Running title: Noise correlation and multivariate responses

Understanding multivariate brain activity: evaluating the effect of voxelwise noise correlations on population codes in functional magnetic resonance imaging

Ru-Yuan Zhang[1], Xue-Xin Wei[2], Kendrick Kay[3]

[1]Center for Magnetic Resonance Research, Department of Neuroscience, University of Minnesota, Minneapolis, MN 55455

[2]Department of Statistics, Center for Theoretical Neuroscience, and Zuckerman Mind Brain Behavior Institute, Columbia University, New York. NY. 10027

[3]Center for Magnetic Resonance Research, Department of Radiology, University of Minnesota, Minneapolis, MN 55455

Correspondence:
Ru-Yuan Zhang
Center for Magnetic Resonance Research
University of Minnesota
2021 6h St SE
Falcon Heights, MN 55455-3007
585-752-6673
zhan1217@umn.edu

ABSTRACT

Previous studies have shown that neurons exhibit trial-by-trial correlated activity and that such noise correlations (NCs) greatly impact the accuracy of population codes. Meanwhile, multivariate pattern analysis (MVPA) has become a mainstream approach in functional magnetic resonance imaging (fMRI), but it remains unclear how NCs between voxels affect MVPA accuracy. Here, we tackle this issue by combining voxel-encoding modeling and MVPA. We focus on a well-established form of NC, tuning compatible noise correlation (TCNC), whose sign and magnitude are systematically related to the tuning similarity between two units. We first replicate the classical finding that TCNCs impair population codes in a standard neuronal population. We then extend our analysis to fMRI data and show that voxelwise TCNCs do not impair and can even improve MVPA accuracy when TCNCs are strong or the number of voxels (i.e., pool size) is large. Further computational analyses demonstrate that the discrepancy between the effect of TCNCs in both neuronal and voxel populations can be explained by units' tuning heterogeneity and pool sizes. Our results provide a theoretical foundation to understand the effect of correlated activity on population codes under different neural measurements. Our results also suggest that the conventional MVPA approach conflates several key aspects of neural coding, and future fMRI research may need to characterize different properties (e.g., NC) of multivariate responses to assess the accuracy of population codes.

Keywords:

Brain decoding, Noise correlation, Multivariate pattern analysis, Voxel-encoding model, Neural variability

SIGNIFICANCE STATEMENT

Noise correlations (NC) is the key component of the multivariate response distribution and thus characterizing their effect on population codes is the cornerstone for understanding probabilistic computation in the brain. Despite extensive studies on NCs in unit recording experiments, little is known with respect to their role in functional magnetic resonance imaging (fMRI). We characterize the effect of voxelwise NC by building voxel-encoding models and directly quantifying the amount of information in multivariate fMRI data. In contrast to the detrimental effects of NC implied in unit recording data, we find that the voxelwise NCs can enhance information codes in several situations. Our work provides a new approach to characterize population responses in fMRI research.

INTRODUCTION

Understanding how neural populations encode information and guide behavior is the central question in cognitive neuroscience. In a neural population, many units exhibit correlated activity. In computational neuroscience, *signal correlation* (SC) refers to the similarity between the tuning functions of two neurons, and *noise correlation* (NC) refers to the correlation between two neurons' trial-by-trial responses towards repetitive presentations of the same stimulus (Averbeck et al., 2006; Cohen and Kohn, 2011; Kohn et al., 2016).

Prior studies in neurophysiology have discovered that many neurons that share similar tuning functions (i.e., a positive SC) also tend to have a weak positive NC, a phenomenon pervasively existing across several brain regions (Lee et al., 1998; Constantinidis and Goldman-Rakic, 2002; Averbeck and Lee, 2003; Gutnisky and Dragoi, 2008; Smith and Kohn, 2008; Huang and Lisberger, 2009; Jermakowicz et al., 2009; Ecker et al., 2010). In this paper we denote this type of NC as *tuning-compatible noise correlation* (TCNC) because the sign and the magnitude of the NC are systematically related to the SC between a pair of neurons. A bulk of theoretical and empirical work has shown that NCs have substantial impacts on population codes. For example, the seminal study by Zohary et al. (1994) has demonstrated that TCNCs limit the amount of information in a neural population as the noise is shared by neurons and

cannot be simply averaged away. Later on researchers realized that this detrimental effect of TCNC is mediated by other factors, such as the form of NC, heterogeneity of tuning functions (Ecker et al., 2011) and its relevance to behavior (Haefner et al., 2013). Given that NC is the key component of the trial-by-trial multivariate response distribution, understanding the effect of NC on population codes is the cornerstone to uncover the generative model that predicts the response distribution given a stimulus (Ma and Jazayeri, 2014).

The study of NC was historically impeded by the technical barrier to simultaneously measure the activity of many neurons in unit recording experiments. In contrast, functional magnetic resonance imaging (fMRI) is advantageous since it can naturally measure the activity of many voxels throughout the entire brain. On the other hand, imaging scientists often use multivariate pattern analysis (MVPA) to assess the accuracy of population codes (Tong and Pratte, 2012; Haxby et al., 2014). However, an above-chance decoding accuracy in MVPA does not specify the detailed representational structure underlying multivariate voxel responses. For example, Fig. 1 illustrates a simple two-voxel model of multivariate decoding. The decoding accuracy in the original state (Fig. 1A) can be improved (e.g., by attention, learning) via either the further separation of mean responses (Fig. 1B) or the warped covariance geometry (Fig. 1C). This example highlights the significant impacts of the shape of the response distribution on population codes and these effects cannot be easily disentangled by the conventional MVPA approach (Naselaris and Kay, 2015).

The magnitude and the structure of NCs in fMRI data still remain largely unknown. People have realized that the NCs can influence MVPA accuracy (Cox and Savoy, 2003) but their precise quantitative relationship has yet been established. A recent study by Ryu and Lee (2018) symmetrically evaluated the impact of three factors—retinotopic distance, cortical distance and tuning similarity—on NCs in early visual cortex and found that the tuning similarity is the major determinant for NCs. However, the NCs were defined as resting-state functional connectivities in visual cortex. Furthermore, van Bergen and Jehee (2018) systematically evaluated voxelwise NCs in human V1 to V3 and showed that the magnitude of NCs monotonically increases as tuning similarity (i.e., SC) increases. They also found that voxelwise NCs in the striate

cortex are in general higher than the NCs in the extrastriate cortex. This provides specific evidence supporting the existence of voxelwise TCNC. Furthermore, one recent study also found that a multivariate classifier can exploit the voxelwise NCs to decode population information (Bejjanki et al., 2017). These results implied that TCNCs in macroscopic fMRI data might have substantial and similar impacts on population codes as in microscopic unit recording data.

In the present study, we combined MVPA and the voxel-encoding modeling approach to assess how the magnitude and form of NCs impact population codes in fMRI data. We assess the accuracy of population codes by calculating MVPA decoding accuracy and the amount of information in a voxel population. The voxel-encoding model proposed in this study allows us to systematically manipulate response parameters (i.e., voxel tuning) so as to form generative models in different scenarios (Naselaris et al., 2011). We first derived the quantitative relationship between the decoding accuracy and the strength of NCs. We then directly calculated the amount of information as a function of NCs in a voxel population. Both methods demonstrated that the accuracy of population codes in fMRI data acts as U-shaped functions as the strength of TCNC increases. Notably, all these analyses in voxel populations were compared against the classical findings in neuronal populations. We will show that the effects of NCs on population codes, no matter in a voxel or a neuronal population, are strongly mediated by the unit tuning heterogeneity. Our results provide a comprehensive theoretical understanding of population codes in fMRI research.

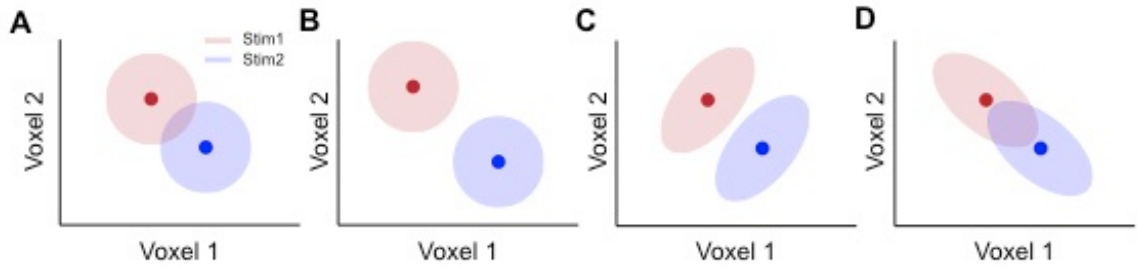

Figure 1. A two-voxel model of MVPA. The pool consists of two responsive voxels and two color disks represent the trial-by-trial response distributions towards two different stimuli. Panel A illustrates the original state of the population responses. The decoding accuracy can be potentially improved via either a bigger separation of the mean population response (panel B) or a warped covariability structure (panel C). Representational structures in panels B and C can be both treated as

"improved population codes" but they have distinct underlying mechanisms. Panel D illustrates that changing the covariability in a specific direction can also worsen decoding.

MATERIALS AND METHODS

Previous endeavors of brain decoding generally fell into two broad categories: classification of stimuli into discrete categories (Kamitani and Tong, 2005) and estimation of a continuous stimulus variable (Vintch and Gardner, 2014). We thus evaluated the effect of NC in brain decoding in two hypothetic tasks—a stimulus-classification task and a stimulus-estimation task. We will first introduce the simulation on a neuronal population and then specify the voxel-encoding model used to generate responses of a voxel population.

Simulation of multivariate neuronal responses

Neuron-encoding model. The neuron-encoding model assumes 180 orientation-selective neurons whose preferred orientations equally span between (0, 180°]. Their tuning curves can be described as:

$$g_k(s) = \alpha + \beta * e^{\gamma*(\cos(\frac{\pi}{90}(s-\varphi_k))-1)} \quad (1)$$

where s is the stimulus and φ_k indicates the preferred orientation of the k-th neuron. a is the baseline firing rate, β is the dynamic range of the response and γ controls the width of the tuning curve. We set the parameter values a = 1, β = 19 and γ = 2, resulting in a tuning curve with the maximum firing rate at 20 Hz. This tuning curve is consistent with previous theoretical work (Ecker et al., 2011) and empirical measurements in the primary visual cortex in primates (Ecker et al., 2010).

Based on this setting, the mean of neuronal population response can be represented by $\boldsymbol{G}(s)=[g_i(s)]$. However, empirically measured neuronal responses varied trial-by-trial. Thus trial-by-trial population responses of neurons follow a multivariate Poisson distribution with mean as $\boldsymbol{G}(s)$. We detail the covariance of this distribution in the following section.

Noise correlation and covariance. We proposed three types of NCs for neuronal data: angular-based tuning compatible noise correlation (aTCNC), curve-based tuning compatible noise correlation (cTCNC) and shuffled noise correlation (SFNC).

