TECHNICAL REPORTS

nature neuroscience

Structure in neural population recordings: an expected byproduct of simpler phenomena?

Gamaleldin F Elsayed[1–3] & John P Cunningham[1,3,4]

Neuroscientists increasingly analyze the joint activity of multineuron recordings to identify population-level structures believed to be significant and scientifically novel. Claims of significant population structure support hypotheses in many brain areas. However, these claims require first investigating the possibility that the population structure in question is an expected byproduct of simpler features known to exist in data. Classically, this critical examination can be either intuited or addressed with conventional controls. However, these approaches fail when considering population data, raising concerns about the scientific merit of population-level studies. Here we develop a framework to test the novelty of population-level findings against simpler features such as correlations across times, neurons and conditions. We apply this framework to test two recent population findings in prefrontal and motor cortices, providing essential context to those studies. More broadly, the methodologies we introduce provide a general neural population control for many population-level hypotheses.

A fundamental challenge of neuroscience is to understand how interconnected populations of neurons give rise to the remarkable computational abilities of our brains. To answer this challenge, advances in recording technologies have produced datasets containing the activity of large neural populations. Population-level analysis techniques have similarly proliferated[1–3] to draw scientific insight from this class of data, and as a result, researchers now generate and study hypotheses about structure in neural population activity. These 'population structures' describe scientifically interesting findings at the population level that elucidate properties or features of neural activity that, ostensibly, can neither be studied with traditional single-neuron analyses nor be predicted from existing knowledge about single-neuron responses. Claims of significant population structures support results in many brain areas including the retina[4], the olfactory system[5,6], frontal cortex[7,8], motor cortex[9,10], parietal cortex[11–13] and more[1,3].

While promising, these advances are also perilous. Population datasets are remarkably complex, and the population structures found in these data are often the result of novel data analysis methods with unclear behavior or guarantees. Specifically, many analysis techniques do not consider the very real concern that the observed population structure may be an expected byproduct of some simpler, already-known feature of single-neuron responses. **Figure 1** shows four examples of population structure from the literature, to demonstrate how this concern may arise. In rodent posterior parietal cortex (**Fig. 1a**), Raposo and colleagues[11] recorded single neurons tuned[14,15] to multiple task parameters (often called mixed selectivity[16]): neural responses in a decision-making task modulated to both the choice and stimulus modality (auditory or visual). They used a machine-learning algorithm to find individual readouts of the population that represented choice only and modality only (plotted against each other in **Fig. 1a**). However, one might ask, is this population structure truly a novel finding, or should we expect to find such readouts given our knowledge that single neurons are tuned to choice and modality? In primate prefrontal cortex (PFC), Murray and colleagues[17] analyzed a neural population during a working-memory task and found a readout that is more stable in time than the single-neuron responses themselves (**Fig. 1b**). Again we may ask, is this stability significant, or is it expected as a byproduct of the temporal smoothness (or, correlations) of single-neuron responses? Population-level neural dynamics have also been studied: low-dimensional projections of neural population responses seemingly evolve over time depending on their response history and initial conditions; examples include the locust antennal lobe[18] (**Fig. 1c**) and primate motor cortex[9] (**Fig. 1d**). Are these population findings novel signatures of dynamical systems[19,20], or is this structure an expected byproduct of the temporal, neural and condition correlations of the neural data? This and the previous concerns of course depend on the subjective assumption that these simpler features are known a priori (i.e., not a consequence of the population structure, to which some researchers give primacy). Nonetheless, in the face of these concerns and a spate of prominent population-level results, the neuroscience community has begun to raise significant doubts about the extent to which population-level findings are an expected byproduct of simpler phenomena. This debate will remain unresolved in the absence of rigorous methodology for evaluating the novelty of population findings.

To address this challenge, we developed a methodological framework—the 'neural population control'—to test whether or not a given population structure is an expected byproduct of a set of primary features: the tuning of single neurons[14,15], temporal correlations of firing rates (regardless of whether one views that temporal correlation as fundamental or a result of smoothing[21–23]) and signal correlations across neurons[24,25] (also called the low dimensionality of neural

[1]Center for Theoretical Neuroscience, Columbia University, New York, New York, USA. [2]Department of Neuroscience, Columbia University Medical Center, New York, New York, USA. [3]Grossman Center for the Statistics of Mind, Columbia University, New York, New York, USA. [4]Department of Statistics, Columbia University, New York, New York, USA. Correspondence should be addressed to J.P.C. (jpc2181@columbia.edu).

Received 3 February; accepted 30 June; published online 7 August 2017; doi:10.1038/nn.4617

populations[10,26]). The central element of this neural population control is a set of algorithms that generate surrogate datasets that share the specified set of primary features with the original neural data but are otherwise random. Accordingly, these surrogates will express any population structure to the extent expected by the specified primary features (since there is by definition no additional structure). We extended Fisher randomization and maximum entropy modeling to generate these surrogate datasets, and we chose the primary features to be the mean and covariance of the data across times, neurons and experimental conditions. This choice is justified: the use of first[15] and second[24,27] moments is standard in neuroscience, and further, these are the lowest-order moments that can produce responses with qualitative similarity to real data in terms of temporal smoothness, low dimensionality and tuning to conditions. These surrogate datasets formed the basis for a statistical test, giving a precise probability (a P value) that a population structure is an expected byproduct of the specified primary features. Critically, careful inspection of this problem also revealed the inadequacy of typical statistical controls and validation techniques, and our results showed the extent to which ignoring such primary features can misstate statistical confidence, perhaps drastically, in a population-level result.

The neural population control can be applied to population structures and to datasets from almost any brain area. To show its utility, we used it to test two recent influential results. First, using data from macaque PFC engaged in a working-memory task[28,29], we found that the presence of strong stimulus-specific population readouts is expected from the robust tuning of single neurons. In contrast, we found that the decision-specific readouts are not expected. Second, using multielectrode array recordings from the macaque motor cortex[9], we found that population-level dynamical structure[30] is not an expected byproduct. The results of the neural population control framework contextualize and clarify these studies, quantitatively resolving skepticism and pointing to how this framework can be used throughout systems neuroscience.

RESULTS

Motivation for the neural population control

Consideration of conventional controls clarifies the need for the neural population control. Traditionally, one begins with a choice of a summary statistic, a number that quantifies the structure in question. In population studies, some common choices are variance explained[11] or a goodness-of-fit metric such as the coefficient of determination[9,13]. This statistic is calculated for the data and then compared to a null distribution, producing a P value, which gives the likelihood of that statistic value (or greater) arising by chance under the null hypothesis. Critically, one requires a null distribution, and the most common approach is to shuffle the neural data so as to disrupt any special coordination that might have given rise to the population structure in question. Then, the summary statistic is calculated for each shuffled surrogate dataset, and the null distribution is built from the calculated statistics of many surrogate datasets. Should the summary statistic of the original data be likely under the null distribution, then, the argument goes, population structure was not surprisingly different than expected by chance. In principle, this procedure is appropriate only if the surrogate datasets conserve all the primary features of the original neural data, such that the surrogates remain a plausible comparison. However, often this essential requirement is not met, in which case one has a major problem of interpretation: is the difference in the summary statistic between the original and surrogate datasets due to disruption of the population structure itself, due to distortion of the primary features, or both?

Failure to account for known features presents a significant challenge and can lead to misinterpreting results. To elucidate this pitfall and highlight what the neural population control offers, **Figure 2** presents two illustrative examples. **Figure 2a** shows responses from two simulated neurons, each tuned to eight stimuli. Suppose a population-level analysis found a readout (**Fig. 2a**) in which the data was well tuned to the stimulus and that this subspace accounted for a great deal of the population signal (99% of data variance captured, here). Of course (by construction) this finding is an expected byproduct of the fact that both neurons are well tuned to this stimulus. However, a standard shuffle (**Fig. 2b**) will corrupt tuning and suggest that population structure is in fact significant (variance of that tuned readout has dropped to 61%). Indeed, repeated shuffles produce a null distribution (**Fig. 2c**) erroneously implying significance (with $P < 0.001$). The neural population control, using algorithms we will shortly introduce, produces surrogates that maintain neural tuning and other primary features, leading to the correct conclusion both qualitatively (**Fig. 2b**) and quantitatively (**Fig. 2c**): here the variance explained by this subspace is an expected byproduct of tuning, not a novel population-level result.

Even when a population-level result is not an expected byproduct of simpler features, conventional controls can meaningfully misstate confidence. **Figure 2d** shows two simulated neurons coupled as an oscillator, where eight stimuli set initial states of the given differential equations. Population-level neural responses thus evolve in time according to a dynamical flow field (**Fig. 2d**). Under standard shuffling (**Fig. 2e**), correlations across neurons and conditions change, and the consistency with the dynamical flow field is considerably reduced, both qualitatively (**Fig. 2e**) and quantitatively (using the coefficient of determination R^2; **Fig. 2f**). The neural population control produces surrogate data with the appropriate primary features, producing the correct null distribution and confidence level (**Fig. 2f**). Thus, the essential remaining challenge for rigorously testing the novelty of population-level results is to develop methods for producing random surrogate datasets (**Fig. 2b,e**) that match the primary features of the original data.

