Representational structure or task structure? Bias in neural representational similarity analysis and a Bayesian method for reducing bias

Ming Bo Cai[1*], Nicolas W. Schuck [1,3], Jonathan W. Pillow[1,2], Yael Niv[1,2],

1 Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544 USA
2 Department of Psychology, Princeton University, Princeton, NJ 08544 USA
3 Max Planck Research Group Neurocode, Max Planck Institute for Human Development, 14195 Berlin, Germany

* mcai@princeton.edu

Abstract

The activity of neural populations in the brains of humans and animals can exhibit vastly different spatial patterns when faced with different tasks or environmental stimuli. The degree of similarity between these neural activity patterns in response to different events is used to characterize the representational structure of cognitive states in a neural population. The dominant methods of investigating this similarity structure first estimate neural activity patterns from noisy neural imaging data using linear regression, and then examine the similarity between the estimated patterns. Here, we show that this approach introduces spurious bias structure in the resulting similarity matrix, in particular when applied to fMRI data. This problem is especially severe when the signal-to-noise ratio is low and in cases where experimental conditions cannot be fully randomized in a task. We propose Bayesian Representational Similarity Analysis (BRSA), an alternative method for computing representational similarity, in which we treat the covariance structure of neural activity patterns as a hyper-parameter in a generative model of the neural data. By marginalizing over the unknown activity patterns, we can directly estimate this covariance structure from imaging data. This method offers significant reductions in bias and allows estimation of neural representational similarity with previously unattained levels of precision at low signal-to-noise ratio. The probabilistic framework allows for jointly analyzing data from a group of participants. The method can also simultaneously estimate a signal-to-noise ratio map that shows where the learned representational structure is supported more strongly. Both this map and the learned covariance matrix can be used as a structured prior for maximum *a posteriori* estimation of neural activity patterns, which can be further used for fMRI decoding. We make our tool freely available in Brain Imaging Analysis Kit (BrainIAK).

Author summary

We show the severity of the bias introduced when performing representational similarity analysis (RSA) based on neural activity pattern estimated within imaging runs. Our Bayesian RSA method significantly reduces the bias and can learn a shared representational structure across multiple participants. We also demonstrate its extension as a new multi-class decoding tool.

Introduction

Functional magnetic resonance imaging (fMRI) measures the blood-oxygen-level-dependent (BOLD) signals [1], which rise to peak ~ 6 seconds after neuronal activity increases in a local region [2]. Because of its non-invasiveness, full-brain coverage, and relatively favorable trade-off between spatial and temporal resolution, fMRI has been a powerful tool to study the neural correlates of cognition [3–5] . In the last decade, research has moved beyond simply localizing the brain regions selectively activated by the cognitive processes and focus has been increasingly placed on the relationship between the detailed spatial patterns of neural activity and cognitive processes [6, 7].

An important tool for characterizing the functional architecture of sensory cortex is representational similarity analysis (RSA) [8]. This classic method first estimates the neural activity pattern from fMRI data recorded as participants observe a set of stimuli or experience a set of task conditions, and then calculates the similarity (e.g., by Pearson correlation) between each pair of the estimated patterns. The rationale is that if two stimuli are represented with similar codes in a brain region, the spatial patterns of neural activation in that region would be similar when processing these two stimuli.

After the similarity matrix between all pairs of activity patterns is calculated in an ROI, it can be compared against similarity matrices predicted by candidate computational models. Researchers can also convert the similarity matrix into a representational dissimilarity matrix (RDM, e.g., $1 - C$, for similarity C based on correlation) and visualize the structure of the representational space in the ROI by projecting the dissimilarity matrix to a low dimensional space [8]. Researchers might also test whether certain experimental manipulations changes the degrees of similarity between neural patterns of interest [9, 10]. To list just a few application of this method in the domain of visual neuroscience, RSA has revealed that humans and monkeys have highly similar representational structures in the inferotemporal (IT) cortex for images across various semantic categories [11]. It also revealed a continuum in the abstract representation of biological classes in human ventral object visual cortex [12] and that basic categorical structure gradually emerges through the hierarchy of visual cortex [13]. Because of the additional flexibility of exploring the structure of neural representation without building explicit computational models, RSA has also gained popularity among cognitive neuroscientists for studying more complex tasks beyond perception, such as decision making.

While RSA has been widely adopted in many fields of cognitive neuroscience, a few recent studies have revealed that the similarity structure estimated by standard RSA might be confounded by various factors. First, the calculated similarity between two neural patterns strongly depends on the time that elapsed between the two measured patterns: the closer the two patterns are in time, the more similar they are [14] [15]. Second, it was found that because different brain regions share some common time course of fluctuation independent of the stimuli being presented (intrinsic fluctuations), RDMs between regions are highly similar when calculated based on patterns of the same trials of tasks but not when they are calculated based on separate trials (thus the intrinsic fluctuation are not shared across regions). This indicates that RSA can be strongly influenced by intrinsic fluctuation [14]. Lastly, Diedrichsen et al. (2011) pointed out that the noise in the estimated activity patterns can add a diagonal component to the condition-by-condition covariance matrix of the spatial patterns. This leads to over-estimation of the variance of the neural pattern and underestimation of correlation between true patterns, and this underestimation depends on signal-to-noise ratio in each ROI, making it difficult to make comparison of RDMs between regions [16].