Table 1. List of symbols

Symbol	Meaning
NC	Noise correlation
SC	Signal correlation
fMRI	Functional magnetic resonance imaging
MVPA	Multivariate pattern analysis
TCNC	Tuning-compatible noise correlation
aTCNC	Angular-based tuning-compatible noise correlation
cTCNC	Curve-based tuning-compatible noise correlation
SFNC	Shuffled noise correlation
R^{aTCNC}	Angular-based tuning-compatible noise correlation matrix
$R^{c\mathrm{TCNC}}$	Curve-based tuning-compatible noise correlation matrix
R^{SFNC}	Shuffled noise correlation matrix
c_{nc}	Noise correlation coefficient
W	Linear weighting matrix from neuronal to voxel responses

Several theoretical studies assumed the NC between a pair of neurons is an exponential function of the angular difference between their preferred orientations:

$$r_{ij}^{aTCNC} = e^{(-\frac{|\varphi_i - \varphi_j|}{L} * 90/pi)}, \quad (2)$$

This equation specifies that the NC between two neurons diminishes as their preferred orientations fall apart. The parameter L controls the magnitude of such decay. We denote this type of NC as angular-based tuning-compatible noise correlation and the correlation matrix as $R^{c\mathrm{TCNC}}$. Here we set L = 1 for simplicity. Ecker et al. (2011) have shown that the parametric form of NC and the value of L does not qualitatively change the result of the simulation, as long as the generated correlation matrix is positive definite. Note that by this definition aTCNCs are always positive (i.e., range 0~1, see Fig. 3A).

The second type is the curve-based tuning compatible noise correlation (cTCNC). cTCNC follows the formal definition of tuning compatible noise correlation. In this case, the NC between a pair of neurons is proportional to their SC (i.e., correlation of their orientation tuning curves):

$$r_{ij}^{cTCNC} = (1-\delta_{ij}) * corr(g_i(\mathbf{S}), g_j(\mathbf{S})) + \delta_{ij}, \quad (3)$$

where δ_{ij} is the Kronecker delta (δ_{ij}=1 if i=j and δ_{ij}= 0 otherwise). $\boldsymbol{S}$ indicates all possible orientations between (0, 180°) and r_{ij}^{cTCNC} is the NC between the i-th and the j-the neurons. $g_i(\mathbf{S})$ and $g_j(\mathbf{S})$ are their tuning curves (see Eq. 1). We denote $R^{c\mathrm{TCNC}}$ as the correlation matrix. Note that unlike aTCNCs, cTCNCs can be negative (see Fig. 3B). Also, the key difference between cTCNC and aTCNC is that cTCNC does not rely on the functional form of tuning curves. In other words, cTCNC can be computed given any form of tuning curves but aTCNC can only be computed from unimodal tuning curves. This is particularly important for calculations of voxelwise NCs (see below).

In the third case, we shuffled the NCs between all pairs of neurons in $R^{c\mathrm{TCNC}}$ such that its rows and columns were rearranged in the same random order but the diagonal of the matrix kept intact (Fig. 3C). We term this type of NC as shuffled NCs (SFNCs) since the correlation bears no resemblance to the neuronal SC. We denote the correlation matrix of SFNCs as R^{SFNC}. R^{SFNC} can serve as a comparison for $R^{c\mathrm{TCNC}}$ since shuffling does not alter the overall magnitude of NCs in a neuronal population.

Furthermore, we assumed Poisson noise of neural spikes such that the response variance of a neuron is equal to their mean activity of a stimulus.

$$\tau_k^2(s) = g_k(s), \quad (4)$$

where τ_k^2(s) is the response variance of the k-th neuron triggered by the stimulus s. Note that in this case the response variance is stimulus-dependent thus not purely additive. The covariance between two neurons can be expressed as:

$$\mathbf{Q}_{ij} = (1-\delta_{ij}) c_{neuron} * r_{ij} * \tau_i \tau_j + \delta_{ij} \tau_i \tau_j, \quad (5)$$

where c_{neuron} is the coefficient controlling the strength of the neuronal NC. τ_i and τ_j are the standard deviation of responses of the i-th and the j-th neurons (Eq. 4), respectively. δ_{ij} is the Kronecker delta. Given the covariance matrix $\mathbf{Q}_{neuron}$, we can derive the population response noise distribution as:

$$\mathbf{e} \sim N(0, \mathbf{Q}_{neuron}), \quad (6)$$

Simulation of multivariate voxel responses

Voxel-encoding model. The voxel-encoding model uses the same pool of orientation-selective neurons (i.e., 180 neurons with tuning curves defined in Eq. 1) as in the neuron-encoding model. Following the standard approach of voxel-encoding modeling, we further assumed that a single voxel response h_i(s) is the linear combination of all neurons in the neuronal population:

$$h_i(s) = \sum_{k=1}^{180} w_{ki} g_k(s) \quad , \tag{7}$$

where w_{ki} is the connection weight between the k-th neuron to the i-th voxel. We sampled w_{ki} from a uniform distribution:

$$w_{ki} \sim uniform(0, 0.01) \tag{8}$$

This range was used such that the generated fMRI responses less than 10, which is consistent with realistic voxel responses. Given that percent signal change in fMRI rarely exceeds 10%, we believe this range of connection weights can produce more realistic voxel responses.

Again, the mean of voxel population response can be represented by $\boldsymbol{H}$(s)=[h_i(s)]. However, empirically measured voxel responses inevitably vary trial-by-trial and this can be expressed as:

$$\mathbf{b} = H(\mathbf{s}) + \mathbf{e} \, , \tag{9}$$

where **b** is estimated percent signal change from standard general linear modeling and **e** represents the multivariate normal noise distribution:

$$\mathbf{e} \sim N(0, \mathbf{Q}_{vxs}) \, , \tag{10}$$

where $\mathbf{Q}_{vxs}$ is the covariance matrix between voxels, which will be detailed in the following section.

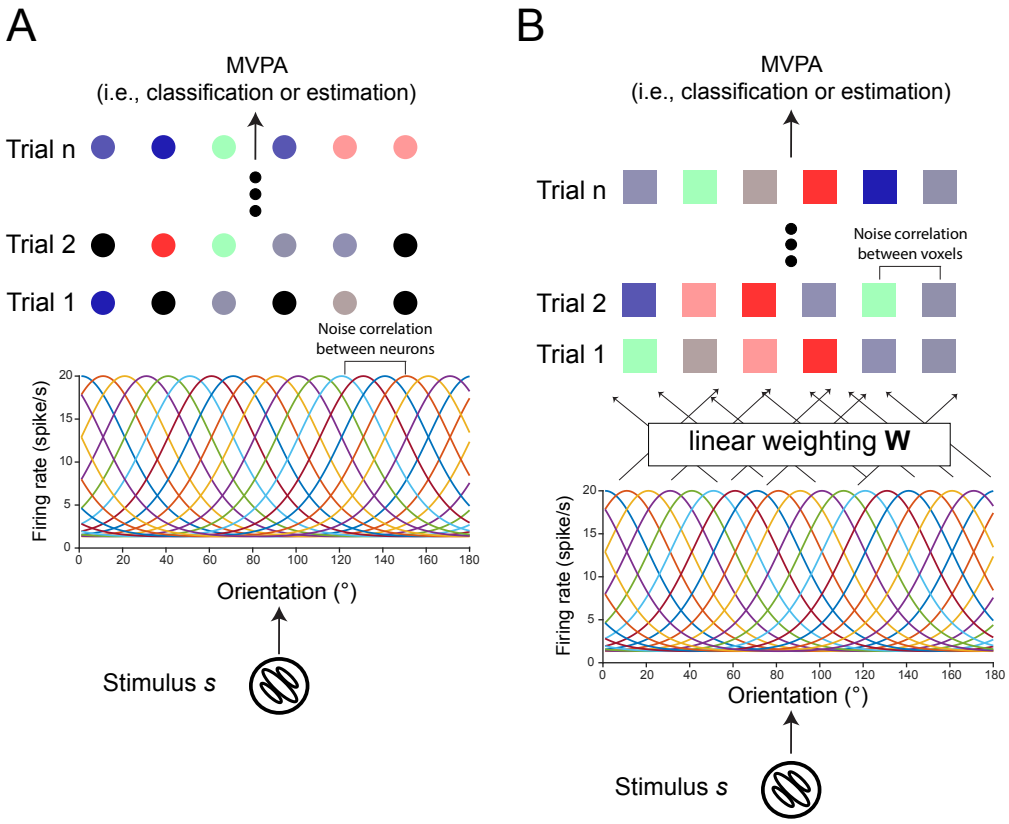

Figure 2. Neuron- and voxel-encoding models. The neuron-encoding model (panel A) proposes a neuronal population with homogenous orientation-selective tuning curves. Each neuron has Poisson-like response variance and the noise correlation between two neurons can be specified with different structures and strength (see Methods). The voxel-encoding model proposes a similar neuronal population and the response of a single voxel is the linear combination of the responses of multiple neurons. The noise correlation between two voxels can be specified using similar methods (see Methods). Note that voxelwise NCs can come from the response variability at both neuronal and voxel levels (see Fig. 6). Using the neural- and the voxel-encoding model, we can generate many trials of neuronal and voxel population responses and perform conventional MVPA on the simulated data. The goal is to examine the multivariate decoding performance as a function of the NC structure and strength between either neurons or voxels. The sample voxel tuning functions will be shown in Fig. 8.

Noise correlation and covariance. We proposed two types of NCs for simulated fMRI data: cTCNC and SFNC. Previous studies have implied the existence of TCNC between voxels (Ryu and Lee, 2018; van Bergen and Jehee, 2018). In addition, the voxelwise NCs inevitably include other non-tuning factors such as head motion, thermal noise, and

vascular noises. These sources of noise might be unrelated to voxel tuning thus more closed to SFNCs.

In the first case, we defined the cTCNC using a similar method as Eq. 3

$$r_{ij}^{cTCNC} = (1-\delta_{ij}) * corr(h_i(\mathbf{S}), h_j(\mathbf{S})) + \delta_{ij}, \tag{11}$$

Note that the TCNC here is based on the tuning curves of two voxels ($h_i(\mathbf{S})$ and $h_j(\mathbf{S})$) not between two neurons.

In the second case, SFNCs were generated using a similar way as in the neuron-encoding model—shuffling the rows and columns in $R^{c\mathrm{TCNC}}$.

