant correlation between PS and GA. Therefore, an important question is whether there is a residual effect of GA on the amount of encoded information in the previous studies after the contributions of PS and PP are eliminated. Our theory predicts that there should not be, and the conditioned bootstrapping approach that we have developed would be useful for teasing apart the statistical features (such as PS and PP) that fundamentally affect the amount of encoded information and behavioral performance from heuristic parameters such as GA that may have only secondary effects on information or behavior.

In conclusion, based on information metrics that are readily applicable to neuronal data, we developed a theory-driven analysis that has identified the statistical features of
neural population responses that modulate the amount of information encoded by cortical neuronal populations. We found an excellent agreement between the theory and the experimental results, which indicates that the assumptions of our approach are reasonable. A critical finding is that PS and PP are the primary features that correlate with behavioral performance and that one must be careful when interpreting effects of other statistical features that may co-vary with PS or PP. Finally, we showed that our results are consistent with predictions of a model that optimally decodes population responses to produce behavior.

References


Ni AM, Ruff DA, Alberts JJ, Symmonds J, Cohen MR (2018) Learning and attention reveal a...


Legends

**Figure 1. The information encoded by a neuronal ensemble is fully characterized by population signal and projected precision.** (A) The trial-by-trial joint activity of a network consisting of $N$ neurons can be characterized by an ellipsoid embedded in an $N$-dimensional space (a representative population of two MT neurons from monkey 4 is shown here; see Fig. 2 and Methods). The covariance matrix and mean activity of the population determine the shape and location of the ellipsoids corresponding to stimulus 1 ($s_1$; green) and stimulus 2 ($s_2$; blue). The linear classifier is characterized by a linear boundary that separates the two clouds of points as well as possible (red line). The difference in mean value for the decision variable $z$ and its standard deviation are represented by $\Delta \mu_z$ and $\sigma_z$, respectively (signal-to-noise ratio of $d'$). (B) Graphical depiction of population signal (PS) and projected precision.
PS is defined as the norm of the tuning vector, $\Delta f$, which corresponds to the distance between the mean responses associated with $s_1$ and with $s_2$ (distance between the centers of the green and blue ellipsoids). PP is calculated from the angles ($\tilde{\theta}_i$) between each eigenvector of the covariance matrix and the tuning vector ($\Delta f$) (orientation of each axis of the ellipsoid with respect to $\Delta f$) as well as from their eigenvalues ($\tilde{\sigma}_i^2$; length of each axis). Greater information is encoded when the longest axis of the ellipsoid becomes orthogonal to the direction of $\Delta f$. Mean pairwise correlations (MPC) is defined as the mean across all neuronal pairs of the trial-by-trial co-variability (Pearson correlation) for a given set of trials. Global activity (GA) is defined as the mean activity across trials neurons and neurons for a set of trials. (C) When information-limiting correlations are added into the system, the amount of encoded information is harmed due to a reduction of PP. Because PS accounts for the contribution of the population tuning on encoded information, it is not affected by the presence of information-limiting correlations. The magnitude of the reduction in PP will depend on the strength of information-limiting correlations ($\epsilon$), the sensitivity of the population tuning to stimulus changes ($|f'|^2$) and the projected precision without information-limiting correlations (PP$_0$) (see Methods). See also Extended Data Figs. 1-1, 1-2 and 1-3.