Recognizing the first two issues, several groups have recently suggested modifications to RSA such as calculating similarity or distance between activity patterns estimated

from separate fMRI runs [15, 17], henceforth referred to as cross-run RSA, and using a Taylor expansion to approximate and regress out the dependency of pattern similarity on the interval between events [15]. For the last issue, Diedrichsen et al. (2011) proposed modeling the condition-by-condition covariance matrix between estimated neural patterns as the sum of a diagonal component that models the contribution of noise in the estimated neural patterns to the covariance matrix and components reflecting the researcher's hypothetical representational structure in the ROI [16] ("pattern-component model"; PCM). These methods improve on traditional RSA, but are not explicitly directed at the source of the bias, and therefore only offer partial solutions.

Indeed, the severity of confounds in traditional RSA is not yet widely recognized. RSA based on neural patterns estimated within an imaging run is still commonly performed. Furthermore, sometimes a study might need to examine the representational similarity between task conditions within an imaging run, such that cross-run RSA is not feasible. The Taylor expansion approach to model the effect of event-interval can be difficult to set up when a task condition repeats several times in an experiment. There also lacks a detailed mathematical examination of the source of the bias and how different ways of applying RSA affect the bias. Researchers sometimes hold the view that RSA of raw fMRI patterns instead of activity patterns ($\boldsymbol{\beta}$) estimated through a general linear model (GLM) [18] does not suffer from the confounds mentioned above. Last but not least, the contribution of noise in the estimated neural patterns to the sample covariance matrix between patterns may not be restricted to the diagonal elements, as we will demonstrate below.

In this paper, we first compare the result of performing traditional RSA on a task-based fMRI dataset with the results obtained when performing the same analysis on white noise, to illustrate the severe bias and spurious similarity structure that can result from that performing RSA on pattern estimates within imaging runs. By applying task-specific RSA on irrelevant resting-state fMRI data, we show that spurious structure also emerges when RSA is performed on the raw fMRI pattern rather than estimated task activation patterns. We sow that the spurious structure can be far from a diagonal matrix, and masks any true similarity structure. We then provide an analytic derivation to help understand the source of the bias in traditional RSA. Previously, we have proposed a method named Bayesian RSA (BRSA), which significantly reduced this bias and allows analysis within imaging runs [19]. Here, we further extend the method to explicitly model spatial noise correlation, thereby mitigating the second issue identified by Heriksson et al. [14], namely the intrinsic fluctuation not modelled by task events in an experiment. Furthermore, inspired by the methods of hyper-alignment [20] and shared response models [21], we extend our method to learn a shared representational similarity structure across multiple participants (Group BRSA) and demonstrate improved accuracy of this approach. Since our method significantly reduces bias in the estimated similarity matrix but does not fully eliminate it at regimes of very low signal-to-noise ratio (SNR), we further provide a cross-validation approach to detecting over-fitting to the data. Finally, we show that the learned representational structure can serve as an empirical prior to constrain the posterior estimation of activity patterns, which can be used to decode the cognitive state underlying activity observed in new fMRI data.

The algorithm in this paper is publicly available in the python package Brain Imaging Analysis Kit (BrainIAK). [1]

[1] Under the *brainiak.reprsimil.brsa* module. Our previous version of Bayesian RSA method [19] with newly added modeling of spatial noise correlation is in the *BRSA* class of the module. The new version described in this paper is implemented in the *GBRSA* class and can be applied to either a single participant or a group of participants.

Results

Traditional RSA translates structured noise in estimated activity patterns into spurious similarity structure

Traditional RSA [8] first estimates the response amplitudes (β) of each voxel in an ROI and then calculates the similarity between the estimated spatial response patterns of that ROI to different task conditions.

The estimation of β is based on a GLM. We denote the fMRI time series from an experiment as $\mathbf{Y} \in \mathbb{R}^{n_T \times n_V}$, with n_T being the number of time points and n_V the number of voxels. The GLM assumes that

$$\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{1}$$

$\mathbf{X} \in \mathbb{R}^{n_T \times n_C}$ is the "design matrix," where n_C is the number of task conditions. Each column of the design matrix is constructed by convolving a hemodynamic response function (HRF) with a time series describing the onsets and duration of all events belonging to one task condition. The regressors composing the design matrix express the hypothesized response time course elicited by each task condition. Each voxel's response amplitudes to different task conditions can differ. All voxels' response profiles form a matrix of spatial activity patterns $\boldsymbol{\beta} \in \mathbb{R}^{n_C \times n_V}$, with each row representing the spatial pattern of activity elicited by one task condition. The responses to all conditions are assumed to contribute linearly to the spatio-temporal fMRI signal through the temporal profile of hemodynamic response expressed in $\mathbf{X}$. Thus, the measured $\mathbf{Y}$ is assumed to be a linear sum of $\mathbf{X}$ weighted by response amplitude $\boldsymbol{\beta}$, corrupted by zero-mean noise $\boldsymbol{\epsilon}$.