We assume the response variances (i.e., τ^2) of all voxels follow a Gamma distribution:

$$\tau_k^2 \sim Gamma(u, v), \tag{12}$$

where u = 9, v = 0.33 are the scale and the shape parameters corresponding to the distribution mean = 3 and variance = 1. Given the response variance of individual voxels and the NC between them, we can calculate the voxelwise covariance matrix $\mathbf{Q}_{vxs}$ using Eq. 5

$$\mathbf{Q}_{ij} = (1-\delta_{ij}) c_{vxs} * r_{ij} * \tau_i \tau_j + \delta_{ij} \tau_i \tau_j, \tag{13}$$

where c_{vxs} is the coefficient controlling the strength of the voxelwise NCs. τ_i and τ_j are the standard deviation of responses of the i-th and the j-th voxels (Eq. 12), respectively. is the Kronecker delta. Similarly, we can further obtain the population response variance **e** using Eq. 10. Note that Eq. 12 only described the variance of response variance across voxels. The multivariate responses of the whole voxel population still follow multivariate Gaussian distribution (Eq. 10).

Differences between the neuron- and the voxel-encoding models

In this section we summarize three key differences between the two encoding models proposed above. The biggest difference is that decoding was performed directly on the simulated neuronal responses in the neuron-encoding model but on the voxel responses, which are the linear combination of the underlying neuronal responses, in the voxel-encoding model. Second, the NCs we manipulated were between neurons in the neuron-

encoding model. In the voxel-encoding model, we either only manipulated the NCs at the stage of voxel activity (Fig. 5) or the NCs in both neuron and voxel stages (Fig. 6). But note that in empirical fMRI studies neuronal NCs are inaccessible so the former case is more related to realistic fMRI data analysis and the latter case only serves to derive theoretical predictions. Third, we assumed the Poisson-like response variance for individual neurons in the neuron-encoding model, which is consistent with the previous theoretical work and empirical findings (Kohn et al., 2016). In this regime, the magnitude of response variance of individual neurons is stimulus-dependent. In the voxel-encoding model, we assumed stimulus-independent additive Gaussian variance for voxels, which is also consistent with one recent computational study (van Bergen and Jehee, 2018).

Data simulation and multivariate pattern analysis

Stimulus-classification task. In the stimulus-classification task, we attempted to classify two stimuli. We manipulated two independent variables: population size (i.e., the number of neurons or voxels) and the strength level of NC coefficient (c_{neuron} in Eq. 5 or c_{vxs} in Eq. 13). We built a linear discriminant using the Matlab function *classify.m*. The linear discriminant assumes that the conditional probability density functions $p(\mathbf{b} \mid s = s_1)$ and $p(\mathbf{b} \mid s = s_2)$ are both normally distributed and the mean and covariance can be estimated from the training data. The classifier was trained on half of the data and tested on the other half.

In the neuron-encoding model, we attempted to classify two stimuli: $s_1 = 92°$, $s_2 = 88°$. The two stimuli were chosen to control the overall task difficulty (i.e., avoid the ceiling or floor effect in classification accuracy). 100 independent simulations were performed and the averaged classification accuracy values across 100 simulations are shown in Fig. 4. In each simulation, we iterated each combination of pool size and c_{neuron} values and performed classification for three forms of NCs. To compensate for potential overfitting as the pool size increases, we set the number of trials as 100 times of the pool size and split the data into two independent parts for training and testing.

In the voxel-encoding model, we reduced the task difficulty and set $s_1 = 70°$, $s_2 = 110°$. 100 independent simulations were performed and averaged. In each simulation, we iterated each combination of pool size and c_{vxs} values. Note that in each simulation and

each pool size we recreated the linear weighting matrix **W** (see Eq. 7) and thus regenerated a new set of voxels. But the same **W** was kept across different c_{vxs} values such that the classification accuracy values were c comparable across different c_{vxs} values.

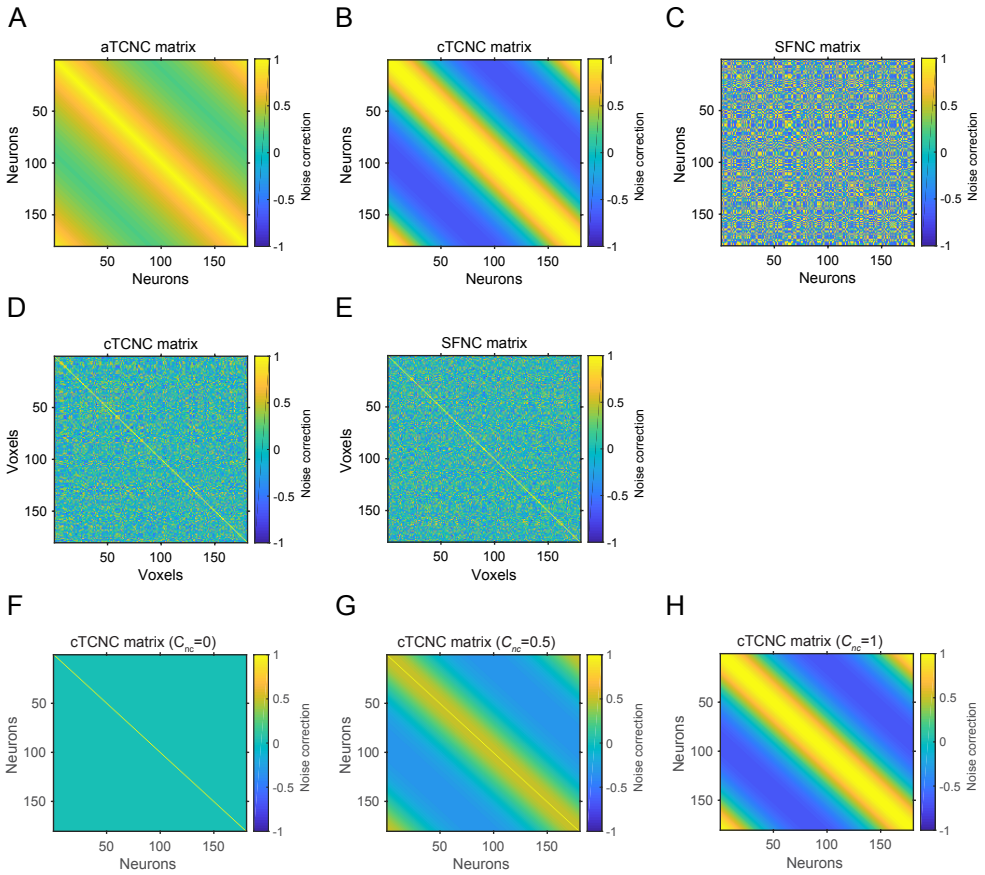

Figure 3. Example noise correlation matrices simulated in a neuronal (panels A-C) and a voxel population (D, E). In the neuronal population (180 neurons), the angular-based TCNC matrix, the curve-based TCNC matrix, and the SFNC matrix are illustrated from left to right. Neurons are sorted according to their preferred orientation from 1 to 180°. In the voxel population (180 voxels), the curve-based TCNC matrix and the SFNC matrix are illustrated. Note that we do not sort the voxels according to their tuning preference. The NC coefficients (c_{nc}) are set to 1 in matrices from A-E. Panels F-H illustrate the cTCNC matrices with NC coefficient values 0, 0.5 and 1, respectively. Note that the panels C and H are identical.

Stimulus-estimation task. In the stimulus-estimation task, a maximum likelihood estimator (MLE) was used to reconstruct the orientation value randomly drawn from (0,

180°] in a trial. Given a population response pattern **b** in the trial, we aimed to find the stimulus s that maximizes the likelihood:

$$\arg\max_{s\in[0,180)} p(\mathbf{b}\,|\,s)\,, \tag{14}$$

. Note that the likelihood functions have been introduced above in the data generative models. Specifically, the data generation distributions are given by the Eqs. 5&6 for neuronal data and the Eqs.9&10 for the voxel data. We numerically evaluated the likelihood of a pattern response **b** in the trial on each of 180 integer stimulus values (i.e., 1-180°) and chose the stimulus that yielded maximum likelihood value. It is worth noting that, in contrast to classification, the MLE method did not involve any model training and the estimation was directly performed based on the generative models. We summed the mean circular square errors (MSE_{circ}) between the estimated stimuli and the true stimuli across all trials:

$$MSE_{circ} = \frac{1}{n}\sum_{i=1}^{n}(\hat{s}_i - s_i)^2\,, \tag{15}$$

We took the inverse of the MSE_{circ} as the estimation efficiency (see Fig. 4&5). A higher estimation efficiency value indicates more accurate estimations.

In the neuron-encoding model, we randomly sampled 1000 stimuli (i.e., 1000 trials) from (0,180°] for each combination of pool sizes and c_{neuron} values, and estimated the stimulus (see Methods below) in each trial.

In the voxel-encoding model, we performed 10 independent simulations and recreated the linear weight **W** in each pool size and each simulation but kept the same **W** across c_{vxs} values. We also resampled a new set of 1000 stimuli in each simulation.

Simultaneously manipulating neuronal and voxelwise NCs

In previous simulations, we only either manipulated the neuronal NCs in the neuron-encoding model or the voxelwise NCs in the voxel-encoding model. The voxel-encoding model contains both the neuronal and voxel activity stages. To further examine the interaction effect of two levels of NCs, we simultaneously manipulate both neuronal and voxelwise NCs in the voxel-encoding model (Fig. 6). In this simulation, we kept similar

settings as the above simulations in the voxel-encoding model except the following changes. First, we used 200 voxels and did not vary the voxel pool size. Second, in every trial of classification and estimation, we first generated a neuronal population response and linearly transformed it into a voxel population response.

Calculation of information

We calculated Fisher information in the stimulus-estimation task as it is one of the standard methods to quantify information in computational neuroscience and it avoids the dependency on an arbitrary decoder (Brunel and Nadal, 1998). Specifically, we use linear Fisher information, which is an approlinearon of Fisher information, and which can be expressed as:

$$I(s) = \mathbf{f}'(s)^T * \mathbf{Q}^{-1}(s) * \mathbf{f}'(s), \tag{16}$$

where $\mathbf{f}'(s)$ is the derivative of the mean population responses with respect to stimulus s and $\mathbf{Q}$ (i.e., $\mathbf{Q}_{\text{neuron}}$ in Eq. 6 or $\mathbf{Q}_{\text{vxs}}$ in Eq. 10) is the covariance matrix given stimulus s. Previous studies have shown it is a good approximation (Moreno-Bote et al., 2014; Kohn et al., 2016). Note that linear Fisher information can be calculated from both simulated neuron- and voxel-based models as long as the tuning curves and the covariance matrix are known. Notably, in neuronal data, complete Fisher information is stimulus-dependent because of the assumed Poisson noise distribution and the covariance matrix $\mathbf{Q}$ varies across stimuli. In the simulated voxel data, we assumed additive Gaussian noise and thus the covariance matrix $\mathbf{Q}$ is identical for all orientations (i.e., stimulus-invariant) and thus linear Fisher information is equal to complete Fisher information. In this paper, we simplify them as "Fisher information" or "information". We computed the averaged linear Fisher information for all 180 discrete orientations:

$$I = \frac{1}{180}\sum_{\theta=1}^{180} I(\theta), \tag{17}$$

Note that theoretically linear Fisher information above only applies to the stimulus-estimation task or a fine-discrimination task. For a general classification task (i.e., classify two stimuli s_1 and s_2), especially for two quite different stimuli, Eq.16 can be rewritten as the discrete format:

$$I = (\Delta\mathbf{f})^T \mathbf{Q}^{-1} \Delta\mathbf{f} \ , \tag{18}$$

where $\Delta\mathbf{f} = \mathbf{f}(s_1) - \mathbf{f}(s_2)$ is the population response difference for two stimuli and $\mathbf{Q} = \frac{\mathbf{Q}(s_1) + \mathbf{Q}(s_2)}{2}$ is the covariance. This metric by definition is not Fisher information and typically called "linear discriminability" (Lin et al., 2015). To avoid confusion in terminologies, we denoted both metrics as "information" in this paper as they indicate the quality of population codes in the two tasks respectively. In the main text, we only show the Fisher information in the stimulus-estimation task (Fig. 6).