Corrected Fisher randomization and tensor maximum entropy

We need to generate surrogate datasets that share the primary features of the original neural data but are otherwise random. We developed two complementary methods that achieve that goal, termed 'corrected Fisher randomization' (CFR) and 'tensor maximum entropy' (TME). CFR adds an optimization step to traditional shuffling, to maintain the primary features, whereas TME samples random datasets from a probability distribution with the correct average primary features (Online Methods). The high-level mechanics of these methods are illustrated schematically in **Supplementary Figure 1**. As in traditional shuffling, the first step of CFR is to randomly shuffle the responses of each neuron across experimental conditions. Because this standard shuffling step destroys the primary features of the original data, we then construct and apply an optimized neural readout, a matrix that reweighs the shuffled neural responses, to minimize the error between the primary features of the new shuffled responses and the primary features of the original neural data. The strength of CFR is that each surrogate dataset preserves the primary features of the original data (up to the optimization error, which is empirically quite minor; **Supplementary Figs. 2** and **3**). However, as with most shuffling techniques, CFR is conservative as it operates on a finite dataset (i.e., it shuffles the finite set of recorded neural responses). Hence, some structures that are not stipulated by the null hypothesis may persist in shuffled data (e.g., if a neural trace is nonsmooth at one time point, the trace after shuffling will still be nonsmooth at this time point). Owing to this potential shortcoming, we also extended the maximum entropy principle (which has been widely used in neuroscience[27,31–33]) to develop the complementary TME method.

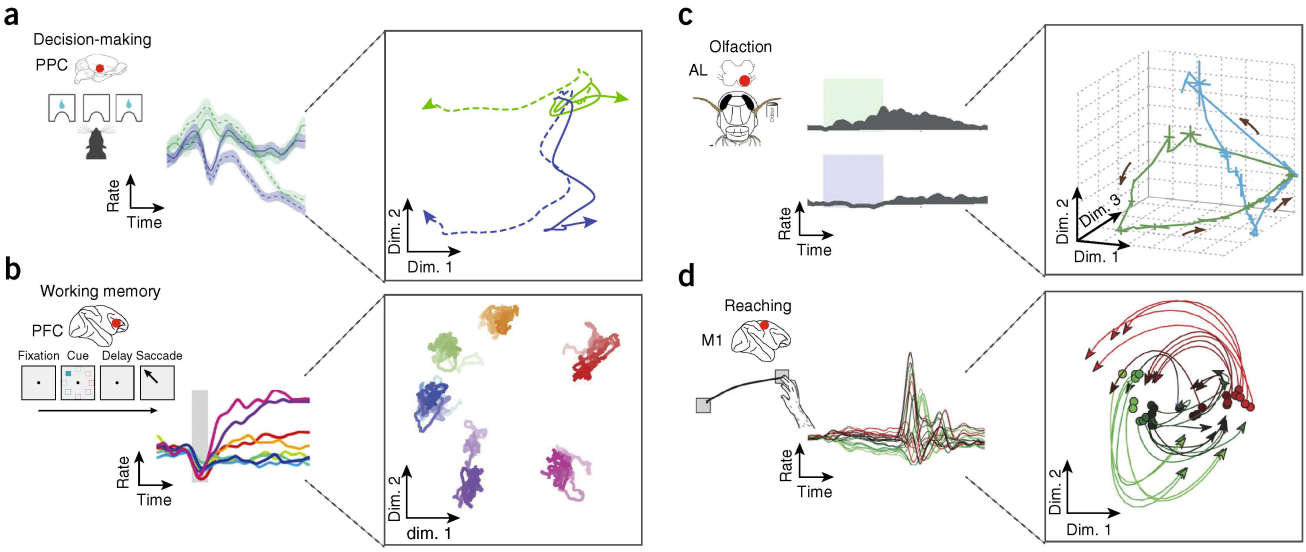

Figure 1 Population structure in systems neuroscience: examples from studies investigating structure at the level of the population. (**a**) Left: an example firing rate response from a rat posterior parietal cortex (PPC) neuron during a multimodality decision-making task (adapted from Raposo *et al.*[11], Nature Publishing Group). The single-neuron responses show mixed selectivities to cue modality (blue, visual cue; green, auditory cue) and decision (dashed lines, right lick port; solid lines, left lick port). Right: a two-dimensional projection of the population response, where choice information is separated along dimension 1 (Dim.1; horizontal) from the modality information, which is separated along dimension 2 (Dim. 2; vertical). (**b**) Left: an example firing rate response from a primate PFC neuron during a working-memory task (adapted from Murray *et al.*[17], National Academy of Sciences). The single-neuron responses to the six stimuli (illustrated by different colors) show temporal dynamics. Right: a two-dimensional projection of the population, in which stimulus information is stably represented across time. (**c**) Left: an example firing rate response from a locust antennal lobe (AL) projection neuron responding to two odors (adapted from Broome *et al.*[18], Elsevier). Right: a three-dimensional projection of the population data with neural trajectories corresponding to the two odor stimuli. (**d**) Left: an example firing rate response from a primate motor cortex (M1) neuron during a delayed-reach task (adapted from Churchland *et al.*[9], Nature Publishing Group). Right: a two-dimensional projection of the population data, with neural trajectories corresponding to each reaching condition.

We derived a probability distribution defined over random tensors (datasets) that maximizes Shannon entropy, subject to the constraints that the expected primary features of the distribution are those of the original neural data (Online Methods). This distribution is a non-trivial and (to our knowledge) novel extension of classic maximum entropy distributions, both in terms of extending to tensor random variables and in terms of the computational techniques required to sample from this distribution. The primary strength of the maximum entropy principle is that higher-order structures in surrogate datasets are completely determined by the primary features (the distribution is by definition maximally unstructured beyond those primary feature constraints). On the other hand, since the constraints are enforced in expectation, variations in the primary features of each surrogate dataset will appear due to finite sampling. Thus, CFR and TME offer complementary and well-balanced techniques for generating surrogate datasets properly according to the null hypothesis.

To demonstrate the framework, we used CFR and TME to generate surrogate datasets based on neural responses recorded from primate PFC during a working memory task (**Fig. 3a**). **Figure 3b** shows the firing rates of one example neuron from the original neural data along with its primary covariance features. To illustrate the ability of CFR and TME to preserve the primary features, we generated three types of surrogate datasets (Online Methods). The first we term surrogate-T, which preserves only the primary features across time similar to the conventional shuffle control from **Figure 2**. The second, surrogate-TN, preserves the primary features across both times and neurons. The third, surrogate-TNC, simultaneously preserves the primary features across times, neurons and conditions. Qualitatively, single-neuron responses from the surrogate datasets appear realistic (**Fig. 3c–e** and **Supplementary Figs. 4** and **5**). Quantitatively, the estimated covariances across times from all surrogate types were similar to the covariance across times of the original neural data (Σ_T; **Fig. 3b–e** and **Supplementary Figs. 2** and **3**). Additionally, the estimated covariances across neurons from surrogate-TN and surrogate-TNC were similar to the covariance across neurons of the original data (Σ_N; **Fig. 3b–e** and **Supplementary Figs. 2** and **3**). Finally, the estimated covariances across conditions from surrogate-TNC were also similar to the covariance across conditions of the original data (Σ_C; **Fig. 3b–e** and **Supplementary Figs. 2** and **3**). Thus, both CFR and TME successfully generated random surrogate data that preserved the specified primary features. These surrogate datasets are then appropriate for generating suitable null distributions for a statistical test of population structures.

In all that follows, we consider surrogate-TNC as the basis for the neural population control, as it addresses the full null hypothesis that temporal, neural and condition means and covariances give rise to the population structures in question. That said, the inclusion of surrogate-T and surrogate-TN here remains important: first, surrogate-T connects to conventional shuffling and will demonstrate the inadequacies of that standard method; second, surrogate-TN demonstrates an alternative null hypothesis that is appropriate in other settings[34], and it allows us to analyze empirically the benefits of adding each of the primary features to the null hypothesis. It is worth noting that other surrogate types, such as surrogate-NC, are easily generated (our software implementation accepts this choice as an input), but they appear visually implausible due to the standard of plotting responses over time.

Population representations and mixed selectivity in PFC

Previous studies have demonstrated that neurons in a number of brain areas respond to multiple task parameters[7,11,16,29,35,36]. These