The goal of RSA is to understand the degree of similarity between each pair of spatial response patterns (i.e., between the rows of $\boldsymbol{\beta}$). But because the true $\boldsymbol{\beta}$ is not accessible, a point estimate of $\boldsymbol{\beta}$, derived through linear regression, is usually used as a surrogate:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\boldsymbol{T}}\mathbf{X})^{-1}\mathbf{X}^{\boldsymbol{T}}\mathbf{Y} \tag{2}$$

Similarity is then calculated between rows of $\hat{\boldsymbol{\beta}}$. For instance, one measure of similarity that is frequently used is Pearson correlation:

$$\boldsymbol{C_{ij}} = \frac{(\hat{\boldsymbol{\beta}}_{\boldsymbol{i}} - \overline{\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}})(\hat{\boldsymbol{\beta}}_{\boldsymbol{i}} - \overline{\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}})^{\boldsymbol{T}}}{\boldsymbol{n_V}\boldsymbol{\sigma}_{\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}}\boldsymbol{\sigma}_{\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}}} \tag{3}$$

To demonstrate the spurious structure that may appear in the result of traditional RSA, we first performed RSA on the fMRI data in one ROI, the orbitofrontal cortex, in a previous dataset involving a decision-making task [22]. The task included 16 different task conditions, or "states." In each state, participants paid attention to one of two overlapping images (face or house) and made judgments about the image in the attended category. The transition between the 16 task states followed the Markov chain shown in Fig **1A**, thus some states often preceded certain other states. The 16 states could be grouped into 3 categories according to the structure of transitions among states (the exact meaning of the states, or the 3 categories, are not important in the context of the discussion here.) We performed traditional RSA on the 16 estimated spatial response patterns corresponding to the 16 task states. To visualize the structure of the neural representation of the task states in the ROI, we used multi-dimensional scaling (MDS) [23] to project the 16-dimensional space defined by the distance between states (1 - correlation) onto a 3-dimensional space (Fig **1B**).

This projection appears to show clear grouping of the states in the orbitofrontal cortex consistent with the 3 categories, suggesting that this brain area represents this aspect of the task. However, a similar representational structure was also observed in

other ROIs. In addition, when we applied the same GLM to randomly generated white noise and performed RSA on the resulting parameter estimates, the similarity matrix closely resembled the result found in the real fMRI data (Fig **1C**). Since there is no task-related activity in the white noise, the structure obtained from white noise is clearly spurious and must reflect a bias introduced by the analysis. In fact, we found that the off-diagonal structure obtained from white noise (Fig **1C**) explained $\mathbf{84 \pm 12\%}$ of the variance of the off-diagonals obtained from real data (Fig **1B**). This shows that the bias introduced by traditional RSA can dominate the result, masking the real representational structure in the data.

To help understand this observation, we provide an analytic derivation of the bias with a few simplifying assumptions [19]. The calculation of the sample correlation of $\hat{\boldsymbol{\beta}}$ in traditional RSA implies the implicit assumption that an underlying covariance structure exists that describe the distribution of $\boldsymbol{\beta}$, and the activity profile of each voxel is one sample from this distribution. Therefore, examining the relation between the covariance of $\hat{\boldsymbol{\beta}}$ and that of true $\boldsymbol{\beta}$ will help us understand the bias in traditional RSA.

We assume that a covariance matrix $\mathbf{U}$ (of size $\boldsymbol{n_C} \times \boldsymbol{n_C}$) captures the true covariance structure of $\boldsymbol{\beta}$ across all voxels in the ROI: $\boldsymbol{\beta} \sim \mathbf{N(0, U)}$. Similarity measures such as correlation are derived from $\mathbf{U}$ by normalizing the diagonal elements to 1. It is well known that temporal autocorrelation exists in fMRI noise [24, 25]. To capture this, we assume that in each voxel $\boldsymbol{\epsilon} \sim \boldsymbol{N(0, \Sigma_\epsilon)}$, where $\boldsymbol{\Sigma_\epsilon} \in \mathbb{R}^{\boldsymbol{n_T} \times \boldsymbol{n_T}}$ is the temporal covariance of the noise (for illustration purposes, here we assume that all voxels have the same noise variance and autocorrelation, and temporarily assume the noise is spatially independent).