Tuning heterogeneity as a mediator for the effect of TCNC

Sompolinsky et al. (2001) have proved that the effect of TCNC on population codes roughly depends on the high-frequency components of the variance-normalized derivative of mean population response, which can be defined as:

$$m_k = \frac{\mathbf{f}'(s)}{\boldsymbol{\tau}(s)} \ , \tag{19}$$

where $\mathbf{f}'(s)$ is the derivative of the mean population responses and $\tau(s)$ is a vector of response variance of the population with respect to stimulus s. Given additive noise (i.e., with stimulus independent covariance matrix) in our simulated voxel response data, the Fisher information can be rewritten as:

$$I(s) = \frac{1}{n} \sum_{k=0}^{n-1} \frac{|\bar{m}_k|^2}{\tilde{r}_k} \ , \tag{20}$$

where n is the pool size (i.e., the number of units), $\bar{\boldsymbol{m}}$ is the discrete Fourier transform of $\boldsymbol{m}$ defined in Eq. 19， $\tilde{\boldsymbol{r}}$ are the eigenvalues of the covariance matrix $\mathbf{Q}$. Note that the key term here is $\bar{m}_k$. Combining Eqs. 19&20, we can see that, given the same covariance matrix (i.e., $\tilde{r}_k$ and $\tau(s)$), the amount of information increases as the spectral amplitudes (i.e., $\bar{m}_k$) of individual neurons increase. Intuitively, the more heterogeneous the tuning curves are, the overall population response towards a stimulus will have larger powers

across frequency bands (Fourier transform of m_k), resulting in an overall larger amount of information.

Varying voxel tuning heterogeneity

To illustrate the effect of tuning heterogeneity, we performed an additional analysis (Fig. 8). In this analysis, we kept almost all settings in our previous simulations except two major changes listed as follows. First, we designed the heterogeneity coefficient (c_{homo}) that controls the tuning heterogeneity. The key to manipulating heterogeneity is to adjust the linear weighting from neuronal to voxel responses. For each voxel, we first randomly selected one neuron from all 180 neurons and assigned the c_{homo} as the linear weight for this neuron. The weights for other neurons were then assigned by a number between 0~1 scaled by (1- c_{homo}) (i.e., (1- c_{homo})*rand in Matlab). In this case, if c_{homo} = 1, the voxel tuning curve is homogeneous and identical to the neuronal tuning curve chosen in the first step; if c_{homo} = 0, the tuning curve is heterogeneous as it is the linear combination of all other neurons with random weights (see Fig. 8C). Second, to make our analysis more comparable to the neuronal simulation, we normalized the range of tuning curves between [1, 20], as the same for the neural tuning curves (see Eq. 1). Note that in this case voxel response amplitude is larger than that in the previous voxel simulation (<10). To compensate for this amplitude change, we also scaled the response variance (i.e., u in Eq. 12) of all voxels.

454 **RESULTS**

455 **1. Effects of noise correlation in neuronal and voxel populations on both stimulus**
456 **estimation and classification**

457 In the first part, we will show the effects of noise correlation on population codes in both
458 a neuronal and a voxel population. We simulated multivariate responses with three and
459 two forms of noise correlation in a neuronal population and a voxel population
460 respectively. We performed two routinely used brain decoding tasks—a stimuli-
461 classification task and a stimulus-estimation task. In the stimulus-classification task, we
462 used a linear classifier to categorize evoked population responses into one of two discrete
463 orientation stimuli. In the stimulus-estimation task, we trained a maximum likelihood
464 estimator (MLE) to reconstruct the continuous orientation value based on the population
465 response in a trial. Both tasks are two routinely used forms of MVPA in the literature
466 (Kamitani and Tong, 2005; Tong and Pratte, 2012; Haxby et al., 2014).

467

468 ***1.1 TCNCs between neurons impair decoding accuracy in multivariate neuronal***
469 ***responses***

470 Before examining the effect of TCNC in a voxel population, we first attempted to
471 replicate the classical findings in a standard neuronal population. In the simulation of
472 neuronal population responses, all neurons shared the same tuning curve except that their
473 preferred orientations were equally spaced in the continuous orientation space (see
474 Methods). We manipulated three types of NCs between neurons—angular-based tuning-
475 compatible noise correlation (aTCNC), curve-based tuning-compatible noise correlation
476 (cTCNC) and shuffled noise correlation (SFNC). The first one is also called 'limited-
477 range correlation' in neurophysiological literature (Zohary et al., 1994; Cohen and Kohn,
478 2011). aTCNCs are based on the angular difference between the preferred orientations of
479 two neurons (Fig. 3A). Particularly, we defined that the strength of NC between two
480 neurons follows an exponential decay function (Eq. 2) of the absolute angular difference
481 between their preferred orientations. This approach has often been used to establish
482 population coding models (Zohary et al., 1994; Bair et al., 2001; Ecker et al., 2011). The
483 second type, cTCNC, is based on the similarity between the tuning curves (i.e., SC) of
484 two neurons (Fig. 3B). We defined that the sign and magnitude of cTCNCs are the same

as and proportional to the SC between two neurons. This is consistent with empirical measurement in electrophysiology (Zohary et al., 1994; Bair et al., 2001; Ecker et al., 2011). Note that both aTCNC and cTCNC are related to the tuning similarity between two neurons since the larger angular difference between the two neurons' preferred orientations, the less their tuning curves are correlated. SFNCs served as a control condition and were generated by randomly shuffling the cTCNCs between neurons (Fig. 3C, see Methods) such that they had no relationship with the tuning properties of neurons.

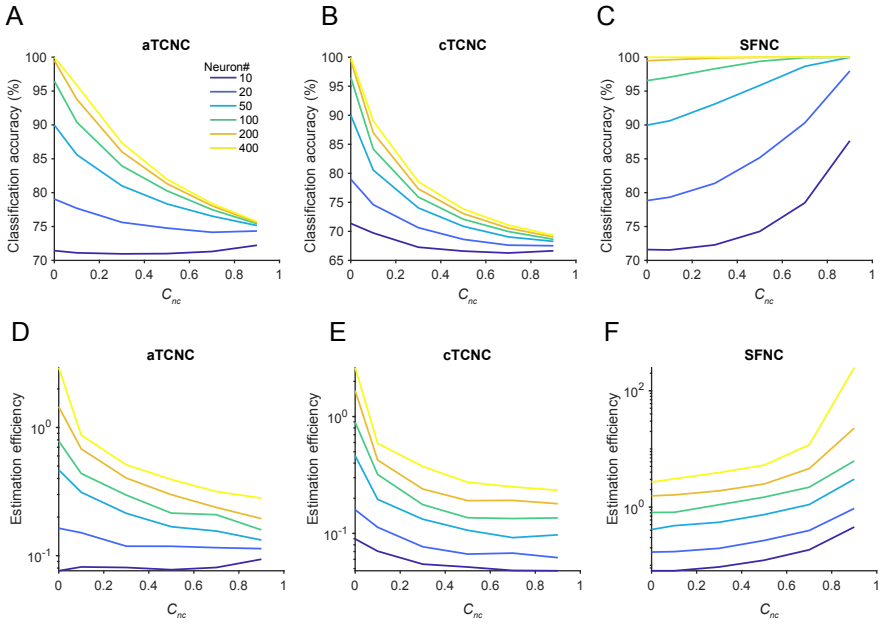

Figure 4. TCNCs impair population codes in a neuronal population. The multivariate classification accuracy (panels A-C) and maximum likelihood estimation efficiency (panels D-F) are depicted as a function of the magnitude of the aTCNC (panels A, D), TCNC (panels B, E) and the SFNC (panels C, F). Both classification accuracy and estimation efficiency decline as the strength of aTCNC and cTCNC increases. Conversely, increasing the strength of SFNC improves decoding accuracy.

We manipulated two variables of the neuronal population—the pool size (i.e., the number of neurons) and the strength of NCs between neurons (i.e., c_{nc}). For every combination of pool sizes and NC strength levels, we simulated population responses in many trials and performed the MVPA decoding (i.e., classification and estimation) on the simulated responses.

Results indeed replicated the findings from previous theoretical work (Zohary et al., 1994). aTCNCs and cTCNCs impaired decoding performance in both tasks: the classification accuracy (Fig. 4A-B) and the efficiency of the MLE (Fig. 4D-E) declined as the strength of aTCNCs and cTCNCs increased. Decoding performance always rose as the strength of SFNCs increased. This result is similar to finding in (Wilke and Eurich, 2002). We will explain this phenomenon in the later sections.

1.2 Decoding accuracy as U-shaped functions of TCNCs in multivariate voxel responses

We next turned to examine the impact of NC on population codes in fMRI data. We simulated responses of a population of voxels using a voxel-encoding model also in both classification and estimation tasks.

We again manipulated two types of NCs—cTCNC and SFNC. The cTCNCs here are similar to above except that they are between voxels rather than neurons. Similarly, cTCNCs here are defined with respect to the similarity of their orientation tuning curves. SFNCs were also generated by randomly shuffling the cTCNCs between voxels (see Methods). Note that we cannot parametrically derive aTCNCs for voxels as we did for neurons since unlike unimodal orientation tuning curves of cortical neurons, orientation tuning curves of voxels might be irregular (i.e., multi-modal) due to the unknown transformation from neuronal to a voxel's activity (see Eq. 7). We will return to this point in a later section.