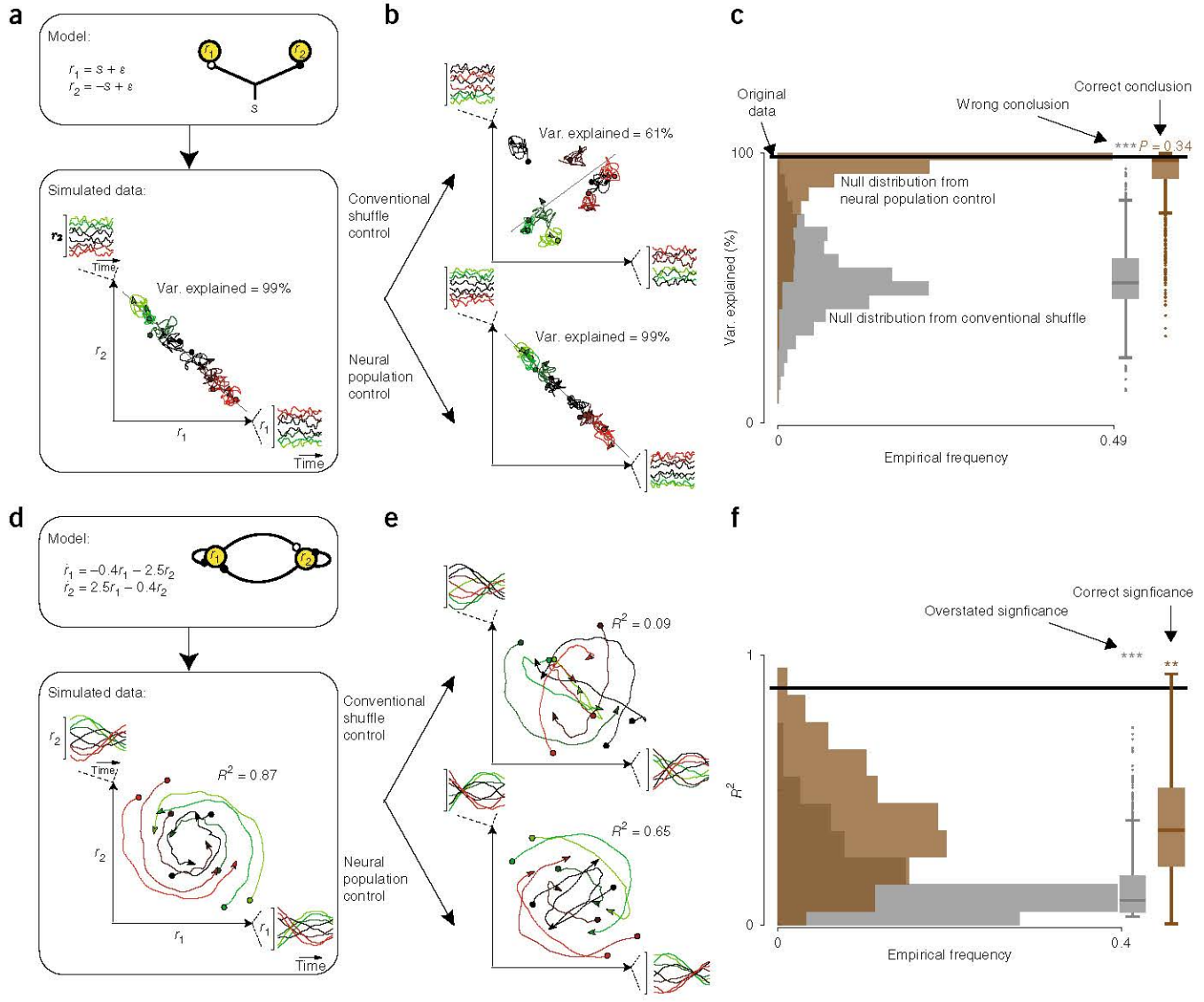

Figure 2 Motivation for the neural population control. (**a**) Simulated firing rates (*r*) from two neurons encoding a hypothetical stimulus at eight conditions (high, moderate and low stimuli correspond to red, black and green traces, respectively), along with the corresponding neural trajectories in the population space (here two-dimensional). Black diagonal line illustrates a one-dimensional projection of the data that represents stimulus, identified by the target dimensionality reduction method (from Mante *et al.*[8]). The data variance (var.) explained by this projection is shown (as a percentage). (**b**) Top: shuffled surrogate data generated by shuffling the single neuron responses from **a** across conditions. The same data analysis method was then used to identify a one-dimensional projection of the data (black line) that represents the stimulus in the shuffled data. Bottom: random surrogate data from the neural population control (Online Methods) and the identified projection that represents the stimulus (black line). The data variance (var.) explained by this projection is shown as a percentage, as in **a**. (**c**) Distribution of variance-explained values from stimulus projections identified from 1,000 surrogate datasets (gray) and another distribution of variance values from 1,000 surrogate datasets from the neural population control (brown). Black line is the percentage of variance explained from the neural data from **a**. Box-and-whisker plots summarize the two distributions (Tukey conventions; box lower border, middle line and upper border show 25th percentile, median, and 75th percentile, respectively, and whiskers show lowest and highest points within 1.5× the interquartile range). (**d**) Firing rates for two neurons are solutions (*r*) to the given differential equations (modeling an oscillator), with eight different initial conditions. The fit of these data to a linear system (R^2) is shown. (**e**) Shuffled data (top) and surrogate data from the neural population control. (**f**) Distributions of R^2 values from 1,000 shuffled datasets and 1,000 surrogate datasets from the neural population control (conventions as in **c**). Smoothed Gaussian noise was added to all simulated data. In **c**, **f** and all subsequent figures, $*P < 0.05$, $**P < 0.01$ and $***P < 0.001$.

mixed responses may obscure representation at the level of single neurons. Dimensionality reduction methods[8,37] are widely used to identify neural readouts (projections of the population) that separate the representation of each task parameter. Further, these readouts often are found to explain substantial data variance. Should we always expect such a finding from any collection of neurons that have these mixed responses? Not necessarily: one can produce toy examples in which the representations fundamentally cannot be separated and other examples in which separation would be possible but only with small variance explained (i.e., in the noise). The suggestion then typically follows that these robust readouts are thus evidence of a collective code in the population: neural responses are coordinated in such a way to produce these readouts, though that coordination is invisible at the level of single neurons. However, this line of reasoning misses the critical concern that these task-parameter-specific readouts may be an expected byproduct of simpler features in the data itself: tuning,

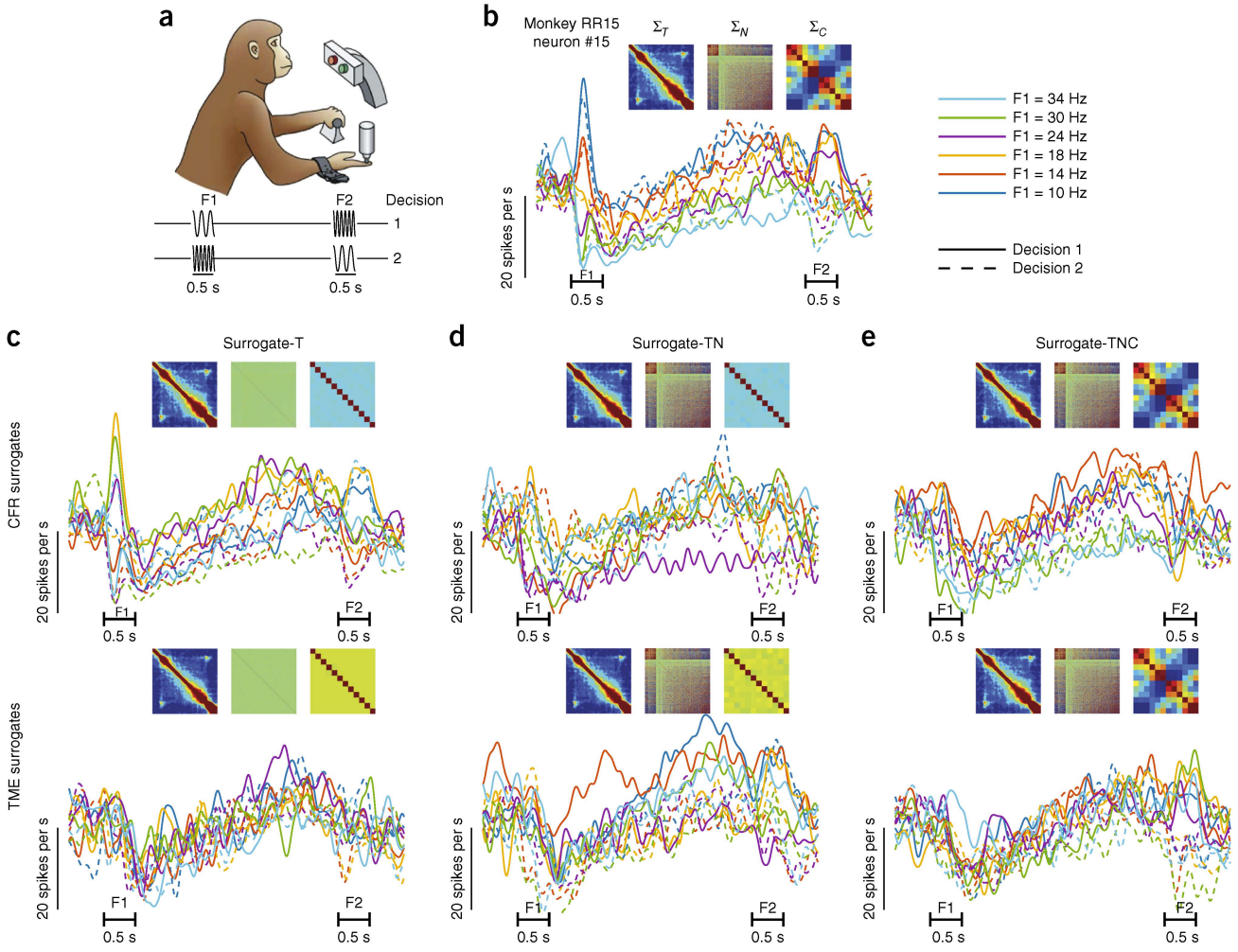

Figure 3 CFR and TME surrogate datasets preserve the specified primary features. (**a**) Working-memory task (adapted from Romo and Salinas[38], Nature Publishing Group). (**b**) Example neuron (neuron number 15 of 571 total) from PFC. Each trace is the trial-averaged firing of the 12 task conditions (six stimuli and two decisions; one trace color and style for each). Horizontal bars denote the times of first (F1) and second (F2) vibrotactile stimuli. Heatmaps in the inset show three covariance matrices across times (Σ_T), neurons (Σ_N) and conditions (Σ_C) of all neurons in this dataset. (**c**–**e**) Example neurons from one (**c**) surrogate-T, (**d**) surrogate-TN and (**e**) surrogate-TNC dataset, respectively; conventions as in **b**. Top panels are surrogate datasets generated using CFR and bottom panels are surrogate datasets generated using TME. Covariance matrices in the insets are obtained by averaging the primary features from 100 surrogate datasets.

temporal smoothness and neural correlations. Asserting this claim of collective code requires the neural population control.

We tested recordings from PFC during a working-memory task[38] (**Fig. 3a**). In this task, subjects (two rhesus macaques) received two vibrotactile stimuli with different frequencies. A delay period separated the presentation of the two stimuli, during which the monkeys were required to maintain a memory of the first stimulus. After the delay period, subjects reported whether the frequency of the first stimulus was higher or lower than the frequency of the second stimulus. Thus, the two relevant task parameters encoded by PFC were the decision and the first stimulus frequency.