By substituting the expression for $\mathbf{Y}$ from equation 1 we obtain

$$\hat{\boldsymbol{\beta}} = (\mathbf{X^T X})^{-1}\mathbf{X^T X}\boldsymbol{\beta} + (\mathbf{X^T X})^{-1}\mathbf{X^T}\boldsymbol{\epsilon} = \boldsymbol{\beta} + (\mathbf{X^T X})^{-1}\mathbf{X^T}\boldsymbol{\epsilon} \tag{4}$$

which means the point estimate of $\boldsymbol{\beta}$ is contaminated by a noise term $\boldsymbol{(\mathbf{X^T X})^{-1}\mathbf{X^T}\epsilon}$. Assuming that the signal $\boldsymbol{\beta}$ is independent from the noise $\boldsymbol{\epsilon}$, it is then also independent from the linear transformation of the noise, $\boldsymbol{(\mathbf{X^T X})^{-1}\mathbf{X^T}\epsilon}$. Thus the covariance of $\hat{\boldsymbol{\beta}}$ is the sum of the covariance of true $\boldsymbol{\beta}$ and the covariance of $\boldsymbol{(\mathbf{X^T X})^{-1}\mathbf{X^T}\epsilon}$:

$$\hat{\boldsymbol{\beta}} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{U} + (\mathbf{X^T X})^{-1}\mathbf{X^T}\boldsymbol{\Sigma_\epsilon}\mathbf{X}(\mathbf{X^T X})^{-1}) \tag{5}$$

The term $\boldsymbol{(\mathbf{X^T X})^{-1}\mathbf{X^T}\Sigma_\epsilon}\mathbf{X}\boldsymbol{(\mathbf{X^T X})^{-1}}$ is the source of the bias in RSA. This bias originates from the structured noise $\boldsymbol{(\mathbf{X^T X})^{-1}\mathbf{X^T}\epsilon}$ in estimating $\hat{\boldsymbol{\beta}}$. It depends on both the design matrix $\mathbf{X}$ and the temporal autocorrelation of the noise $\boldsymbol{\epsilon}$. Fig **1F** illustrates how structured noise can alter the correlation of noisy pattern estimates in a simple case of just two task conditions. Even if we assume the noise is spatially and temporally independent (i.e., $\boldsymbol{\Sigma_\epsilon}$ is a diagonal matrix, which may be a valid assumption if one "pre-whitens" the data before further analysis [25]), the bias structure still exists but reduces to $\boldsymbol{(\mathbf{X^T X})^{-1}\sigma^2}$, where $\boldsymbol{\sigma^2}$ is the variance of the noise.

Since the covariance matrix of $\hat{\boldsymbol{\beta}}$ is biased, its correlation is also distorted from the true correlation structure. This is because correlation is merely a rescaling of rows and columns of a covariance matrix. Fig **1C** essentially illustrates this bias structure after being converted to correlation matrix (in this case, $\boldsymbol{\sigma}$=1 and $\boldsymbol{\beta} = \mathbf{0}$) as this RSA structure, by virtue of being derived for white noise, can only result from structure in the design matrix $\mathbf{X}$. In reality, both spatial and temporal correlations exist in fMRI noise, which complicates the structure of the bias. But the fact that bias in Fig **1C** arises even when applying RSA to white noise which itself has no spatial-temporal correlation helps to emphasize the first contributor to the bias: the timing structure of the task, which is exhibited in the correlations between the regressors in the design matrix. Whenever the interval between events of two task conditions is shorter than the length of the HRF (which typically outlasts 12 s), correlation is introduced between