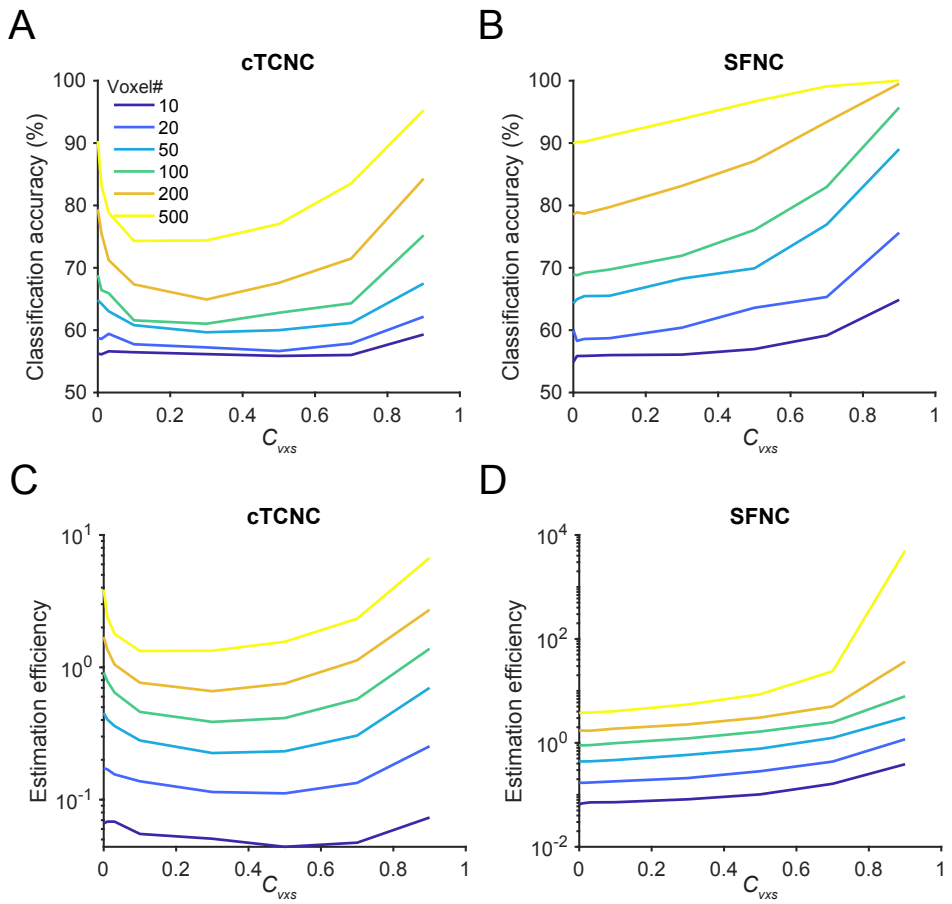

Figure 5. Decoding accuracy as U-shaped functions of cTCNCs in a voxel population. The multivariate classification accuracy (panels A, B) and estimation efficiency (panels C, D) are depicted as a function of the magnitude of cTCNCs (panels A, C) and SFNCs (panels B, D). Decoding accuracy exhibits U-shaped functions as the strength of cTCNCs increases. Similar to a neuronal population, SFNCs always improve MVPA decoding accuracy.

Surprisingly, we found that the decoding performance exhibited U-shaped functions of the increasing amount of cTCNCs: both classification accuracy (Fig. 5A) and estimation efficiency (Fig. 5C) first declined and then rose in both tasks. This is puzzling since the predominant view in neurophysiology regards cTCNCs as detrimental but here we demonstrated that to some extent cTCNCs also improve population codes. SFNCs in general improved the decoding accuracy, similar to the effect observed in a neuronal population.

1.3 Simultaneously varying neuronal and voxelwise noise correlations

In empirical fMRI studies, we can only measure the voxelwise NC but the sources of these NCs are unclear. One important source might be the neuronal NC because the neuronal NC could propagate to the voxel level if voxel responses are believed to be the aggregation of neuronal responses. fMRI data might also involve other MRI specific noise (e.g., hemodynamic fluctuations, thermal noise, and head motion). It is thus reasonable to assume that voxel-level NCs reflect the combinations of the neuronal and other voxel-level NCs. Systematically disentangling these factors is intractable but we can derive the theoretical bounds using our analytical framework here. In previous analyses, we only manipulated either the NCs between neurons in a neuronal population or the NCs between voxels in a voxel population. Here we manipulate late both neuronal and voxelwise TCNCs in the voxel-encoding model.

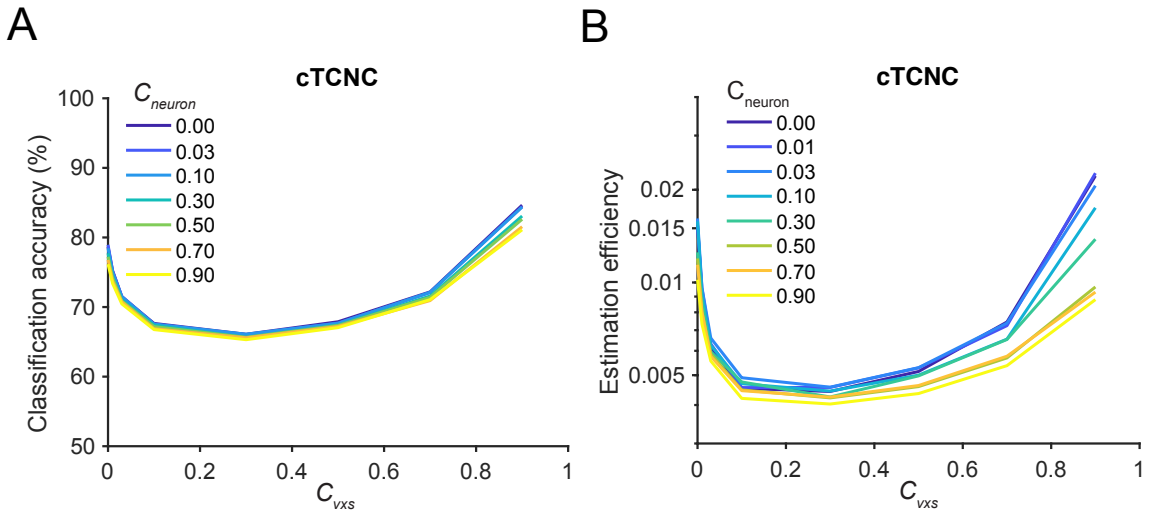

Figure 6. The impacts of neuronal and voxelwise cTCNCs on stimulus classification (A) and estimation (B). In both tasks, decoding performance exhibits U-shaped functions of the strength of voxelwise cTCNCs. Neuronal cTCNCs have small impacts on classification accuracy because the voxel-level noise primarily limits information. Neuronal cTCNCs have a more prominent detrimental effect in the estimation task. These results are consistent with the results when two levels of cTCNCs are manipulated independently.

We repeated the classification and the estimation task on a voxel population (see Methods for details). Results showed that increasing neuronal-level cTCNCs had a small impact on classification accuracy and the change in the classification accuracy values was primarily determined by voxel-level cTCNCs. This is because the classification problem (s_1 = 70°, s_2 = 110°) is an easy coarse discrimination task given relatively good signal-to-

noise level at the neuron level. The classification accuracy here is primarily bottlenecked by the noise at voxel level thus even though neuronal cTCNCs attenuate the information at the neuron level, they have relatively small impacts on overall classification accuracy.

In the estimation task, the neuronal cTCNCs dampened estimation efficiency and the voxelwise cTCNCs impacts estimation efficiency as U-shaped functions, both results consistent with previous analysis of manipulating two levels of NCs independently. The effect of neuronal cTCNCs is more prominent in the estimation task because stimulus estimation requires precise resolution of the local signal, which is only related to the tuning curves and cannot be adjusted by task difficulty.

2. Results of information-theoretical analyses explain the effects of noise correlation on population codes

In the second part, we will show how to use information-theoretical analyses to explain the simulation results above. Especially, we want to highlight the unit tuning heterogeneity as a mediator for the effect of noise correlation in a population. Unit tuning heterogeneity also acts as the key factor to explain the differential effects of TCNC in a neuronal and a voxel population.

2.1 Amount of information echoes decoding accuracy in population codes

Above analyses focused on assessing the population codes from the decoding perspective (i.e., MVPA), the approach that almost all previous fMRI decoding studies used. However, merely looking at the accuracy of an arbitrary decoder might be insufficient since decoding performance critically depends on the decoder. For example, a decoder might have bias or wrong assumptions about data structure. Here, we propose an alternative approach—directly calculate the amount of Fisher information for the stimulus-estimation task and linear discriminability for the stimulus-classification task, which is also the standard metric for information coding in computational neuroscience (Brunel and Nadal, 1998). For the estimation task, Fisher information indicates the low bound of decoding variance that any unbiased decoder can achieve and thus does not depend on a particular form of a decoder. For the classification task, linear

discriminability measures the magnitude of separation of two multivariate response distributions. It is also called a variant of linear Fisher information for a classification task (Kohn et al., 2016). Here, we define both metrics as "information" for corresponding tasks as they both indicate the quality of population codes. The analysis of information has two major advantages over the conventional MVPA approach. First, the assessment of information coding is not related to the assumptions or efficacy of a particular decoder. This is particularly useful for the stimulus-estimation task because an appropriate decoder might be hard to derive (e.g., the typical regression approach is inappropriate) given the continuous and circular nature of orientation. Second, in theory more information in a population usually indicates a higher decoding accuracy. But their relationship is nonlinear. Classification accuracy can reach the floor (e.g., 50% for binary classification) and ceiling (i.e., 100%) but the amount of information has a relatively broad range thus more sensitive to population codes.

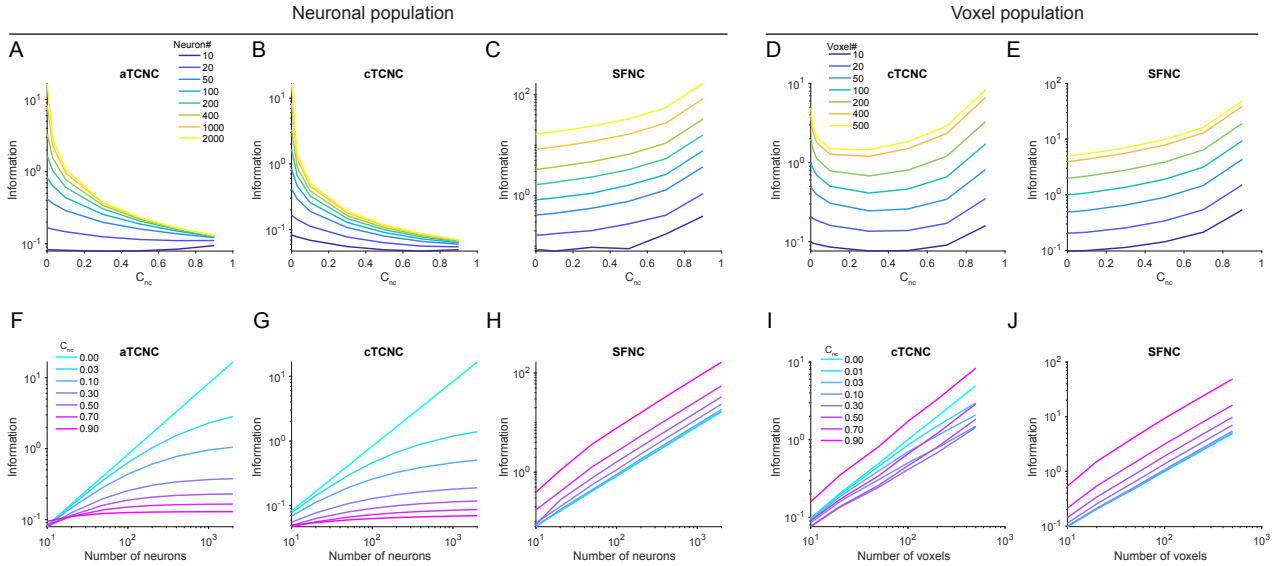

Figure 7. Fisher information in neuronal and voxel populations with diverse types and strength of NCs. The upper and the bottom rows depict the amount of information as a function of increasing strength of NCs and the increasing number of units in the population, respectively. Panels A-C and F-H illustrate the amount of information in a neuronal population and correspond to Fig. 4. Panels D-E and I-J illustrate the amount of information in a voxel population and correspond to Fig. 5. Note that here we only plot the information in a stimulus-estimation task. We consider three types of NC—aTCNC (panels A, F), cTCNC (panels B, G) and SFNCs (panels C, H) in the neuronal population, as already shown in Fig. 4. Similar treatments are performed for the voxel population, as shown in Fig. 5. The calculation of