Responses in PFC showed mixed selectivity to the two task parameters (**Fig. 3b**). We used demixed principal component analysis[37] to identify decision-specific and stimulus-specific population readouts in both the original neural dataset and our surrogate datasets. The projection of the population activity onto the decision (stimulus) readout reflects the population representation of the decision (stimulus). We then compared these projections to those found in surrogate datasets generated by CFR and TME. Qualitatively, the projections, from both the original and the surrogate-TNC datasets, appeared to be tuned to the decision and the stimulus (**Fig. 4a–d**). Quantitatively, we calculated the percentage variance explained by the decision and stimulus projections, which summarized the degree to which each projection accounts for the population response.

Figure 4e demonstrates that the variance captured by the decision projection from the original neural data was significantly higher than the variance captured by the decision projections from the surrogate datasets ($P = 0.015$ for RR15 TME surrogate-TNC; $P < 0.001$ in all other subjects and tests). This finding demonstrates that the population representation of the decision was not an expected byproduct of the primary features. However, the same procedure for the stimulus projection demonstrates that the population representation of the stimulus could not be distinguished from an expected byproduct of the primary features, as surrogate-TNC data generated with only those primary features displayed comparable population structure (**Fig. 4f**). The variance captured by the stimulus projection from the original neural data reached significance only when compared to the variance captured by the stimulus projections from surrogate-T and

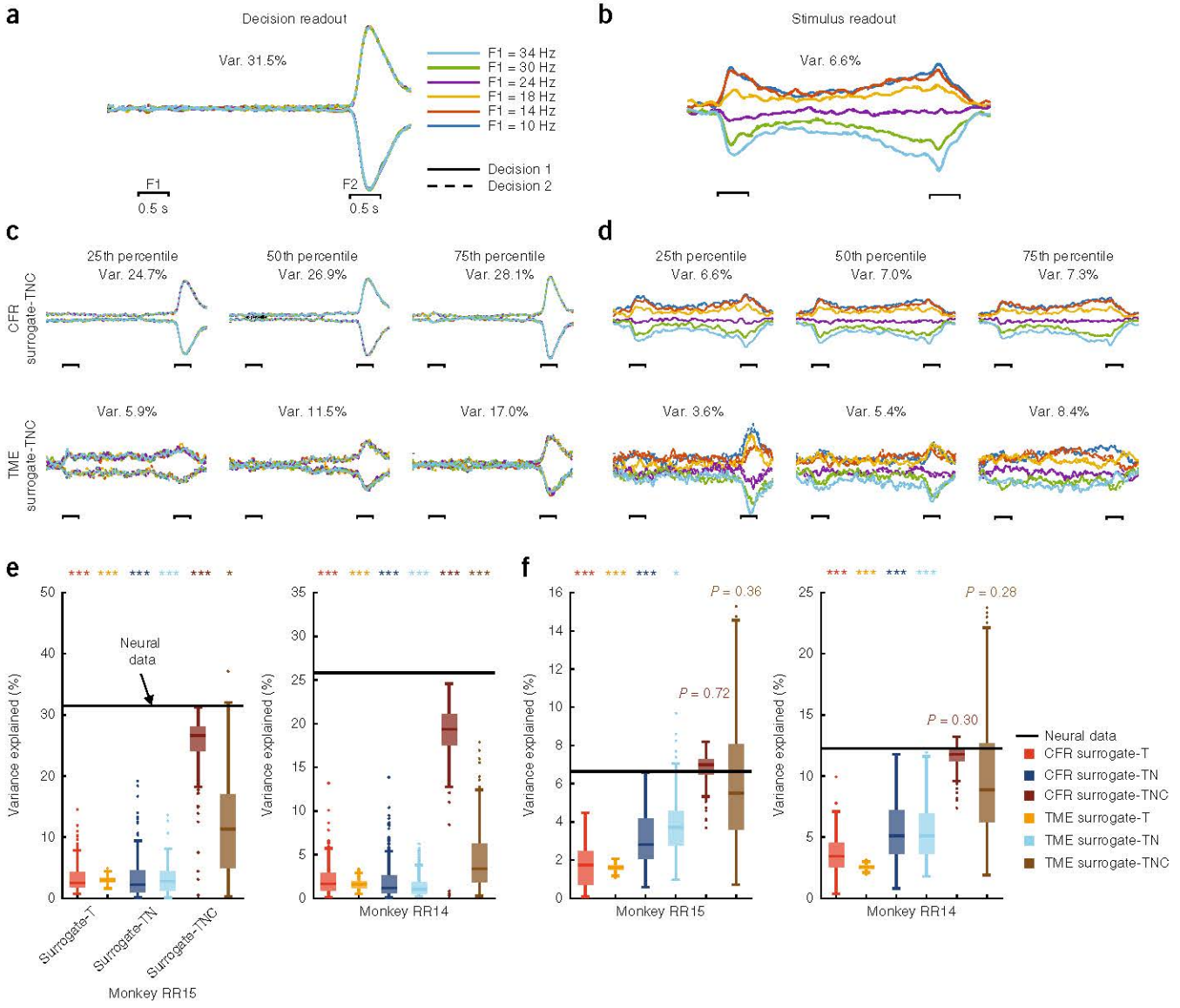

Figure 4 Decision (or stimulus) readouts in PFC are not (or are) an expected byproduct. Population readouts for decision and stimulus identified using demixed principal component analysis. (**a**) Projections of the original population responses from monkey RR15 onto the top decision-specific readout. (**b**) Projections of neural responses from monkey RR15 onto the top stimulus-specific readout. Trace colors and style follow the same conventions as in **Figure 3b**. (**c**) As in **a** but for decision readouts from surrogate datasets generated by CFE (top) and TME (bottom). We show surrogates at various points in the distribution of variance explained (25th, 50th and 75th percentiles; 200 surrogate datasets). (**d**) As in **b** but for the surrogate datasets from CFE (top) and TME (bottom) methods; as in **c** the 25th, 50th and 75th percentile examples are shown (200 surrogate datasets). Scale bars and color scheme in **b**–**d** as in **a**. (**e**) Percent variance-explained of the population projection onto the top decision readout. Black lines show the percent variance explained from the original neural data; colored box-and-whisker plots show the variance explained distribution from 200 surrogate samples (significance levels denoted by asterisks; conventions as in **Fig. 2c,f**; upper-tail test). The variance of the decision projection is calculated during the decision epoch (from 100 ms after the second stimulus onset until the second stimulus offset). (**f**) As in **e** but for percent variance explained of the population projection onto the top stimulus readout. The variance of the stimulus projection is calculated during the stimulus epoch (from 100 ms after first stimulus onset until the second stimulus onset).

surrogate-TN datasets but not when compared to surrogate-TNC datasets. This result was similar when we repeated the same analysis using another statistic (the explained variance metric used in Kobak *et al.*[37]; **Supplementary Fig. 6**). This negative result contextualizes our understanding of population-level representations: sometimes, despite qualitative appearances of a collective population code, such a readout can exist simply because of a powerful algorithm and simpler known features in the data. Note that this result does not mean that the stimulus readout is absent or wrong in any way but rather that it is expected, given the primary features of the data.

One advantage of this framework is that the contribution of each primary feature to the population structure can be quantified by studying the null distributions across different surrogate types. This inspection indicates that tuning across conditions was probably the feature giving rise to the stimulus readout. Although single neurons in PFC showed mixed responses to the stimulus and decision, the tuning of the stimulus was prominent (**Supplementary Fig. 7**). Due to the task structure, neurons in PFC responded to the first stimulus at all times of the task, except during the brief period starting at the second stimulus onset (**Fig. 3a,b**). Thus, the population representation

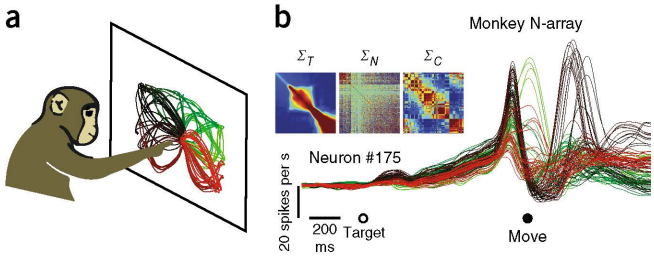

Figure 5 Motor cortex responses during a delayed-reach task. (**a**) Delayed-reach task. Monkeys performed straight and curved reaches to targets displayed on a frontoparallel screen. Trajectories represent the average hand position during each of 108 reaching conditions. (**b**) Example neuron (neuron number 175 of 218 total) recorded from the motor cortex of one monkey during the delayed-reach task. Each trace is the smoothed, trial-averaged firing rate during one of the reaching conditions. The trace color indicates the reach condition from **a**. Heatmaps in the inset represent three covariance matrices that quantify the primary features across time (Σ_T), neurons (Σ_N) and conditions (Σ_C) of the entire population dataset.