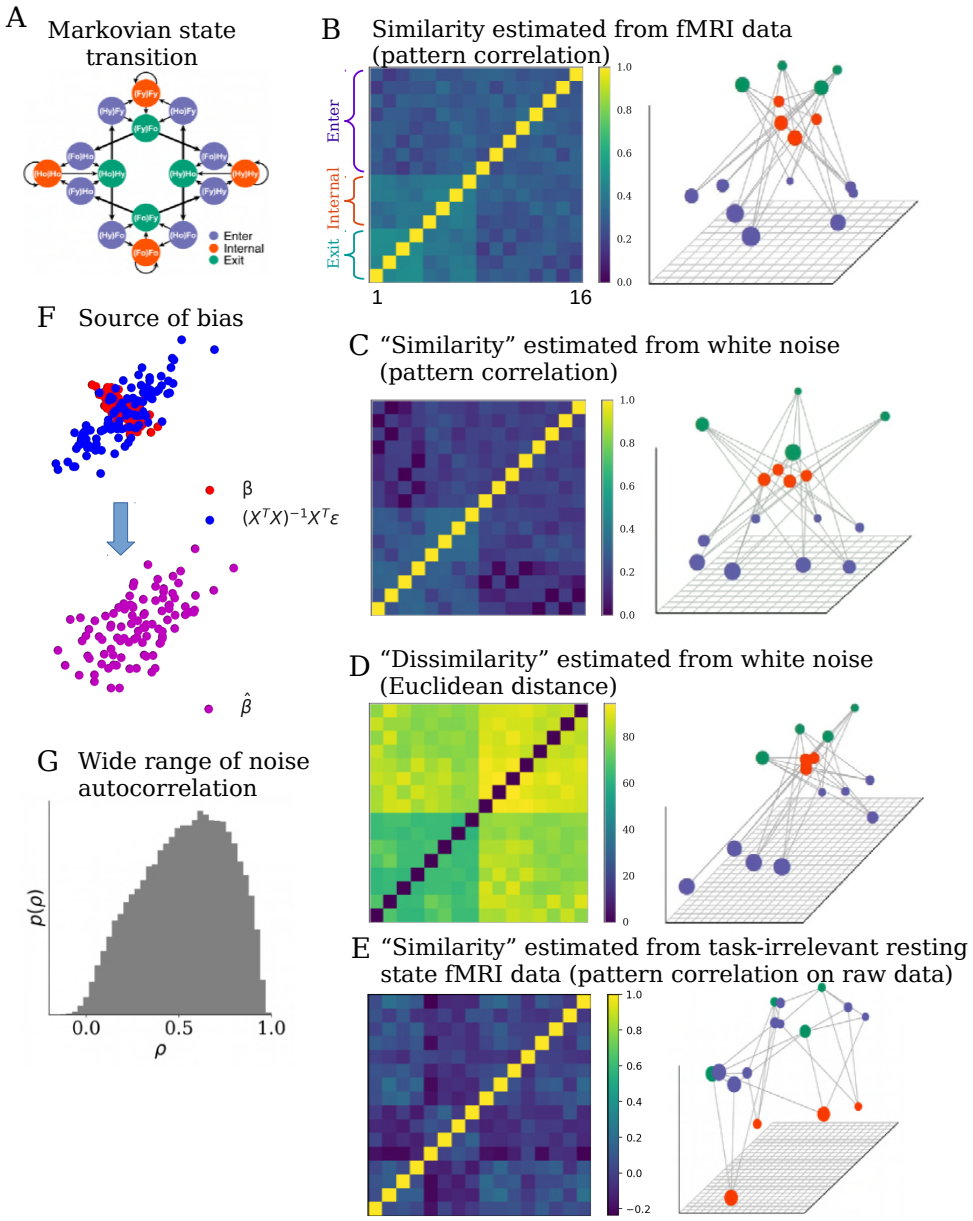

Figure 1. Standard RSA introduces bias structure to the similarity matrix. **(A)** A cognitive task including 16 different experimental conditions. Transitions between conditions follow a Markov process. Arrows indicate possible transitions, each with $\boldsymbol{p = 0.5}$. The task conditions can be grouped into 3 categories (color coded) according to their characteristic transition structure. **(B)** Standard RSA of activity patterns corresponding to each condition estimated from a brain region reveals a highly structured similarity matrix (left) that reflects aspects of the transition structure in the task. Converting the similarity matrix $\boldsymbol{C}$ to a distance matrix $\mathbf{1} - \boldsymbol{C}$ and projecting it to a low-dimensional space using MDS reveals a highly regular structure (right). Seeing such a result, one may infer that representational structure in the ROI strongly reflects the task structure.

Figure 1. **(C)** However, applying RSA to regression estimates of of patterns obtained from pure white noise generates a very similar similarity matrix (left), with a similar low-dimensional projection (right). This indicates that standard RSA can introduce spurious structure in the similarity matrix that does not exist in the data. **(D)** RSA Using Euclidean distance as a similarity metric applied to patterns estimated from the same noise (left) yields a slightly different, but still structured, similarity structure (right). **(E)** Calculating the correlation between raw patterns of resting state fMRI data (instead of patterns estimated by a GLM), assuming the same task structure as in (A), also generates spurious similarity structure, albeit different from those in (B-D). A permutation test shows that many of the high correlation values are not expected in a null distribution (details in main text). **(F)** The bias in this case comes from structured noise introduced during the GLM analysis. Assuming the true patterns $\boldsymbol{\beta}$ (red dots) of two task conditions are anti-correlated (the horizontal and vertical coordinates of each dot represent the response amplitudes of one voxel to the two task conditions), regression turns the noise $\boldsymbol{\epsilon}$ in fMRI data into structured noise $\mathbf{(X^T X)^{-1} X^T \boldsymbol{\epsilon}}$ (blue dots). The correlation between the noises in the estimated patterns is often non-zero (assumed to be positive correlation here) due to the correlation structure in the design matrix and the autocorrelation property of the noise. The estimated patterns $\hat{\boldsymbol{\beta}}$ (purple dots) are the sum of $\boldsymbol{\beta}$ and $\mathbf{(X^T X)^{-1} X^T \boldsymbol{\epsilon}}$. The correlation structure between estimated activity vectors for each condition will therefore differ from the correlation structure between the true patterns $\boldsymbol{\beta}$. **(G)** Distribution of the autocorrelation coefficients in a resting state fMRI dataset, estimated by fitting AR(1) model to the time series of each voxel resampled at TR=2.4s. The wide range of degree of autocorrelation across voxels makes it difficulty to calculate a simple analytic form of the bias structure introduced by the structured noise, and calls for modeling the noise structure of each voxel separately.

their corresponding columns in the design matrix. The degree of correlation depends on the overlapping of the HRFs. If one task condition often closely precedes another, which is the case here as a consequence of the Markovian property of the task, their corresponding columns in the design matrix are more strongly correlated. As a result of these correlations, $\mathbf{X^T X}$ is not a diagonal matrix, and neither is its inverse $\mathbf{(X^T X)^{-1}}$.