information largely mirrors the decoding results shown in Fig. 4 and Fig. 5. Critically, the amount of information in the voxel population exhibits U-shaped functions of increasing strength of cTCNCs (panel D) and cTCNCs do not limit information as the number of voxels increases (panel I). These results drastically differ from the effect of aTCNCs (panels A, F) and cTCNCs (panels B, G) in a neuronal population.

We calculated Fisher information in both the neuronal and the voxel populations (see Methods) as functions of the pool size and NC strength. Results largely mirrored the previous decoding results. In the neuronal population, we replicated the key signatures of detrimental effects of TCNCs: the amount of information saturated as the pool size increased given the presence of aTCNCs (Fig. 6F) and cTCNCs (Fig. 6G) but not SFNCs (Fig. 6H). Also, the amount of information declined as the magnitude of aTCNCs (Fig. 6A) and cTCNCs increased (Fig. 6B). This pattern was reversed as the magnitude of SFNCs increased (Fig. 6C). In the voxel population, the amount of information always increased as the pool size expanded in both cTCNCs (Fig. 6I) and SFNCs (Fig. 6J) conditions. Similar to the decoding results, the amount of information exhibited U-shaped functions as the magnitude of cTCNCs (Fig. 6D) increased and always grew as the magnitude of SFNCs (Fig. 6E) increased.

The effect of NC on the capacity of the neural population code have been investigated in various studies over the past two decades (e.g., Abbott and Dayan, 1999; Sompolinsky et al., 2001; Shamir and Sompolinsky, 2006; Ecker et al., 2011), leading to somewhat mixed results. Early results in neurophysiology suggest that cTCNCs could be detrimental (Zohary et al., 1994), but later studies suggest that the results may be more complicated, depending on the detailed of the configurations of the neural codes. There are regimes where the cTCNCs could be beneficial (Wilke and Eurich (2002), Shamir and Sompolinsky, 2006; Ecker et al., 2011). Wilke and Eurich (2002) found that adding jitter to a uniform positive correlation structure would benefit the code. They further provided an intuitive argument on why variability in the noise correlation might increase the coding capacity. Their results are consistent with our simulation results on the benefit of SFNCs.

We proposed that Fisher information could provide potential explanations of the effect of different kinds of NCs on estimation task and fine discrimination task. As

659 described earlier, Fisher information provides a bound on the estimation error.
660 Furthermore, it also provides a bound for discrimination threshold (Seung and
661 Sompolinsky, 1993), which determines the performance of the fine discrimination task
662 we tested. Indeed, we found that the results on Fisher information echo well with the
663 performance on the two tasks under various manipulations on the Fisher information,
664 confirming our proposal.
665
666 ***2.2 Voxel tuning heterogeneity and pool sizes explain the effect of TCNC on***
667 ***population codes***
668 Why TCNCs manifest differently in neuronal and voxel responses? We reason that the
669 neuron-to voxel transformation (i.e., linear weighting **W**) might be the key factor that
670 alters the effect of TCNCs. Unlike the homogeneous neuronal tuning curves (i.e., same
671 width, amplitude, and baseline), voxel tuning curves might be multi-modal or have
672 diverse tuning widths and amplitudes (Fig. 8C). This is due to the unknown distribution
673 (i.e., the weighting matrix **W** in Eq. 7) of orientation-selective neurons within a voxel.
674 Even though individual neurons follow a uniform bell-shape tuning property, the
675 aggregation of them can produce tuning functions with diverse forms. Because of the
676 tuning heterogeneity, TCNCs do not limit information anymore (see mathematical
677 derivation in Methods). The effect of tuning heterogeneity has been suggested in (Wilke
678 and Eurich, 2002; Shamir and Sompolinsky, 2006; Ecker et al., 2011). This assumption
679 of homogeneous neuronal tuning curves is apparently not realistic because in the primate
680 brain it has been known that the shape of tuning curves varies drastically across neurons.
681 Such tuning heterogeneity removes TCNCs' limitation on information. In a large
682 population, TCNCs even always benefit the population codes (Fig. 8B). This theoretical
683 implication has been also corroborated by an empirical study on orientation decoding in
684 primate V1 (Graf et al., 2011). Most importantly, the principle of tuning heterogeneity
685 applies both neuronal and voxel populations.

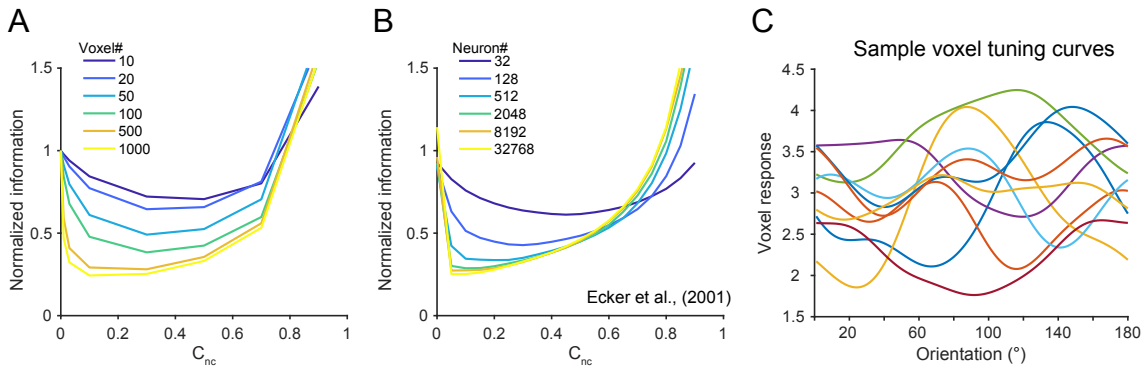

Figure 8. Tuning heterogeneity might be one of the major mediators for the effect of TCNC in a voxel population. In panels A, y-axis indicates the absolute amount of information is normalized by the amount of information when NCs are 0. Normalized information in our simulation is directly reminiscent of the information on a neuronal population with heterogeneity tuning curves (panel B, reproduced from (Ecker et al., 2011)). This result implies that tuning heterogeneity mediates the effect of TCNC in both types of data. Panel C illustrates some sample tuning curves of the simulated voxels. Due to the uncertain neuron-to-voxel connections (i.e., linear weight matrix **W**), the endowed voxel tuning curves also exhibit irregular forms.

To further substantiate the interactive effect between tuning heterogeneity and NCs on population codes, we performed two additional analyses. First, in the voxel population (Fig, 7A), we normalized the information when cTCNCs are present (i.e., NC > 0) by the information when cTCNCs are absent (i.e., NC = 0). We found a similar pattern as in a heterogeneous neuronal population (Fig. 8B). Importantly, the relative benefits of TCNCs are all U-shaped functions. Such beneficial effects are more pronounced in a large population (i.e., the lowest point of the function shifts to left as the pool size increases in Fig. 8A).

In the second analysis, we manipulated the degree of voxel tuning heterogeneity and the strength of cTCNCs in the voxel population. The amount of information was calculated as a function of these two variables (Fig. 9A&B). Results showed that the amount of information follows U-shaped functions if voxel tuning is highly heterogeneous (i.e., c_{homo}=0 in Fig. 9). However, as the voxel tuning becomes progressively homogeneous, TCNCs always decrease information, which is consistent with the results obtained in a standard neuronal population (Fig. 7B).

Ecker et al. (2011) derived a mathematical foundation for the effects of tuning heterogeneity, pool size, and TCNC on population codes, built up the earlier work by Sompolinsky et al. (2001). Briefly, tuning heterogeneity poses more energy on the high-frequency components of the variance-normalized derivative of mean population response (see Eq. 20). The more energy on high-frequency components, the more prominent the beneficial effect of TCNC will be. Here, we extend previous work and demonstrate several novel aspects of NC in a population. First, previous theoretical work in neurophysiology primarily focused on the estimation task (but see (Moreno-Bote et al., 2014)) while the majority of neuroimaging research focused on the classification task. We compared both tasks in both populations. Second, previous work only analyzed one type of TCNC (i.e., aTCNC in our case) and we systematically compared three types of NC. Third, we manipulated cTCNC at both neuronal and voxel activity stages to simulate more realistic interaction between neuronal and fMRI responses.

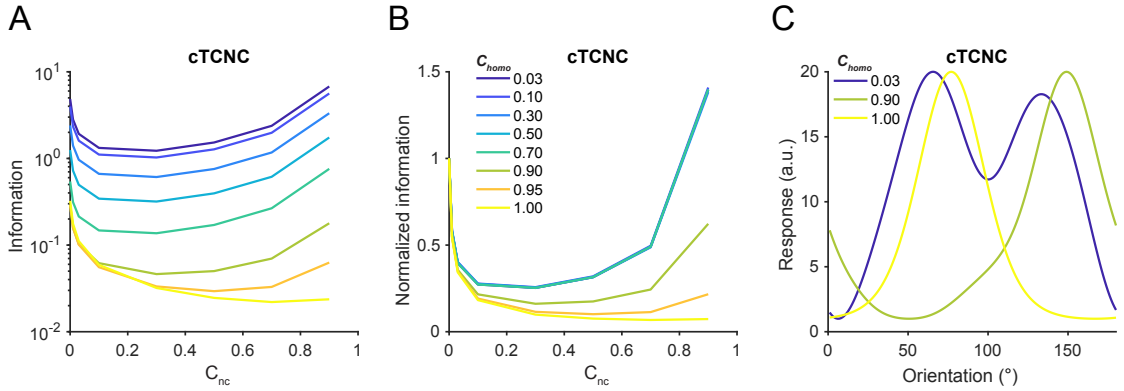

Figure 9. Interaction effect between cTCNC and tuning heterogeneity on population codes. 500 voxels were simulated (see Methods). A larger value of heterogeneity coefficient c_{homo} indicates more homogeneous voxel tuning curves. Note that the simulated voxel tuning curves are identical to neuronal tuning curves when c_{homo} =1. Panel A and B illustrate the raw and normalized amount of information as a function of TCNC under different tuning heterogeneity levels. The raw information is normalized in the same way as in Fig. 8. As voxel tuning heterogeneity increases, the shape of the functions changes from U-shaped to monotonically decreasing. Panel C illustrates sample tuning curves with different heterogeneity levels.