**Figure 2.** Three behavioral tasks used to test theoretical predictions in macaque monkeys. **(A)** One monkey performed a coarse direction discrimination task (monkey 1) while pairs of units were recorded in the middle temporal (MT) area (Britten et al., 1992). After stimulus presentation (random dots moving toward the preferred or null direction of the neurons under study), the monkey reported the direction of motion by making a saccade to one of two targets. Difficulty was controlled by varying the percentage of coherently moving dots in the stimulus. **(B)** Two monkeys performed an attentional task (monkeys 2 and 3)
while ~50 units were recorded simultaneously from the lateral prefrontal cortex (LPFC, area 8a) (Tremblay et al., 2015). Four Gabor patterns were presented on the screen and the task was to make a saccade to the attended location after a change in orientation of the cued Gabor. (C) One monkey performed a fine direction discrimination task (monkey 4) while ~25 units were recorded simultaneously from area MT. After presentation of a fully coherent random dot stimulus, the monkey had to report whether dots moved leftward or rightward of vertical by making a saccade to one of two targets. Difficulty was controlled by making the left/right component of motion very small. See Methods for details.

Figure 3. Dependencies between encoded information and statistical features of neuronal responses can be evaluated by the conditioned bootstrapping method. (A, B) By subsampling trials with replacement (bootstrap) from the original dataset, distributions of the values of encoded information (DP\textsubscript{\textsuperscript{cv}}) (A) and statistical features of the neuronal activity (B) are generated. Example distributions of the values of population signal (PS), projected precision (PP), mean pairwise correlations (MPC) and global activity (GA) for monkey 4 (ensemble size = 10 units). (C) A conditioning analysis is performed to determine the impact of fluctuations of each feature of the neuronal response on information by selecting subsets of bootstraps in which the other features are held close to their distributional medians. For instance, to study the effect of MPCs on information, bootstraps are selected such that PS, PP, and GA are fixed near their median values (dark regions in the distributions). If bootstrap fluctuations in MPC (while fixing PS, PP, and GA) do not modulate DP\textsubscript{\textsuperscript{cv}}, we can conclude that MPC do not play a role in the amount of information encoded by a neural network. This approach is also used to study the effects of bootstrap fluctuations of neuronal activity on behavioral performance.
Figure 4. Mean pairwise correlations and global activity correlate with encoded information. (A) Percentage change in the amount of encoded information (DPcv) related to fluctuations of mean pairwise correlations (MPC; blue) and global activity (GA; orange). Fluctuations of MPC and GA are produced through a bootstrap process in which virtual instances of the same experiment are generated by sampling trials with replacement from a particular dataset (see Fig. 3B and Methods; not to be confused with the conditioning analysis described in Fig. 3C). Smaller values of MPC produce significantly larger values of DPcv; thus, they are negatively correlated. Data are shown from pairs of neurons recorded in MT during a coarse direction discrimination task (monkey 1; Figure 2A). (B, C) Percentage change in DPcv, as a function of ensemble size (2, 4, 6, 8 and 10 units), for recordings from LPFC 8a during an attentional task (monkeys 2 and 3; see Figure 2B and Methods). GA has a strong modulatory effect on encoded information on both monkeys. (D) Analogous results for MT ensembles recorded during a fine direction discrimination task (monkey 4; see Figure 2C and Methods). MPC has a consistent negative effect on encoded information. In all panels, error bars correspond to the 25th -75th percentile of the distribution of bootstrap medians and significant deviations from zero are indicated by high contrast colored bars (Wilcoxon signed rank test, $P < 0.05$).

Figure 5. Bootstrap fluctuations in population signal, projected precision, mean pairwise correlations, and global activity are correlated. (A) Examples of how bootstrap fluctuations (without conditioning) in GA are correlated with bootstrap fluctuations in PS (left panel), and fluctuations of MPC are correlated with PP (right panel) for pairs of neurons recorded in MT during a coarse direction discrimination task (monkey 1; Figure 2A). (B)
Median Pearson correlation between fluctuations in PS and MPC (red), PS and GA (green), PP and MPC (blue) and PP and GA (orange) for each monkey and ensemble size (see Methods). Error bars correspond to the 25th -75th percentile of the distribution of bootstrap medians and significant deviations from zero are indicated by high contrast colored bars (Wilcoxon signed rank test, P < 0.05). Correlations between these different statistical features of the neuronal activity are likely to explain reported dependencies of encoded information on MPC and GA.