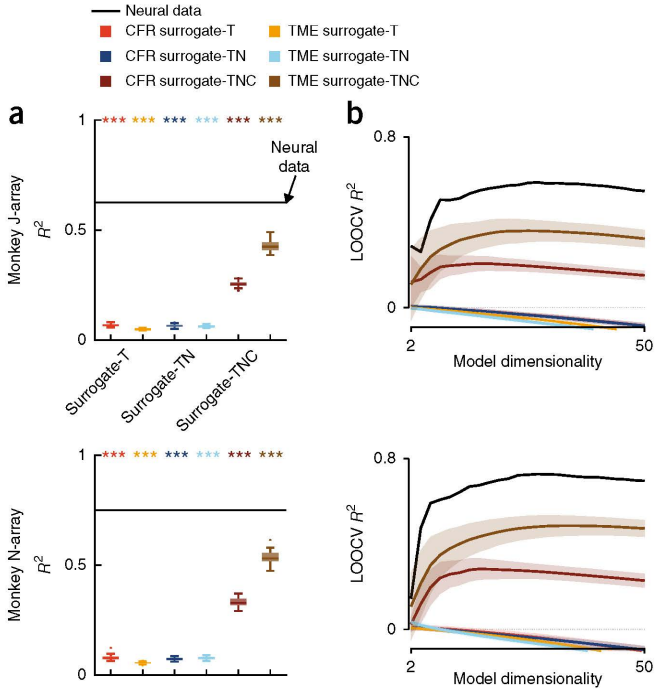

Figure 6 Population dynamics in motor cortex are not an expected byproduct. We projected 400 ms of movement-related neural activity in the motor cortex on the top principal components (PCs) and then fitted them to a linear dynamical system. (**a**) Quality of fit (R^2) of the original neural responses projected onto the top 28 PCs (determined by cross validation; **Supplementary Fig. 8**) to the dynamical system model. Black lines denote the R^2 from the original neural data. Colored box-and-whisker plots denote the R^2 distributions from 100 surrogate datasets from each surrogate type (conventions as in **Fig. 2c**,**f**). Asterisks denote significantly higher R^2 than the surrogates (***$P < 0.001$; upper-tail test). (**b**) Leave-one-condition-out cross-validation (LOOCV) for the R^2 measure of fitting data to a dynamical system model with various choices of model dimensionalities (numbers of PCs). Black trace denotes the R^2 value from the original neural data and colored traces are the mean R^2 values from the surrogate datasets (color conventions as in **a**; shaded areas represent ±2 s.d.).

of the stimulus arose from the prominent tuning of single neurons, as expressed by the mean and covariance across conditions. Unlike the stimulus, the neural responses to the decision were briefer (only after the second stimulus onset) and overlapped with (and were dominated by) the neural responses to the stimulus (**Fig. 3a**,**b** and **Supplementary Fig. 7**). Hence, the population representation of the decision is not an expected byproduct of the underlying primary features and as such uncovers additional information about the representation of decision in PFC.

Population representations of task parameters are often useful for summarizing large datasets, and novel and rigorous methods like demixed principal component analysis are effective in finding those representations. The present result reminds us that care should be taken when interpreting the population representations found by these methods: in some cases, these representations may be an important indication of collective population codes hidden at the single-neuron level, while in other cases they may be simply a redescription of single-neuron tuning.

Primary features alone do not explain dynamical structure in motor cortex

To highlight the broad applicability of our framework, we next applied the neural population control to motor cortex responses during a delayed-reach task (**Fig. 5a**,**b**) to test the dynamical systems hypothesis[9,30]. Classical studies have assumed motor cortex activity represents movement kinematics[15]. Other studies have argued that the complexity of neural responses is beyond what is expected from coding models[39–41] and is more consistent with dynamical systems models[9,35]. In this view, motor cortex generates simple dynamical patterns of activity that are initialized by preparatory activity[39,42,43], and these patterns are then combined to produce complex muscle activity[39,44]. As in **Figure 2d**, a dynamical system implies a particular population structure, and recent studies have shown neural trajectories evolving (approximately) according to a low-dimensional linear dynamical system[9,35]. However, despite controls and comparisons with other hypotheses[9,39,45], the concern persists that this population structure may be an expected byproduct of simpler features in the data. That counterargument goes as follows: the temporal smoothness of neural responses in motor cortex data will give rise to temporally smooth neural trajectories, and correlated responses across neurons and conditions will give rise to low-dimensional neural trajectories that are also spatially smooth (in the sense that tuning implies that each neuron's response changes smoothly from one condition to the next, and thus population trajectories must also change smoothly in neural space from one condition to the next). Together, it is quite reasonable to suppose that this population structure (low-dimensional linear dynamical system fit) will arise as an expected byproduct of primary features of data.

We fit low-dimensional linear dynamical systems to population responses from multiple monkeys (dimensionality was chosen by cross-validation; **Supplementary Fig. 8**) and quantified the quality of fit by the coefficient of determination R^2 (Online Methods). We then generated a null distribution of R^2 values by fitting surrogate datasets generated by CFR and TME to the same dynamical model. Our results show that the R^2 from the original neural data was significantly higher than the R^2 from every surrogate type (**Fig. 6a**; $P < 0.001$). This result was consistent across different monkeys during different reaching tasks (**Supplementary Fig. 9**) and held similarly for oscillatory linear dynamics (**Supplementary Fig. 10**). Our neural population control demonstrates that the recently reported dynamical structure in motor cortical responses is not an expected byproduct of the specified primary features.

We can again use the different surrogate types (surrogate-T, -TN and -TNC) to quantify the contribution of each primary feature.

R^2 values from the surrogate-T and surrogate-TN datasets are similar, whereas the R^2 values from surrogate-TNC are much higher (**Fig. 6a**). The surrogate-TNC datasets are the only ones that preserve tuning (to experimental reach condition), from which we conclude that tuning contributes meaningfully to any appearance of dynamical structure in surrogate data (albeit substantially less than the original data). In other words, tuning inherently produces some degree of spatial smoothness that, if ignored, would have led to meaningful overstatement of the test significance of this dynamical structure.

To assess the sensitivity of this test result to model dimensionality, we performed leave-one-condition-out cross-validation and quantified the R^2 values of test conditions, both from the original neural data and the surrogate datasets, based on dynamical models with different dimensionalities. The same results hold: the R^2 value from the original neural data was still significantly higher than the R^2 values from all types of surrogate data, across a wide range of model dimensionalities (**Fig. 6b**; $P < 0.001$ for all dimensionalities above 6). While a very low-dimensional model leads to low R^2 values in both the original neural data and the surrogate datasets, as model dimensionality increases, the R^2 value of the original neural dataset increases disproportionately, separating from the surrogate datasets.

The structure we investigated here is consistent with a simple class of dynamical systems[19,46], but certainly the underlying mechanism generating population responses is more complex. It is essential to note that the present neural population control does not attempt to distinguish between linear dynamical models and other models (dynamical or otherwise); it specifically tests whether there is more linear dynamical structure than expected from the primary features in neural data.

DISCUSSION

Neural populations are increasingly studied, compelling the analysis of large neural datasets and the consideration of new scientific hypotheses. However, the future of these analyses hinges on our ability to reliably distinguish novel population-level findings from redescriptions of simpler features of the data. To that end, we developed a neural population control to statistically test whether a population-level result is an expected byproduct of the primary features of temporal, neural and condition correlations. The CFR and TME methods generate surrogate datasets that preserve the primary features but are otherwise random and can thus be meaningfully compared to the original neural data. We applied the neural population control to data from PFC during a working memory task. We found that the presence of a neural readout specific to the decision was significant and may be an interesting form of collective code, whereas the presence of a neural readout specific to the stimulus could be explained by the tuning of single-neurons. Further, we applied this framework to data from motor cortex during a reaching task, demonstrating that population-level dynamics are not an expected byproduct of primary features.

When applying the neural population control framework, interpretational precision is critical. Specifically, consider our finding that the presence of a stimulus readout in PFC is expected from single-neuron tuning. First, this finding does not assert that the stimulus readout is incorrect or absent, nor does it indicate any technical flaw in the analysis method; rather it indicates that we cannot rule out the possibility that the readout is merely a redescription of tuning and thus that we should not necessarily infer evidence of a collective code. Second, and more subtly, any claim that a population-level readout is an expected byproduct is conditioned on the subjective belief that single-neuron tuning is known to be a fundamental feature that exists in data. Should one believe that, instead, the population-level readout of the stimulus is the fundamental feature, one could instead ask if single-neuron tuning is an expected byproduct to that assumption (indeed, some might quite sensibly argue this direction to be more scientifically plausible). Our framework makes no claims as to which features are fundamental but rather quantifies the extent to which structure will appear at the level of the population as a result of a set of specified primary features. Indeed, our framework is conservative, as it assumes the existence of primary features without any mechanistic underpinning; in other words, we do not require the existence of a competing scientific model to produce data with these features (and finding such a model might be difficult). This assumption presents a high bar when compared to specific mechanistic models that correspond to the population structure in question.

At the broadest interpretational level, rejection of the null hypothesis does not prove the existence of a specific population structure. Instead, such a finding rules out a simpler explanation of observing that structure in data. We do not claim that a test that fails to reject this null hypothesis would somehow negate the scientific significance of a population structure. Indeed, these simpler explanations may themselves be scientifically interesting. For example, studies have demonstrated that minimal models of correlations among neurons provide accurate and nontrivial predictions of population activity patterns in primate[32,33] and other vertebrate[27] retina. Additionally, failing to reject this null hypothesis may simply imply that current data or the complexity of experimental behavior is inadequate to elucidate that structure[2].

At a technical level, the CFR and TME methods are complementary and exploit principles that have a long history in neuroscience[27,31–33]. These methods can be applied interchangeably, and their minor differences have little effect on the hypothesis being tested. That said, certainly each method possesses its advantages. CFR generally better preserves the primary features for each surrogate dataset, while TME has the exact primary features in expectation (**Supplementary Figs. 2** and **3**). On the other hand, TME produces more thoroughly randomized surrogates than does CFR. By construction, CFR operates on a finite set of original neural responses, which may allow structure to persist in the surrogate datasets even if it was not stipulated by the null hypothesis. In contrast, TME surrogate datasets are maximally random in the Shannon entropy sense and have no unintended structure. Thus, if it is most crucial to eliminate any structure beyond the primary features, TME is preferred. If it is most important that each surrogate dataset preserves the primary features of the original neural data as close as possible, CFR is preferred. For practical purposes, note also that CFR is more computationally expensive because it requires optimization for each surrogate dataset, whereas TME requires an optimization only once (Online Methods). On another technical point, in this work both the prefrontal and motor applications involved firing rates that were averaged across trials within a given condition. One natural technical question is how this framework works in the single-trial setting. If one works with single-trial time histograms (a single-trial peristimulus time histogram) or rate estimates (as is often done[47–49]), then the neural population control works without further modification. Should one wish to work with spike trains directly, then further assumptions must be made so that means and covariances can be meaningfully calculated (as these features do not apply to point-process data). A rate estimate is one choice; other nonrate choices such as a spike train metric[50] would require further development.