DISCUSSION

Characterizing the effect of noise correlation on population codes has attracted vast research attention in the past years as it is related to several key topics (e.g., probabilistic computation, uncertainty) in neuroscience research. But the majority of relevant studies are confined to the field of neurophysiology. On the other hand, fMRI can measure many responsive units in the brain but most prior fMRI studies only employed MVPA to evaluate population codes. MVPA accuracy, however, is merely a coarse description of population codes and the precise quantitative relationship between the voxelwise NCs and population codes still remains unclear. Here, we conducted a series of theoretical analyses to systematically examine how NCs with different forms and strength level influences MVPA accuracy and amount of information on multivariate fMRI responses. We made three major observations: (1) decoding accuracy and the amount of information follow U-shaped functions of TCNCs in a voxel population and this effect is mediated by voxel tuning heterogeneity and pool sizes; (2) we also found that even randomly injecting some NCs between voxels improves population codes and this effect can be intuitively understood through a simple winner-take-all principle of decoding; (3) furthermore, the comparisons against a standard neuronal population demonstrate that the effect of NC in both neuronal and voxel populations can be understood within a unified computational framework related to tuning heterogeneity and pool sizes. Our work extends the research on noise correlations in neurophysiology to fMRI. These results suggest that conventional MVPA is insufficient for fully characterizing the details of population codes and it is better to uncover the full distribution of multivariate responses in fMRI data. The method proposed here can be used in future fMRI research.

Implications for sensory processing

Though well-established in empirical experiments, how NCs contribute to in perceptual or cognitive processes? One important function of NC might be to represent sensory uncertainty (Pouget et al., 2013; Ma and Jazayeri, 2014).

The behavioral signatures of sensory uncertainty have been shown in different perceptual and cognitive domains (Ernst and Banks, 2002; Dayan and Daw, 2008; Barthelme and Mamassian, 2010; Gifford et al., 2014). From the perspective of Bayesian

decision theory, a perceptual decision is not merely a point estimation but is associated with a full probabilistic distribution of possible options computed in the brain (Pouget et al., 2013). One class of theories propose that the internal distribution can be approximated by a Monte Carlo sampling process and the trial-by-trial or moment-by-moment response fluctuations might reflect samples from the distribution (Hoyer and Hyvarinen, 2003; Fiser et al., 2010; Shivkumar et al., 2018). This view has been corroborated by the findings that neuronal NCs can be captured by stochastic sampling processes from the posterior belief of task structures or stimuli (Haefner et al., 2016; Orban et al., 2016). In the field of neuroimaging, a recent fMRI study directly estimated sensory uncertainty in the human brain based on a voxel-encoding model and demonstrated that the decoded uncertainty is indeed correlated with behavioral variability (van Bergen et al., 2015). These results suggest that characterizing the effects of the noise correlation might hold the key to uncover the inference process in the brain.

Towards a generative understanding of multivariate fMRI responses

In contrast to the enthusiasms in characterizing generative processes of stimulus-evoked responses in neurophysiology, the generative modeling of fMRI data has still been lacking and one of the key missing components is the response covariance between voxels. Many imaging scientists only focus on using MVPA to decode information from fMRI data (Tong and Pratte, 2012; Haxby et al., 2014). In recent years people have increasingly realized the limitations of MVPA as a discriminative modeling approach, calculating the probability p(stimulus | response). Rich representational information can be easily buried by merely examining decoding accuracy (Naselaris and Kay, 2015).

From a probabilistic modeling approach, understanding the generative computation in the brain is equivalent to calculation of the joint probability p(response, stimulus), equal to p(response | stimulus) $\times$ p(stimulus) according to Bayes' rule. Existing voxel-encoding modeling tackles the mean of the first term p(response | stimulus) as it describes how a stimulus can go through a series of computations to produce population responses in the brain. However, the likelihood function p(response | stimulus) also requires the covariance part. Thus, to work towards a full generative understanding of

802 multivariate voxel responses, we need to combine standard voxel-encoding models and
803 characterizations of the covariance of response distributions.
804
805 ***Quantifying information in fMRI data***
806 In this paper, we focused on two routinely used perceptual tasks—the stimulus-estimation
807 task and the stimulus-classification task. Stimulus estimation is equivalent to a very fine-
808 discrimination task as it needs to discriminate the true stimulus value from nearby stimuli
809 in the feature pace. It is primarily determined by Fisher information. Binary classification
810 is similar to coarse discrimination as it depends on the distance of two stimuli in
811 representational space. In this case, linear discriminability is a better measure than Fisher
812 information (Newsome et al., 1989; Britten et al., 1992; Butts and Goldman, 2006; Lin et
813 al., 2015).
814 For both tasks, we propose an alternative method instead of the conventional
815 MVPA approach to assess the accuracy of population codes. In our analysis, we directly
816 calculate the amount of information and this method can be applied to both neuronal and
817 voxel populations. Our work presents a good example to apply this approach in future
818 fMRI research. For example, in a typical binary classification task, we usually measure
819 many trials of population responses towards each of two stimuli. There must be some
820 NCs among voxels but whether they impair or improve population codes is hard to
821 delineate as we have shown that it depends on several factors such as the strength, form
822 of NCs and the pool size. We can use Eq. 18 to tackle this issue. The response difference
823 and covariance matrices can be directly estimated from empirical data. We can therefore
824 use these variables to directly quantify the amount of information. To illustrate the
825 contribution of NCs, we can compare two conditions, one condition in which estimated
826 response difference, individual response variance, and voxelwise NC are used to
827 calculate the information, the other condition in which the estimated response difference
828 and individual response variance are kept same but the NCs are deliberately removed. We
829 can conclude that these NCs are detrimental if the removal of NCs elevates the amount of
830 information or vice versa.
831

Winner-take-all principle as an intuitive explanation for the effect of SFNC on population codes

Most of the analyses in the result part mainly focus on TCNCs. We next turned to another interesting finding—SFNCs improve population codes in both the voxel and the neuronal populations. At first glance, this might seem surprising since it suggests that decoding accuracy can be improved by randomly injecting some NCs between units. Although the analyses of Fisher information provide a formal explanation for this phenomenon—decoding accuracy increases because the amount of information in the population increases, here we want to highlight a more intuitive explanation—a winner-take-all principle enhances decoding accuracy in the conventional multivariate analysis.

We created a simple three-voxel model to illustrate this effect (Fig. 10). The NC between voxels X and Y improves classification (Fig. 10A), while the NC between voxels Y and Z impairs classification (Fig. 10B). However, when all three voxels are aggregated, the contributions of two NCs do not cancel out each other and the overall decoding performance is still improved by the NC between X and Y, regardless of the NC between Y and Z. Importantly, classification accuracy on X, Y, and Z with NCs is higher than the condition that three units activate independently. This result demonstrated that, as long as there exists at least one pair of "good" voxels whose NC is beneficial, a linear classifier can utilize the NC to achieve a good decoding performance. In other words, the most informative units determine the lower bound of decoding accuracy (Fig. 10).

This winner-take-all principle demonstrates an intuitive explanation for the effects of SFNCs on decoding. SFNCs are generated by randomly shuffling the TCNCs between voxels such that the NCs between some pairs of units might be beneficial whereas in other cases might be detrimental when considered individually. However, the overall decoding accuracy is related to the most beneficial NCs. As the pool size increases, it is more likely that a small portion of units can be assigned with NCs that greatly benefit decoding, resulting in an overall enhanced decoding accuracy on the entire population.

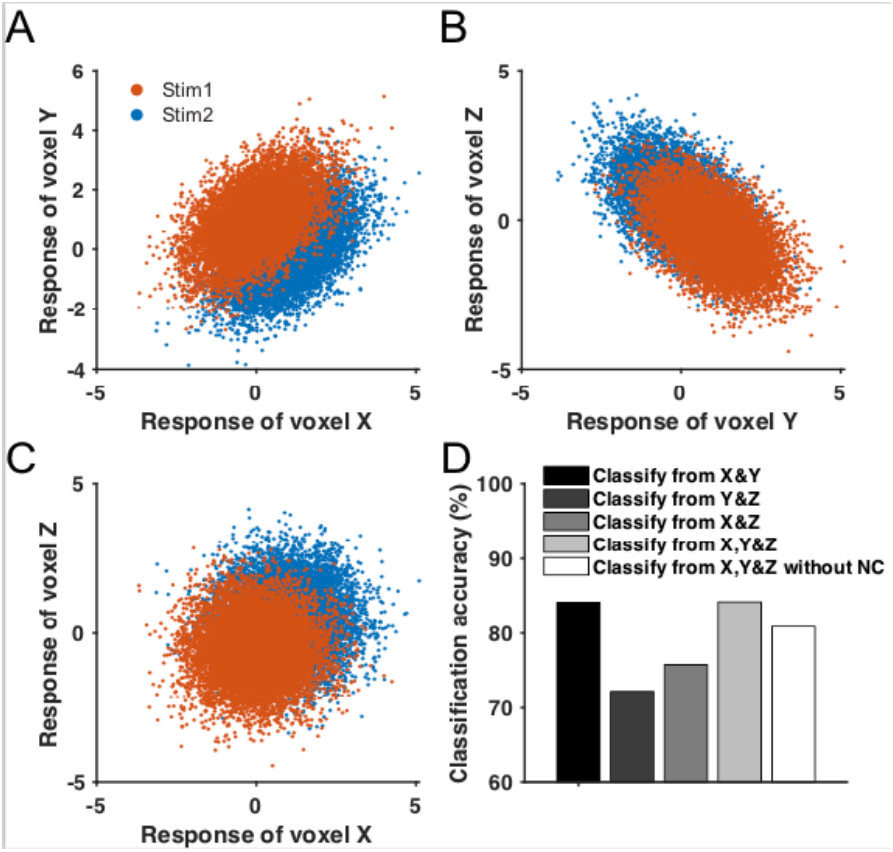

Figure 10. A three-voxel model illustrating the winner-take-all principle of multivariate decoding. Panel A illustrates the trial-by-trial responses of voxels X and Y towards two stimuli. The covariance structure of X and Y enhances classification accuracy. Similarly, panel B illustrates that the covariance structure of voxels Y and Z impairs classification. Voxels X and Z have no systematic NC (panel C). Panel D depicts the classification accuracy based on population responses of X and Y (panel A), Y and Z, X and Z, all three units with and without NCs, respectively. The beneficial and detrimental effects of the covariance structures in panels A and B do not cancel each other if all three voxels are combined. The decoding performance of voxels X and Y determines the classification accuracy based on all three voxels (i.e., winner-take-all principle).