Figure 6. Encoded information depends mainly on population signal and projected precision once other variables are controlled. (A) Percentage change in the amount of encoded information (DP_{cv}) when changes in one statistical feature of neuronal responses are isolated by the conditioned bootstrapping method described in Fig. 3C (population signal (PS): red; projected precision (PP): green; mean pairwise correlation (MPC): blue; global activity (GA): orange). Only PS and PP produce significant changes in DP_{cv} when the other three features are keep constant (e.g., PS is tested by using bootstrap samples for which the values of PP, MPC, and GA are all close to their respective median values). Data are shown from pairs of neurons recorded in MT during a coarse direction discrimination task (monkey 1; see Figure 2A and Methods). (B, C) Percentage change in DP_{cv}, as a function of ensemble size (2, 4, 6, 8 and 10 units), for recordings from LPFC 8a during an attentional task (monkeys 2 and 3; see Figure 2B and Methods). (D) Analogous results for MT ensembles recorded during a fine direction discrimination task (monkey 4; see Figure 2C and Methods). In all panels, error bars correspond to the 25th -75th percentile of the distribution of bootstrap medians and significant deviations from zero are indicated by high contrast colored bars (Wilcoxon signed rank test, P < 0.05). See also Extended Data Figs. 6-1 and 6-2.
**Figure 7. Definitions of behavioral performance for each task.** (A) Psychometric curve averaged across recording sessions for monkey 1. Proportion of choices in favor of the preferred direction as a function of motion coherence (positive values of coherence correspond to the preferred direction of motion). Grey region: s.e.m. (B, C) Distributions of reaction times for monkeys 2 and 3. For the attentional task, reaction time is defined as mean time (across trials in a particular dataset) from the change in stimulus orientation until the saccade to the cued Gabor pattern. (D) Psychometric curve averaged across recording sessions for monkey 4. Proportion of rightward choices as a function of the direction of motion with respect to vertical (positive and negative values denote motions with a rightward and leftward component, respectively). Grey region: s.e.m.

**Figure 8. Amount of encoded information correlates with behavioral performance.** Pearson correlation between fluctuations in encoded information ($D_{PE}$) and fluctuations in monkeys’ performance (see Methods). Across all datasets and ensembles sizes, trials associated with larger encoded information are also significantly associated with better task performance. Error bars correspond to the 25th -75th percentile of the distribution of bootstrap medians and significant deviations from zero (high contrast colored bars) are calculated by a Wilcoxon signed rank test (not significant if $P > 0.05$).

**Figure 9. Population signal and projected precision best predict behavioral performance.** Percentage change in behavioral performance, as a function of ensemble size, is shown for each behavioral task when the conditioned bootstrapping approach is used to isolate fluctuations in PS (red), PP (green), MPC (blue), and GA (orange). See also Extended Data Figs. 9-1, 9-2 and 9-3.
Figure 10. An experimentally-constrained neural population model accounts for the empirical findings. (A) Generative model for the simulated neuronal responses. A population of $N$ model neurons was characterized by linear tuning curves with slope $m_1$. An intermediate activity pattern ($r'_j$) was obtained by drawing an $N$-dimensional sample from a multivariate Gaussian distribution (mean $\mu$ and covariance $\Sigma^{\text{gen}}$) and then corrupting it with sensory noise ($\delta s_j$) on every trial $j$ ($r_j$) ($M$ trials in total). A homogeneous response gain modulation ($g_j$) and a Poisson step were applied to produce the final population spike count ($n_j$). The choice of the virtual agent ($c_j$) was obtained by an optimal read-out of the population activity pattern on each trial (Methods). (B) Percentage change in the amount of information encoded by the model (DP$_{cv}$), as a function of the ensemble size, when bootstrap fluctuations of different features of the neural activity are isolated. Only isolated bootstrap fluctuations in PS and PP influence the amount of information encoded by the population, consistent with the experimental observations (Fig. 6). (C) Percentage change in behavioral performance predicted by fluctuations in different statistical features of model population activity. PS and PP are the most important factors influencing behavioral performance. The magnitude of changes in behavioral performance is similar to that observed experimentally (Fig. 9). See also Extended Data Fig. 10-1.