When studying population-level questions in neuroscience, it is important for our hypotheses to be consistent with existing, simpler

features of neural data. Here we have found that it is equally important to quantitatively investigate whether these simpler features themselves reproduce the population structure being considered by that hypothesis. The neural population control may be applied to test a wide range of population hypotheses from essentially any brain area and thus provides a general framework for rigorously resolving debates in the field about the novelty of population level results.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the laboratories of L. Paninski, M. Churchland and K. Shenoy for discussions. We thank M. Churchland, L. Abbott and K. Miller for comments on the manuscript. We thank D. Kobak and C. Machens for discussions about and assistance with the dPCA algorithm. We thank M. Kaufman for help with **Figure 1a**. We thank M. Churchland, M. Kaufman, S. Ryu and K. Shenoy for the motor cortex data. We thank R. Romo and C. Brody for the prefrontal cortex data, downloaded from the CRCNS (available at the time of publication at https://crcns.org/data-sets/pfc/pfc-4). We thank T. Requarth for comments on the manuscript. We thank the 2016 Modeling Neural Activity conference for discussions and for a travel grant to GFE (MH 064537, NSF-DMS 1612914 and the Burroughs-Wellcome Fund). This work was funded by NIH CRCNS R01 NS100066-01, the Sloan Research Fellowship, the McKnight Fellowship, the Simons Collaboration on the Global Brain SCGB325233, the Grossman Center for the Statistics of Mind, the Center for Theoretical Neuroscience, the Gatsby Charitable Trust and the Zuckerman Mind Brain Behavior Institute.

AUTHOR CONTRIBUTIONS

G.F.E. and J.P.C. contributed to all aspects of this study.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Cunningham, J.P. & Yu, B.M. Dimensionality reduction for large-scale neural recordings. *Nat. Neurosci.* **17**, 1500–1509 (2014).
2. Gao, P. & Ganguli, S. On simplicity and complexity in the brave new world of large-scale neuroscience. *Curr. Opin. Neurobiol.* **32**, 148–155 (2015).
3. Stevenson, I.H. & Kording, K.P. How advances in neural recording affect data analysis. *Nat. Neurosci.* **14**, 139–142 (2011).
4. Pillow, J.W. *et al.* Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature* **454**, 995–999 (2008).
5. Stopfer, M., Jayaraman, V. & Laurent, G. Intensity versus identity coding in an olfactory system. *Neuron* **39**, 991–1004 (2003).
6. Saha, D. *et al.* A spatiotemporal coding mechanism for background-invariant odor recognition. *Nat. Neurosci.* **16**, 1830–1839 (2013).
7. Machens, C.K., Romo, R. & Brody, C.D. Functional, but not anatomical, separation of "what" and "when" in prefrontal cortex. *J. Neurosci.* **30**, 350–360 (2010).
8. Mante, V., Sussillo, D., Shenoy, K.V. & Newsome, W.T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
9. Churchland, M.M. *et al.* Neural population dynamics during reaching. *Nature* **487**, 51–56 (2012).
10. Sadtler, P.T. *et al.* Neural constraints on learning. *Nature* **512**, 423–426 (2014).
11. Raposo, D., Kaufman, M.T. & Churchland, A.K. A category-free neural population supports evolving demands during decision-making. *Nat. Neurosci.* **17**, 1784–1792 (2014).
12. Morcos, A.S. & Harvey, C.D. History-dependent variability in population dynamics during evidence accumulation in cortex. *Nat. Neurosci.* **19**, 1672–1681 (2016).
13. Elsayed, G.F., Lara, A.H., Kaufman, M.T., Churchland, M.M. & Cunningham, J.P. Reorganization between preparatory and movement population responses in motor cortex. *Nat. Commun.* **7**, 13239 (2016).
14. Hubel, D.H. & Wiesel, T.N. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J. Neurophysiol.* **28**, 229–289 (1965).
15. Georgopoulos, A.P., Schwartz, A.B. & Kettner, R.E. Neuronal population coding of movement direction. *Science* **233**, 1416–1419 (1986).
16. Rigotti, M. *et al.* The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
17. Murray, J.D. *et al.* Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proc. Natl. Acad. Sci. USA* **114**, 394–399 (2017).
18. Broome, B.M., Jayaraman, V. & Laurent, G. Encoding and decoding of overlapping odor sequences. *Neuron* **51**, 467–482 (2006).
19. Sussillo, D., Churchland, M.M., Kaufman, M.T. & Shenoy, K.V. A neural network that finds a naturalistic solution for the production of muscle activity. *Nat. Neurosci.* **18**, 1025–1033 (2015).
20. Maass, W., Natschläger, T. & Markram, H. Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Comput.* **14**, 2531–2560 (2002).
21. Cunningham, J.P., Gilja, V., Ryu, S.I. & Shenoy, K.V. Methods for estimating neural firing rates, and their application to brain-machine interfaces. *Neural Netw.* **22**, 1235–1246 (2009).
22. London, M., Roth, A., Beeren, L., Häusser, M. & Latham, P.E. Sensitivity to perturbations *in vivo* implies high noise and suggests rate coding in cortex. *Nature* **466**, 123–127 (2010).
23. Gerstein, G.L. & Perkel, D.H. Simultaneously recorded trains of action potentials: analysis and functional interpretation. *Science* **164**, 828–830 (1969).
24. Cohen, M.R. & Kohn, A. Measuring and interpreting neuronal correlations. *Nat. Neurosci.* **14**, 811–819 (2011).
25. Gawne, T.J. & Richmond, B.J. How independent are the messages carried by adjacent inferior temporal cortical neurons? *J. Neurosci.* **13**, 2758–2771 (1993).
26. Yu, B.M. *et al.* Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *J. Neurophysiol.* **102**, 614–635 (2009).
27. Schneidman, E., Berry, M.J. II, Segev, R. & Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **440**, 1007–1012 (2006).
28. Romo, R., Brody, C.D., Hernández, A. & Lemus, L. Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature* **399**, 470–473 (1999).
29. Brody, C.D., Hernández, A., Zainos, A. & Romo, R. Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. *Cereb. Cortex* **13**, 1196–1207 (2003).
30. Shenoy, K.V., Sahani, M. & Churchland, M.M. Cortical control of arm movements: a dynamical systems perspective. *Annu. Rev. Neurosci.* **36**, 337–359 (2013).
31. Tang, A. *et al.* A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro. *J. Neurosci.* **28**, 505–518 (2008).
32. Shlens, J. *et al.* The structure of multi-neuron firing patterns in primate retina. *J. Neurosci.* **26**, 8254–8266 (2006).
33. Shlens, J. *et al.* The structure of large-scale synchronized firing in primate retina. *J. Neurosci.* **29**, 5022–5031 (2009).
34. Kimmel, D., Elsayed, G.F., Cunningham, J.P., Rangel, A. & Newsome, W.T. Encoding of value and choice as separable, dynamic neural dimensions in orbitofrontal cortex. *Cosyne* 2016.
35. Churchland, M.M. & Shenoy, K.V. Temporal complexity and heterogeneity of single-neuron activity in premotor and motor cortex. *J. Neurophysiol.* **97**, 4235–4257 (2007).
36. Hernández, A. *et al.* Decoding a perceptual decision process across cortex. *Neuron* **66**, 300–314 (2010).
37. Kobak, D. *et al.* Demixed principal component analysis of neural population data. *eLife* **5**, e10989 (2016).
38. Romo, R. & Salinas, E. Flutter discrimination: neural codes, perception, memory and decision making. *Nat. Rev. Neurosci.* **4**, 203–218 (2003).
39. Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Ryu, S.I. & Shenoy, K.V. Cortical preparatory activity: representation of movement or first cog in a dynamical machine? *Neuron* **68**, 387–400 (2010).
40. Fetz, E.E. Are movement parameters recognizably coded in the activity of single neurons. *Behav. Brain Sci.* **15**, 679–690 (1992).
41. Scott, S.H. Population vectors and motor cortex: neural coding or epiphenomenona? *Nat. Neurosci.* **3**, 307–308 (2000).
42. Churchland, M.M., Afshar, A. & Shenoy, K.V. A central source of movement variability. *Neuron* **52**, 1085–1096 (2006).
43. Kaufman, M.T. *et al.* Roles of monkey premotor neuron classes in movement preparation and execution. *J. Neurophysiol.* **104**, 799–810 (2010).
44. Kaufman, M.T., Churchland, M.M., Ryu, S.I. & Shenoy, K.V. Cortical activity in the null space: permitting preparation without movement. *Nat. Neurosci.* **17**, 440–448 (2014).
45. Michaels, J.A., Dann, B. & Scherberger, H. Neural population dynamics during reaching are better explained by a dynamical system than representational tuning. *PLOS Comput. Biol.* **12**, e1005175 (2016).
46. Hennequin, G., Vogels, T.P. & Gerstner, W. Optimal control of transient dynamics in balanced networks supports generation of complex movements. *Neuron* **82**, 1394–1406 (2014).
47. Ecker, A.S. *et al.* State dependence of noise correlations in macaque primary visual cortex. *Neuron* **82**, 235–248 (2014).
48. Park, I.M., Meister, M.L., Huk, A.C. & Pillow, J.W. Encoding and decoding in parietal cortex during sensorimotor decision-making. *Nat. Neurosci.* **17**, 1395–1403 (2014).
49. Kaufman, M.T., Churchland, M.M., Ryu, S.I. & Shenoy, K.V. Vacillation, indecision and hesitation in moment-by-moment decoding of monkey motor cortex. *Elife* **4**, e04677 (2015).
50. Victor, J.D. Spike train metrics. *Curr. Opin. Neurobiol.* **15**, 585–592 (2005).