In general, unless the order of task conditions is very well counter-balanced and randomized across participants, the noise $\mathbf{(X^T X)^{-1} X^T \boldsymbol{\epsilon}}$ in $\hat{\boldsymbol{\beta}}$ is not i.i.d between task conditions. The bias term $\mathbf{B = (X^T X)^{-1} X^T \boldsymbol{\Sigma}_{\epsilon} X (X^T X)^{-1}}$ then deviates from a diagonal matrix and causes unequal distortion of the off-diagonal elements in the resulting correlation matrix of $\hat{\boldsymbol{\beta}}$. These unequal distortions alter the order of ranking of the values of the off-diagonal elements. Therefore, rank correlation between the similarity matrix from traditional RSA and similarity matrix of any candidate computational model is necessarily influenced by the bias. Conclusion based on such comparison between two similarity matrices or based on comparing a pair of off-diagonal elements within a neural similarity matrix becomes problematic, as long as the bias causes unequal distortion. Furthermore, if the design matrices also depend on participants' performance such as errors and reaction time, the bias structure could depend on their performance as well. Comparison between neural representational structure and participants' behavioral performance may also become problematic in such situations.

It is worth pointing out that the bias is not restricted to using correlation as metric of similarity. Because structured noise exists in $\hat{\boldsymbol{\beta}}$, any distance metrics between rows of $\hat{\boldsymbol{\beta}}$ estimated within imaging runs of fMRI data are likely biased. We can take Euclidean distance as an example. For any two task conditions i and j, the expectation of the distance between $\hat{\boldsymbol{\beta}}_i$ and $\hat{\boldsymbol{\beta}}_j$ is $\sum_{k=1}^{n_V} (\beta_{ik} - \beta_{jk})^2 + n_V (B_{ii}^2 + B_{jj}^2 - 2B_{ij}^2)$, where $\mathbf{B}$

is the bias in the covariance structure. Therefore, the bias $\boldsymbol{n_V(B_{ii}^2 + B_{jj}^2 - 2B_{ij}^2)}$ in Euclidean distance also depends on the task timing structure and the property of noise. (See Fig **1D**).

In our derivations above, point estimates of $\hat{\boldsymbol{\beta}}$ introduce structured noise due to the correlation structure in the design matrix. One might think that the bias can be avoided if a design matrix is not used, i.e., if RSA is not performed after GLM analysis, but directly on the raw fMRI patterns. Such an approach still suffers from bias, for two reasons that we detail below.

First, RSA on the raw activity patterns suffers from the second contributor to the bias in RSA that comes from the temporal properties of fMRI noise. To understand this, consider that estimating activity pattern by averaging the raw patterns, for instance 6 sec after each event of a task condition (that is, at the approximate peak of the event-driven HRF) is equivalent to performing an alternative GLM analysis with a design matrix $\mathbf{X_6}$ that has delta functions 6 sec after each event. Although the columns of this design matrix $\mathbf{X_6}$ are orthogonal and $\mathbf{(X_6^T X_6)^{-1}}$ becomes diagonal, the bias term is still not a diagonal matrix. Because of the autocorrelation structure $\boldsymbol{\Sigma_\epsilon}$ in the noise, the bias term $\mathbf{(X_6^T X_6)^{-1} X_6^T \Sigma_\epsilon X_6 (X_6^T X_6)^{-1}}$ essentially becomes a sampling of the temporal covariance structure of noise at the distances of the inter-event intervals. In this way, timing structure of the task and autocorrelation of noise together still cause bias in the RSA result.

To illustrate this, we applied RSA to the raw patterns of an independent set of resting state fMRI data from the Human Connectome Project [26], pretending that the participants experienced events according to the 16-state task in Fig **1A**. As shown in Fig **1E**, even in the absence of any task-related signal spurious similarity structure emerges when RSA is applied to the raw patterns of resting state data. To quantify the extent of spurious structure in Fig **1E**, we computed the null distribution of the average estimated similarity structure by randomly permuting the task condition labels on each simulated participant's estimated similarity structure 10000 times and averaging them. We then compared the absolute values of the off-diagonal elements in Fig **1E** against those in the null distribution. The Bonferroni corrected threshold for incorrectly rejecting at least one true hypothesis that an off-diagonal element in the average similarity matrix is from the null distribution is p=0.0004 for $\boldsymbol{\alpha}$=0.05. In our resting-state fake RSA matrix, 39 out of 120 off-diagonal elements significantly deviated from the null distribution based on this threshold.