Figure 1-1. The expression for the optimal classifier is the best approximation to the amount of encoded information. (A) Ratio between percentage of explained variance (E.V.) of the optimal classifier and the correlation-blind classifier (top panel) and between the optimal and the variability-blind classifier (bottom panel; see Methods). The percentage of explained variance determines how good a given analytical expression of DP$_{th}$ approximates the performance of a cross-validated linear classifier DP$_{cv}$ (see Methods). For all monkeys (mean across all ensemble sizes: 2, 4, 6, 8 and 10 units) the theoretical expression for the optimal classifier (Eq. 1) is the best approximation for the amount of information encoded by
the neural population as expressed by $\text{DP}_{cv}$. (B) Ratio between percentage of explained variance (E.V.) of the correlation-blind classifier and the optimal classifier (top panel) and between the correlation-blind and the variability-blind classifier (bottom panel) after shuffling the activity of each neuron across trials for a fixed condition. The amount of encoded information when pairwise correlations are removed is better approximated by the correlation-blind classifier than by the optimal classifier (before shuffling) and the variability-blind classifier for all monkeys.

**Figure 1-2.** The theoretical expression provides excellent fits for different linear classifiers and goodness-of-fit metrics. (A) The slope and intercept parameters of the linear fit (Type II regression, see Methods) between the theoretical and the cross-validated decoding performance ($\text{DP}_a$ and $\text{DP}_{cv}$) are relatively close to 1.0 and 0.0, respectively, for a Linear Discriminant Analysis (LDA) on all ensemble sizes and monkeys. Deviations from 1.0 (slope) and 0.0 (intercept) can be explained by the limited number of trials used to train and test the LDA, which produces values of DP below 0.5 in some cases. (B) All goodness-of-fit metrics (% explained variance, slope and intercept) are improved when the training set is used to test the performance of LDA. (C) When using Logistic Regression (LR) instead of LDA, the fitting results are similar to panel (A). (D) As in panel (B), when using the training set to evaluate the performance of LR, all goodness-of-fit metrics are improved with respect to panel (C).

**Figure 1-3.** Linear discriminant analysis outperforms logistic regression and quadratic discriminant analysis. (A) Median decoding performance (DP) for linear discriminant analysis (LDA), logistic regression (LR), and quadratic discriminant analysis (QDA) (5-fold cross-validation) for an ensemble size of 2 units from monkey 1. LDA produces a larger DP than both LR ($P = 6.2 \times 10^{-13}$, Wilcoxon signed rank test) and QDA ($P = 2.4 \times 10^{-3}$). Note that significance of the differences is strong due to pairing of the data despite the relatively large error bars displayed. (B, C) Equivalent to (A) for monkeys 2 and 3 and for a range of ensemble sizes (2, 4, 6, 8 and 10 units). As in (A), LDA produces the largest mean DP of the three decoders. (D) Analogous results for monkey 4. In all panels error bars denote 25th -75th percentile of the distribution of bootstrap medians and * indicates significance in the reported differences ($P < 0.05$).
Figure 6-1. Percentage change on the amount of information encoded by the neuronal population (% change DPcv) for all monkeys and ensemble sizes. Population signal and projected precision are the most influential factors on DPcv. * = 0.01 < P < 0.05, ** = 0.001 < P < 0.01, *** = P < 0.001.

Figure 6-2. Difference of percentage change on the amount of information encoded by the neuronal population (% change DPcv) for all monkeys and ensemble sizes for all the pairs of studied features. Population signal and projected precision are the most influential factors on DPcv. * = 0.01 < P < 0.05, ** = 0.001 < P < 0.01, *** = P < 0.001.