ONLINE METHODS

Experimental design and recordings. Motor cortex data was recorded and described in previous work[9]. In brief, four male rhesus monkeys (J, N, A and B) performed delayed reaches to radially arranged targets on a frontoparallel screen. Monkeys A and B performed straight reaches with different speeds and distances (28 reaching conditions); monkeys J and N performed both straight and curved reaches (108 reaching conditions). Recordings were made from primary motor and dorsal premotor cortices with single electrodes (datasets A, B, J1, J2, J3, J4 and N) and chronically implanted 96-electrode arrays (datasets J-array and N-array). Large populations were recorded (64, 74, 50, 58, 55, 50, 170, 118 and 218 neurons for datasets A, B, J1, J2, J3, J4, J-array, N and N-array, respectively). Firing rates were calculated by averaging spiking activity across trials for each reaching condition, smoothing with a 24-ms Gaussian kernel and sampling the result at 10-ms intervals. See Churchland *et al.*[9] for all further details about subjects and experiment. We further excluded one outlier neuron from monkey A that had an unrealistically high firing rate.

PFC data was recorded and described in previous work[28,29]. In brief, two male rhesus monkeys (RR15 and RR14) performed a working-memory task. Two vibrotactile stimuli were delivered to one digit of the hand for 500 ms each, separated by an interstimulus delay period. Monkeys received a juice reward for discriminating and reporting the relative frequency of the two stimuli. Neural responses were recorded from PFC via an array of seven independent microelectrodes. See Romo *et al.*[28] and Brody *et al.*[29] for all further details about subjects and experiment. We followed the neuron selection criteria and firing rates computation method reported in Kobak *et al.*[37]. First, we selected only the sessions in which all six frequencies (10, 14, 18, 24, 30 and 34 Hz) were used for the first stimulus and in which the monkeys made the correct choice. Second, we included only neurons that had responses in all 12 possible conditions (all combination of 6 stimuli and 2 choices) with at least 5 trials per condition and firing rates of less than 50 spikes per s (571 and 217 neurons from monkey RR15 and monkey RR14, respectively). Third, firing rates were calculated by averaging spiking activity across trials for each stimulus condition, smoothing with a 50-ms Gaussian kernel and sampling the result at 10-ms intervals.

For all datasets, the sample sizes were similar to those reported in the field and no randomization or blinding was used to assign subjects and conditions (see further details in Churchland *et al.*[9], Romo *et al.*[28] and Brody *et al.*[29]). We followed two further preprocessing steps used in previous work[9]. First, responses for each neuron were soft-normalized to approximately unity firing rate range (divided by a normalization factor equal to the firing rate range + 5 spikes per s). Second, responses for each neuron were mean-centered at each time by subtracting the mean activity across all conditions from each condition's response, because the analyses in this work focus on aspects of population responses that differ across conditions. In our statistical tests, no assumptions were made about the normality or other assumptions on the distribution of surrogate data (see additional information in the attached **Life Sciences Reporting Summary**).

Quantifying primary features across different modes of the data. Each dataset, processed as above, formed a tensor, $X \in \mathbb{R}^{T\times N\times C}$, across T time points, C conditions and N neurons. To quantify the primary temporal, neural and condition features, we calculated the marginal mean and covariance across each of these three modes. Regarding the mean, we followed standard practice and, without loss of generality, centered the data to form a tensor $\bar{X} \in \mathbb{R}^{T\times N\times C}$, which had zero mean across the temporal mode

$$\left(\frac{1}{NC}\sum_{n=1}^{N}\sum_{c=1}^{C}\bar{X}(:,n,c) = 0\right)$$

and similar for the neuron and condition modes. This mean-centering operation can be accomplished by sequentially calculating and subtracting the mean vectors across each mode (in any mode order) or equivalently by calculating and subtracting the least-norm marginal mean tensor $M \in \mathbb{R}^{T\times N\times C}$, such that $\bar{X} = X - M$ (**Supplementary Note 1**).

With this zero-mean dataset $\bar{X}$, we then calculated the covariance matrices across times, neurons and conditions, specifically:

$$\Sigma_T = \sum_{n=1}^{N}\sum_{c=1}^{C}\bar{X}(:,n,c)\bar{X}(:,n,c)^\top \in \mathbb{R}^{T\times T}$$

$$\Sigma_N = \sum_{t=1}^{T}\sum_{c=1}^{C}\bar{X}(t,:,c)\bar{X}(t,:,c)^\top \in \mathbb{R}^{N\times N}$$

$$\Sigma_C = \sum_{t=1}^{T}\sum_{n=1}^{N}\bar{X}(t,n,:)\bar{X}(t,n,:)^\top \in \mathbb{R}^{C\times C}$$

The marginal mean tensor and covariance matrices quantify the basic univariate and pairwise structure of the data across each of the temporal, neural and condition modes. As a technical point, note that other ways of estimating these moments can also be used without any change to the neural population control. For example, regularized covariance estimators are often computed to incorporate prior beliefs about these moments; should one use such a method, the null hypothesis of neural population control would embody the posterior belief of these moments, given the data.

Generating surrogate data with the corrected Fisher randomization (CFR) method. Starting from the zero marginal mean data tensor $\bar{X} \in \mathbb{R}^{T\times N\times C}$, we randomized the data by shuffling: we permuted the condition labels for the responses of each neuron across time. The standard shuffling procedure was done independently across neurons, resulting in a shuffled tensor $\bar{S}_0 \in \mathbb{R}^{T\times N\times C}$. Forming this tensor will also have destroyed the first-order and second-order features of the original neural data. To retain these primary features, we introduced a readout weight matrix, $K \in \mathbb{R}^{N\times N}$, such that the resulting surrogate tensor $S \in \mathbb{R}^{T\times N\times C}$ had the correct marginal means and covariances. That is, the surrogate tensor S is the readout:

$$\bar{S}(:,:,c) = \bar{S}_0(:,:,c)K,\ \forall c \in [1,\ldots,C]$$

$$S = \bar{S} + M$$

where $\bar{S}_0(:,:,c) \in \mathbb{R}^{T\times N}$ is one condition of the shuffled tensor $\bar{S}_0$ corresponding to condition c and M is the marginal mean tensor. To ensure that $\bar{S}_0$ has mean zero across all modes, we constrained K to have unit eigenvector with zero eigenvalue (**Supplementary Note 2**). This constraint ensures that the shuffled $\bar{S}$ has zero marginal mean across all the tensor modes. What remains is to optimize K such that the marginal covariances of the surrogate datasets are as matched as possible to those of the original data. We quantified the deviation of the original marginal covariances with the following three cost functions:

$$f_T = \frac{\left\|\Sigma_T - \sum_{c=1}^{C}\bar{S}_0(:,:,c)KK^\top\bar{S}_0(:,:,c)^\top\right\|_F^2}{\sum_{t=1}^{T}e_T(t)}$$

$$f_N = \frac{\left\|\Sigma_N - K^\top\left[\sum_{c=1}^{C}\bar{S}_0(:,:,c)^\top\bar{S}_0(:,:,c)\right]K\right\|_F^2}{\sum_{n=1}^{N}e_N(n)}$$

$$f_C = \frac{\left\|\Sigma_C - \sum_{t=1}^{T}\bar{S}_0(t,:,:)^\top KK^\top\bar{S}_0(t,:,:)\right\|_F^2}{\sum_{c=1}^{C}e_C(c)}$$

where Σ_T, Σ_N and Σ_C are the temporal, neural and condition covariance matrices, respectively, with eigenvalue vectors e_T, e_N and e_C. To find the desired linear readout ($\hat{K}$), we solved:

$$\hat{K} = \mathrm{argmin}_{K\in\mathbb{R}^{N\times N}}(f_T + f_N + f_C),\ \text{subject to}\ K1 = 0$$

This objective can be optimized using any standard gradient descent package (we used the Manopt and LDR libraries[51,52]) and will result in a readout matrix

that retains the marginal covariance of the original neural data to the extent possible. The resulting surrogates will thus be random in the sense that population structure beyond these primary features should be absent, but constrained in the sense that the same primary features as in the original neural data are maintained (up to the minimum error achieved by the optimization). As an implementation note, we chose the readout to be in the neural space because it is commonly used in systems neuroscience, but the above approach could be implemented similarly by reading out the condition or temporal modes (our software implementation takes that choice as an input). On a similar point of technical detail, above we chose to initially shuffle neurons across conditions as it is a standard choice (and one that agrees with the applications in prefrontal and motor cortices shown in the results), but again our software implementation takes the shuffle mode (conditions, neurons or time) as an input.