The nature of noise correlations in fMRI data

Although fMRI can naturally measure the activity of many units, the investigation of NCs in fMRI data has just begun recently. Exploring this issue in fMRI data is, however, non-trivial and we summarize the related issues as follows.

First, the definition of "noise correlation" in fMRI research is still under debate. The well-accepted definition of "noise correlation" in computational neuroscience is the

correlation of trial-by-trial responses between two neurons given the repeated presentation of the same stimulus. This definition emphasizes stimulus-evoked responses. In this paper, we strictly follow this definition and assume voxel responses as trial-by-trial responses estimated from the standard general linear modeling. However, one recent study defined the noise correlation between two voxels as their resting-state functional connectivity or background functional connectivity during a task (Bejjanki et al., 2017). In theory, these definitions deviate from the conventional definition in the field of neurophysiology. Future studies might need to examine the relations between resting-, task-based functional connectivity and trial-by-trial variation of voxel responses.

Second, the sources of noise correlations in fMRI data are still unclear. On one hand, the conventional term "noise correlation" itself is somehow misleading since, according to the results in this paper, response variability also contains a substantial amount of stimulus information. In other words, response variability is not purely "noise" and it might reflect some critical aspects of stimulus processing or task structure (Bondy et al., 2018). On the other hand, empirically measured voxel response variability also contains all other non-sensory noises, such as thermal noise, head motion etc. These forms of noise are irrelevant to information processing even though they could affect decoding. This raises a theoretical concern that we can examine the population codes at the macroscopic level using voxel responses, an approach most fMRI studies employed. However, it is challenging to use the macroscopic voxel responses to infer the tuning and NCs at the microscopic neuron level since our linear pooling model here is an oversimplification of real neural processing in the brain. We still lack the precise quantitative relationship between the neuronal and voxel responses. Relevant issues presumably also exist in neurophysiological data. Disentangling sensory and non-sensory noises still remains a challenge with which both fMRI and neurophysiological researchers are confronted.

Third, unlike the established empirical measurements of NC in neurophysiological data, how NCs related to the tuning similarity of voxels still needs to be quantified. A recent study by (Ryu and Lee, 2018) symmetrically evaluated the impact of three factors–retinotopic distance, cortical distance, and tuning similarity—on NCs in early visual cortex and found that the tuning similarity is the major determinant. However,

this study defined noise correlation as resting-state connectivity. van Bergen and Jehee (2018) established a voxel-encoding model similar to our work here and evaluated the relationship between noise and signal correlations. Similar to the well-established finding of TCNC in neurophysiology, they showed that the magnitude of voxelwise NCs in early visual cortices (V1 to V3) is also a monotonically positive function of voxelwise SCs, albeit nonlinear. Interestingly, the magnitude of NCs in V1 is generally higher than the NCs in V2 and V3. This provides evidence supporting the existence of TCNC in voxel response. Future studies are needed to investigate this issue in more brain areas, feature dimensions, and tasks.

Acknowledgments:
The work was supported by NIH Grants P41 EB015894, P30 NS076408, S10 RR026783, S10 OD017974-01, and the W.M. Keck Foundation.

Author contributions: R.Z. conceived, designed and performed research; R.Z. wrote the draft of the paper. R.Z., X.W. and K.K. edited the paper.

Conflict of Interest: The authors declare no competing financial interests.

REFERENCES

Averbeck BB, Lee D (2003) Neural noise and movement-related codes in the macaque supplementary motor area. J Neurosci 23:7630-7641.

Averbeck BB, Latham PE, Pouget A (2006) Neural correlations, population coding and computation. Nat Rev Neurosci 7:358-366.

Bair W, Zohary E, Newsome WT (2001) Correlated firing in macaque visual area MT: time scales and relationship to behavior. J Neurosci 21:1676-1697.

Barthelme S, Mamassian P (2010) Flexible mechanisms underlie the evaluation of visual confidence. Proc Natl Acad Sci U S A 107:20834-20839.

Bejjanki VR, da Silveira RA, Cohen JD, Turk-Browne NB (2017) Noise correlations in the human brain and their impact on pattern classification. PLoS Comput Biol 13:e1005674.

Bondy AG, Haefner RM, Cumming BG (2018) Feedback determines the structure of correlated variability in primary visual cortex. Nat Neurosci 21:598-606.

Britten KH, Shadlen MN, Newsome WT, Movshon JA (1992) The analysis of visual motion: a comparison of neuronal and psychophysical performance. J Neurosci 12:4745-4765.

Brunel N, Nadal JP (1998) Mutual information, Fisher information, and population coding. Neural Comput 10:1731-1757.

Butts DA, Goldman MS (2006) Tuning curves, neuronal variability, and sensory coding. PLoS Biol 4:e92.

Cohen MR, Kohn A (2011) Measuring and interpreting neuronal correlations. Nat Neurosci 14:811-819.

Constantinidis C, Goldman-Rakic PS (2002) Correlated discharges among putative pyramidal neurons and interneurons in the primate prefrontal cortex. J Neurophysiol 88:3487-3497.

Cox DD, Savoy RL (2003) Functional magnetic resonance imaging (fMRI)"brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage 19:261-270.

Dayan P, Daw ND (2008) Decision theory, reinforcement learning, and the brain. Cogn Affect Behav Neurosci 8:429-453.

Ecker AS, Berens P, Tolias AS, Bethge M (2011) The effect of noise correlations in populations of diversely tuned neurons. J Neurosci 31:14272-14283.

Ecker AS, Berens P, Keliris GA, Bethge M, Logothetis NK, Tolias AS (2010) Decorrelated neuronal firing in cortical microcircuits. Science 327:584-587.

Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415:429-433.

Fiser J, Berkes P, Orban G, Lengyel M (2010) Statistically optimal perception and learning: from behavior to neural representations. Trends Cogn Sci 14:119-130.

Gifford AM, Cohen YE, Stocker AA (2014) Characterizing the impact of category uncertainty on human auditory categorization behavior. PLoS Comput Biol 10:e1003715.

Graf AB, Kohn A, Jazayeri M, Movshon JA (2011) Decoding the activity of neuronal populations in macaque primary visual cortex. Nat Neurosci 14:239-245.

Gutnisky DA, Dragoi V (2008) Adaptive coding of visual information in neural populations. Nature 452:220-224.

Haefner RM, Berkes P, Fiser J (2016) Perceptual Decision-Making as Probabilistic Inference by Neural Sampling. Neuron 90:649-660.

Haefner RM, Gerwinn S, Macke JH, Bethge M (2013) Inferring decoding strategies from choice probabilities in the presence of correlated variability. Nat Neurosci 16:235-242.

Haxby JV, Connolly AC, Guntupalli JS (2014) Decoding neural representational spaces using multivariate pattern analysis. Annu Rev Neurosci 37:435-456.

Hoyer PO, Hyvarinen A (2003) Interpreting neural response variability as Monte Carlo sampling of the posterior. In: Advances in neural information processing systems, pp 293-300.

Huang X, Lisberger SG (2009) Noise correlations in cortical area MT and their potential impact on trial-by-trial variation in the direction and speed of smooth-pursuit eye movements. J Neurophysiol 101:3012-3030.

Jermakowicz WJ, Chen X, Khaytin I, Bonds AB, Casagrande VA (2009) Relationship between spontaneous and evoked spike-time correlations in primate visual cortex. J Neurophysiol 101:2279-2289.

Kamitani Y, Tong F (2005) Decoding the visual and subjective contents of the human brain. Nat Neurosci 8:679-685.

Kohn A, Coen-Cagli R, Kanitscheider I, Pouget A (2016) Correlations and Neuronal Population Information. Annu Rev Neurosci 39:237-256.

Lee D, Port NL, Kruse W, Georgopoulos AP (1998) Variability and correlated noise in the discharge of neurons in motor and parietal areas of the primate cortex. J Neurosci 18:1161-1170.

Lin IC, Okun M, Carandini M, Harris KD (2015) The Nature of Shared Cortical Variability. Neuron 87:644-656.

Ma WJ, Jazayeri M (2014) Neural coding of uncertainty and probability. Annu Rev Neurosci 37:205-220.

Moreno-Bote R, Beck J, Kanitscheider I, Pitkow X, Latham P, Pouget A (2014) Information-limiting correlations. Nat Neurosci 17:1410-1417.

Naselaris T, Kay KN (2015) Resolving Ambiguities of MVPA Using Explicit Models of Representation. Trends Cogn Sci 19:551-554.

Naselaris T, Kay KN, Nishimoto S, Gallant JL (2011) Encoding and decoding in fMRI. Neuroimage 56:400-410.

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

Orban G, Berkes P, Fiser J, Lengyel M (2016) Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex. Neuron 92:530-543.

Pouget A, Beck JM, Ma WJ, Latham PE (2013) Probabilistic brains: knowns and unknowns. Nat Neurosci 16:1170-1178.

Ryu J, Lee SH (2018) Stimulus-Tuned Structure of Correlated fMRI Activity in Human Visual Cortex. Cereb Cortex 28:693-712.

Seung HS, Sompolinsky H (1993) Simple models for reading neuronal population codes. Proc Natl Acad Sci U S A 90:10749-10753.

Shamir M, Sompolinsky H (2006) Implications of neuronal diversity on population coding. Neural Comput 18:1951-1986.

Shivkumar S, Lange RD, Chattoraj A, Haefner RM (2018) A probabilistic population code based on neural samples. arXiv.

Smith MA, Kohn A (2008) Spatial and temporal scales of neuronal correlation in primary visual cortex. J Neurosci 28:12591-12603.

Sompolinsky H, Yoon H, Kang K, Shamir M (2001) Population coding in neuronal systems with correlated noise. Phys Rev E Stat Nonlin Soft Matter Phys 64:051904.

Tong F, Pratte MS (2012) Decoding patterns of human brain activity. Annu Rev Psychol 63:483-509.

van Bergen RS, Jehee JFM (2018) Modeling correlated noise is necessary to decode uncertainty. Neuroimage 180:78-87.

van Bergen RS, Ma WJ, Pratte MS, Jehee JF (2015) Sensory uncertainty decoded from visual cortex predicts behavior. Nat Neurosci 18:1728-1730.

Vintch B, Gardner JL (2014) Cortical correlates of human motion perception biases. J Neurosci 34:2592-2604.

Wilke SD, Eurich CW (2002) Representational accuracy of stochastic neural populations. Neural Comput 14:155-189.

Zohary E, Shadlen MN, Newsome WT (1994) Correlated neuronal discharge rate and its implications for psychophysical performance. Nature 370:140-143.