Figure 9-1. Results shown in Fig. 9 are robust to different conditioning parameters and metrics. (A) Equivalent to Fig. 9, but averaged across ensemble sizes. Percentage change in behavior (see Methods) is plotted for bootstrap fluctuations of each statistical feature (conditioned bootstrapping method; see Methods), and results are shown for different conditioning criteria: ± 10% (left column) and ± 20% (right column) from the median value of the bootstrap distribution. (B) Same as (A) but using Pearson correlation between bootstrap fluctuations of the different features of neural activity and behavioral performance instead of percentage change. Data are shown for the same two conditioning criteria and averaged across ensemble sizes. Error bars correspond to 25th - 75th percentile of the distribution of bootstrap medians and significant deviations from zero (high contrast colored bars) are calculated by a Wilcoxon signed rank test (not significant if P > 0.05).

Figure 9-2. Percentage change on behavioral performance (% change Behavior) for all monkeys and ensemble sizes. Population signal and projected precision are the most influential factors on behavior. * = 0.01 < P < 0.05, ** = 0.001 < P < 0.01, *** = P < 0.001.

Figure 9-3. Difference of percentage change on behavioral performance (% change Behavior) for all monkeys and ensemble sizes for all the pairs of studied features. Population
signal and projected precision are the most influential factors on behavior $* = 0.01 < P < 0.05$, $** = 0.001 < P < 0.01$, $*** = P < 0.001$.

**Figure 10-1.** An experimentally-constrained population model replicates experimental results. (A) Graphical model depicting, in full detail, the generative process underlying the surrogate datasets (see Methods). (B) Tuning curve of an example neuron used in the model. This neuron’s firing rate signals the direction of motion with a linear tuning curve that is corrupted by noise. (C) Distribution of Fano Factors for all units in an example surrogate ‘session’ of the model. The mean value of the distribution is in rough agreement with empirical values of Fano factors found in *in vivo* recordings. (D) Distribution of pairwise correlations of the spike counts of all neuronal pairs in an example surrogate ‘session’ of the model. The mean value of the distribution is compatible with the empirical values of mean pairwise correlations found in *in vivo* recordings. (E) The theoretical decoding performance of a linear classifier ($D_{\text{th}}$) was a very good approximation of the cross-validated decoding performance ($D_{\text{cv}}$) on the model data (see Methods and Fig. 1). (F) Amount of encoded information ($D_{\text{cv}}$) as a function of the network size for the three different stimulus intensities in the model. When differential correlations are not included in the model (light colored lines), information grows without limit for all of the stimulus values ($D_{\text{cv}} = 1.0$ upper limit for information in a classification task). When differential correlations are included in the model (darker colored lines), $D_{\text{cv}}$ is lower and reaches an asymptote (dashed colored lines) for large network sizes. (G) Pearson correlation between bootstrap fluctuations in the amount of encoded information ($D_{\text{cv}}$) and behavioral performance of the virtual agent. The model accounts for the weak coupling between these two quantities (see Fig. 10). Error bars correspond to 25th - 75th percentile of the distribution of bootstrap medians.

**Author Contributions**

R.N., J.M.T., G.C.D. and R.M.B. conceived the project. R.N. and R.M.B. derived the analytical calculations. R.N. performed the analysis. N.E.P., A.A., and J.M.T. obtained the recordings (monkeys 2, 3 and 4). All authors wrote the manuscript.
\[ PS = |\Delta f| \]
\[ PP = \sqrt{\sum_{i=1}^{N} \frac{\cos^2 \hat{\theta}_i}{\sigma_i^2}} \]

\[ PS \rightarrow PS \]
\[ PP \rightarrow \frac{PP_0}{\sqrt{1 + \tilde{\epsilon} PP_0^2}} \]
\[ \tilde{\epsilon} = |P'|^2 \tilde{\epsilon} \]