Generating surrogate data with the tensor maximum entropy (TME) method. A complementary approach to generating surrogate datasets that preserve the primary features of the neural data is to follow the principle of maximum entropy modeling. In this context, that principle dictates that surrogate data should be drawn from the distribution that is maximally random (i.e., which requires the fewest additional assumptions) but obeys the constraints of having the correct first-order and second-order marginal moments. Specifically, our maximum entropy objective was:

$$\hat{p}(S) = \text{argmax}_{p(S)} - \int p(S)\log(p(S))dS, \text{ subject to}$$

$$\int p(S)\,dS = 1$$

$$\mathrm{E}_p[S] = M$$

$$\overline{S} = S - M$$

$$\mathrm{E}_p\left[\sum_{n=1}^{N}\sum_{c=1}^{C}\overline{S}(:,n,c)\overline{S}(:,n,c)^\top\right] = \Sigma_T$$

$$\mathrm{E}_p\left[\sum_{t=1}^{T}\sum_{c=1}^{C}\overline{S}(t,:,c)\overline{S}(t,:,c)^\top\right] = \Sigma_N$$

$$\mathrm{E}_p\left[\sum_{t=1}^{T}\sum_{n=1}^{N}\overline{S}(t,n,:)\overline{S}(t,n,:)^\top\right] = \Sigma_C$$

where $S \in \mathbb{R}^{T\times N\times C}$ is the surrogate random variable, and $\mathrm{E}_p[.]$ denotes expectation with respect to the distribution p. Intuitively, with first and second moment constraints, one expects this distribution to be $\mathcal{N}$ Gaussian. While that is true, the solution is nontrivial (**Supplementary Notes 3** and **4**). Using the standard Lagrangian method and Kronecker algebra, we derived the maximum entropy distribution:

$$\hat{p}(\text{vec}(S)) = \mathcal{N}(\text{vec}(M), \Psi)$$

$$\Psi = \frac{1}{2}(Q_C \otimes Q_N \otimes Q_T)(\Lambda_C \oplus \Lambda_N \oplus \Lambda_T)^{-1}(Q_C \otimes Q_N \otimes Q_T)^\top$$

$$e_T(t) = \frac{1}{2}\sum_{n=1}^{N}\sum_{c=1}^{C}\frac{1}{\lambda_T(t) + \lambda_N(n) + \lambda_C(c)}, \forall t \in [1,\ldots,T]$$

$$e_N(n) = \frac{1}{2}\sum_{t=1}^{T}\sum_{c=1}^{C}\frac{1}{\lambda_T(t) + \lambda_N(n) + \lambda_C(c)}, \forall n \in [1,\ldots,N]$$

$$e_C(c) = \frac{1}{2}\sum_{t=1}^{T}\sum_{n=1}^{N}\frac{1}{\lambda_T(t) + \lambda_N(n) + \lambda_C(c)}, \forall c \in [1,\ldots,C]$$

where Q_T, Q_N and Q_C are the known eigenvector matrices, and e_T, e_N and e_C are the known eigenvalues of the true marginal covariance matrices Σ_T, Σ_N and Σ_C, respectively. Λ_T, Λ_N and Λ_C are diagonal matrices with diagonal elements λ_T, λ_N and λ_C, respectively (the Lagrange multipliers). We numerically solved for the multiplier values in terms of the given eigenvalues and reached the exact solution (i.e., zero error, to machine precision). Note that this distribution is defined over a tensor variable, and thus its covariance matrix $\Psi \in \mathrm{R}^{TNC\times TNC}$ can easily be on the order of $10^6 \times 10^6$ for a modest dataset (e.g., a dataset with 100 neurons and 100 conditions recorded from 100 timepoints), which is prohibitively large for memory and runtime considerations. Left unaddressed, sampling surrogate data from this distribution would be infeasible. To address this challenge, we exploited the Kronecker structure[53] to efficiently operate with these matrices and exactly sample from this tensor distribution. It is worth noting that, in contrast to the CFR method, the samples from this maximum entropy distribution maintain the specified primary features in expectation (i.e., each individual surrogate sample will have differences in the primary features due to finite sampling along each mode). On the other hand, TME has the key virtue that, by construction, surrogates will have no structure beyond what is specified, whereas CFR only partially achieves this goal via shuffling.

Extensions to other surrogate types. The procedures described so far generate surrogate-TNC datasets that preserve the primary features across times, neurons and conditions. To constrain for only temporal features, or temporal and neural features, slight modifications were required. In CFR, the optimization objective was accordingly modified (surrogate-T: optimized only f_T and added only the temporal mean; surrogate-TN: optimized both f_T and f_N and added the temporal and neural means). Similarly, in TME, the constraints were modified (surrogate-T: constrained only temporal covariance and added only the temporal mean; surrogate-TN: constrained both temporal and neural covariance and added both the temporal and neural means). This discussion also makes clear that both methods can be easily extended to other modes (and to any number of modes) that might be available in other recording contexts by a similar approach; our software implementation directly handles additional modes.

Quantifying structure in motor cortex: low-dimensional dynamical systems. Per standard practice, we analyzed data during the 400-ms duration reflecting the movement response ($\underline{T}$) and projected the data onto the top $\underline{N}$ principal components (PCs) of the data to produce a reduced tensor $\underline{X} \in \mathbb{R}^{\underline{T}\times\underline{N}\times C}$, where $\underline{N} < N$ obtained by cross-validation (**Supplementary Fig. 8**). The linear dynamical system models the temporal evolution of these low-dimensional neural trajectories as fixed across conditions, namely:

$$\dot{\underline{X}}(:,:,c) \approx \underline{X}(:,:,c)J, \forall c \in [1,\ldots,C].$$

where $J \in \mathrm{R}^{\underline{N}\times\underline{N}}$ is the dynamics matrix determining the flow field. $\underline{N}$ thus determines the dimensionality of the model. We fit the model with:

$$\hat{J} = \text{argmin}_{J\in\mathrm{R}^{\underline{N}\times\underline{N}}}\frac{\sum_{c=1}^{C}\|\dot{\underline{X}}(:,:,c) - \underline{X}(:,:,c)J\|_F^2}{\sum_{c=1}^{C}\|\dot{\underline{X}}(:,:,c)\|_F^2}$$

The solution of the above objective function can be analytically obtained by least-squares, and the quality of the fit is quantified by the coefficient of determination (R^2), which equals one minus the minimum normalized error achieved. We also quantified the generalization performance of the model by performing leave-one-condition-out cross-validation (LOOCV) on the reconstruction of $\dot{\underline{X}}(:,:,c_{\text{test}})$ from $\hat{J}$, which was appropriately estimated from data that did not include c_{test}. We repeated this procedure for $c_{\text{test}} = [1,\ldots,C]$, yielding a LOOCV R^2 statistic.

Quantifying structure in PFC: identifying population readouts. To identify stimulus- and decision-specific population readouts in PFC, we used demixed principal component analysis (dPCA). In brief, dPCA starts by performing different marginalization procedures of data to produce multiple datasets, each reflecting one of the task parameters. Then, dPCA identifies dimensions (dPCs) that minimize the reconstruction error of each marginalization of data. Unlike PCA, which maximizes variance, dPCA produces projections with high variance and good demixing of the specified covariates (see Kobak *et al.*[37] for dPCA details).

For the original data and for each surrogate dataset, we allowed dPCA to find at most 30 dPCs, after which we selected the top component that represented the

doi:10.1038/nn.4617

stimulus and the top component that represented the decision in each dataset. We projected the original and surrogate responses onto their top dPCs and quantified the variance captured by these projections during the relevant epochs. The stimulus-projection variance was based on the epoch starting 100 ms after the first stimulus presentation and ending at the onset of the second stimulus. The decision-projection variance was based on the epoch starting 100 ms after the second stimulus presentation and ending at the second stimulus offset. In addition to the conventional percentage variance explained, we computed another variance statistic, the percentage reconstruction variance (as used in Kobak *et al.*[37]), defined as:

$$\frac{||X_N||_F^2 - ||X_N - \mathbf{v}\mathbf{d}^\top X_N||_F^2}{||X_N||_F^2} \times 100\%$$

where $X_N \in \mathbb{R}^{N \times TC}$ is the data reshaped along the neuron mode, $\mathbf{d} \in \mathbb{R}^N$ is the top dPC and $\mathbf{v} \in \mathbb{R}^N$ is the decoder vector mapping the projection ($\mathbf{d}^\top X_N$) to the neural space (see Kobak *et al.*[37] for the encoder and decoder description).

Data availability. The datasets from motor cortex analyzed in the current study are available upon reasonable request from the authors of Churchland *et al.*[9]. The datasets from PFC analyzed during the current study are available at https://crcns.org/data-sets/pfc/pfc-4.

Code availability. A code package for the CFR method is available at https://github.com/gamaleldin/CFR. A code package for the TME method is available at https://github.com/gamaleldin/TME.

51. Boumal, N., Mishra, B., Absil, P.A. & Sepulchre, R. Manopt, a Matlab toolbox for optimization on manifolds. *J. Mach. Learn. Res.* **15**, 1455–1459 (2014).
52. Cunningham, J.P. & Ghahramani, Z. Linear dimensionality reduction: survey, insights, and generalizations. *J. Mach. Learn. Res.* **16**, 2859–2900 (2015).
53. Gilboa, E., Saatçi, Y. & Cunningham, J.P. Scaling multidimensional inference for structured Gaussian processes. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 424–436 (2015).

nature research

Corresponding author(s): John Cunningham

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

The sample size are comparable to, or higher, than those generally employed in the field.

2. Data exclusions

Describe any data exclusions.

The criteria is described in Methods (subsection 1)

3. Replication

Describe whether the experimental findings were reliably reproduced.

N/A

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

No randomization was used.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

No blinding used.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☒	☐	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Nature Neuroscience: doi:10.1038/nn.4617

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	There are two main custom software packages that were designed and used in this study (CFR and TME packages).

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	N/A

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	N/A

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	N/A
b. Describe the method of cell line authentication used.	N/A
c. Report whether the cell lines were tested for mycoplasma contamination.	N/A
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	N/A

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	N/A

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	N/A

Nature Neuroscience: doi:10.1038/nn.4617