Second, averaging raw data 6 sec after events of interest over-estimates the similarity between neural patterns of adjacent events, an effect independent of the fMRI noise property. This is because the true HRF in the brain has a protracted time course regardless of how one analyzes the data. Thus the estimated patterns (we denote by $\hat{\boldsymbol{\beta}}_{\mathbf{6}}$) in this approach are themselves biased due to the mismatch between the implicit HRF that this averaging assumes and the real HRF. The expectation of $\hat{\boldsymbol{\beta}}_{\mathbf{6}}$ becomes $\boldsymbol{E}[\hat{\boldsymbol{\beta}}_{\mathbf{6}}] = \boldsymbol{E}[\mathbf{(X_6^T X_6)^{-1} X_6^T Y}] = \boldsymbol{E}[\mathbf{(X_6^T X_6)^{-1} X_6^T (X\beta + \epsilon)}] = \mathbf{(X_6^T X_6)^{-1} X_6^T X}\boldsymbol{\beta}$ instead of $\boldsymbol{\beta}$. Intuitively, X temporarily smears the BOLD patterns of neural responses close in time but $\mathbf{(X_6^T X_6)^{-1} X_6^T}$ only averages the smeared BOLD patterns without disentangling the smearing. $\hat{\boldsymbol{\beta}}_{\mathbf{6}}$ thus mixes the BOLD activity patterns elicited by all neural events within a time window of approximately 12 sec (the duration of HRF) around the event of interest, causing over-estimation of the similarity between neural patterns of adjacent events. If the order of task conditions is not fully counter-balanced, this method would therefore still introduce into the estimated similarity matrix a bias caused by the structure of the task.

Similar effect can also be introduced if $\hat{\boldsymbol{\beta}}$ is estimated with regularized least square regression [27]. Regression with regularization of the amplitude of $\hat{\boldsymbol{\beta}}$ trades off bias in the estimates for variance (noise). On the surface, reducing noise in the pattern

estimates may reduce the bias introduces into the similarity matrix. However, the bias in $\hat{\boldsymbol{\beta}}$ itself alters the similarity matrix again. For example, in ridge regression, an additional penalization term $\boldsymbol{\lambda\beta^T\beta}$ is imposed for $\boldsymbol{\beta}$ for each voxel. This turns estimates $\hat{\boldsymbol{\beta}}$ to $\hat{\boldsymbol{\beta}} = (\mathbf{X^TX} + \boldsymbol{\lambda I})^{-1}\mathbf{X^TY}$. The component contributed to $\hat{\boldsymbol{\beta}}$ by the true signal $\boldsymbol{X\beta}$ becomes $(\mathbf{X^TX} + \boldsymbol{\lambda I})^{-1}\mathbf{X^T}\boldsymbol{\beta}$. As $\boldsymbol{\lambda}$ increases, this component increasingly attributes neural activity triggered by other task events near the time of an event of interest to the event's activity. Therefore, this method too would overestimate pattern similarity between adjacent events.

In all the derivations above, we have assumed for simplicity of illustration that the noise in all voxels has the same temporal covariance structure. In reality, the autocorrelation can vary over a large range across voxels (Fig. **1G**). So the structured noise in each voxel would follow a different distribution. Furthermore, the spatial correlation in noise means the noise in $\hat{\boldsymbol{\beta}}$ is also correlated across voxels, which makes the bias even more complicated. At minimum, noise correlation between voxels violates the requirement of Pearson correlation that pairs of observations should be independent.

Bayesian RSA significantly reduces bias in the estimated similarity

As shown above, the covariance structure of the noise in the point estimates of neural activity patterns $\hat{\boldsymbol{\beta}}$ leads to bias in the subsequent similarity measures. The bias can distort off-diagonal elements of the resulting similarity matrix unequally if the order of task conditions is not fully counterbalanced. In order to reduce this bias, we propose a new strategy that aims to infer directly the covariance structure $\boldsymbol{U}$ that underlies the similarity of neural patterns, using raw fMRI data. Our method avoids estimating $\hat{\boldsymbol{\beta}}$ altogether, and instead marginalizes over the unknown activity patterns $\boldsymbol{\beta}$ without discarding uncertainty about them. The marginalization avoids the structured noise introduced by the point estimates, which was the central cause of the bias. Given that the bias comes not only from the experimental design but also from the spatial and temporal correlation in noise, we explicitly model these properties in the data. We name this approach Bayesian RSA (BRSA) as it is an empirical Bayesian method [28] for estimating $\boldsymbol{U}$ as a parameter of the prior distribution of $\boldsymbol{\beta}$ directly from data.

Direct estimation of similarity matrix while marginalizing unknown neural patterns

BRSA assumes a hierarchical generative model of fMRI data. In this generative model, the covariance structure $\boldsymbol{U}$ serves as a hyper-parameter that governs the distribution of $\boldsymbol{\beta}$, which in turn generates the observed fMRI signal $\mathbf{Y}$. Each voxel $\boldsymbol{i}$ has its own noise parameters, including auto-correlation coefficient $\boldsymbol{\rho_i}$, variance $\boldsymbol{\sigma_i^2}$ of innovation noise (the noise component unpredictable from the previous time step) and pseudo-SNR $\boldsymbol{s_i}$ (we use the term 'pseudo-SNR' because the actual SNR depends on both the value of the shared covariance structure $\boldsymbol{U}$ and the voxel-specific scaling factor $\boldsymbol{s_i}$). Given these, $\boldsymbol{(\sigma_i s_i)^2 U}$ is the covariance matrix of the distribution of the activity levels $\boldsymbol{\beta_i}$ in voxel $\boldsymbol{i}$. The model allows different signal and noise parameters for each voxel to accommodate situations in which only a fraction of voxels in an ROI might have high response to tasks [27] and because the noise property can vary widely across voxels (e.g., Fig. **1G**). We denote the voxel-specific parameters ($\boldsymbol{\sigma_i^2}$, $\boldsymbol{\rho_i}$ and $\boldsymbol{s_i}$) of all voxels together as $\boldsymbol{\theta}$.

If the fMRI noise can be assumed to be independent across voxels [19], then for any single voxel $\boldsymbol{i}$, we can marginalize over the unknown latent variable $\boldsymbol{\beta_i}$ to obtain an analytic form of the likelihood of observing the fMRI data $\boldsymbol{Y_i}$ in that voxel $\boldsymbol{p(Y_i|X, U, \theta_i)}$. Multiplying the likelihoods for all voxels will result in the likelihood for the entire dataset: $\boldsymbol{p(Y|X, U, \theta)}$. Note that this computation marginalizes over $\boldsymbol{\beta}$,

avoiding altogether the secondary analysis on the point estimates $\hat{\boldsymbol{\beta}}$ that is at the heart of traditional RSA. Through the marginalization, all the uncertainty about $\boldsymbol{\beta}$ is correctly incorporated into the likelihood. By searching for the optimal $\hat{\boldsymbol{U}}$ and other parameters $\hat{\boldsymbol{\theta}}$ that maximize the data likelihood, we can therefore obtain a much less biased estimate of $\boldsymbol{U}$ for the case of spatially independent noise [19].

However, as illustrated by [14], intrinsic fluctuation shared across brain areas that is not driven by stimuli can dominate the fMRI time series and influence the RSA result. If one labels any fluctuation not captured by the design matrix as noise, then intrinsic fluctuation shared across voxels can manifest as spatial correlation in the noise, which violates our assumption above. To reduce the impact of intrinsic fluctuation on the similarity estimation, we therefore incorporate this activity explicitly into the BRSA method, with inspiration from the GLM denoising approach [29].

We start by assuming that the shared intrinsic fluctuation across voxels can be explained by a finite set of time courses, which we denote as $\boldsymbol{X_0}$, and the rest of the noise in each voxel is spatially independent. If $\boldsymbol{X_0}$ were known, the modulation $\boldsymbol{\beta_0}$ of the fMRI signal $\mathbf{Y}$ by $\boldsymbol{X_0}$ can be marginalized together with the response amplitude $\boldsymbol{\beta}$ to the experimental design matrix $\boldsymbol{X}$ (note that we still infer $\boldsymbol{U}$, the covariance structure of $\boldsymbol{\beta}$, not of $\boldsymbol{\beta_0}$). Since $\boldsymbol{X_0}$ is unknown, BRSA uses an iterative fitting procedure that alternates between a step of fitting the covariance structure $\boldsymbol{U}$ while marginalizing $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta}$, and a step of estimating the intrinsic fluctuation $\boldsymbol{X_0}$ from the residual noise with principal component analysis (PCA). Details of this procedure are described in the Materials and Methods under *Model fitting procedure*.

Since our goal is to estimate $\boldsymbol{U}$, voxel-specific parameters $\boldsymbol{\theta}$ can also be analytically or numerically marginalized so that we only need to fit $\boldsymbol{U}$ for the marginal likelihood $\boldsymbol{p(\mathbf{Y}|\mathbf{X}, \mathbf{X_0}, U)}$. This reduces the number of free parameters in the model and further allows for the extension of estimating a shared representational structure across a group of participants, as shown later. Fig **2** shows a diagram of the generative model. More details regarding the generative model and the marginalization can be found in the Materials and Methods, under *Generative model of Bayesian RSA*.

The covariance matrix $\boldsymbol{U}$ can be parameterized by its Cholesky factor $\boldsymbol{L}$, a lower-triangular matrix. To find the $\hat{\boldsymbol{U}}$ that best explains the data $\boldsymbol{Y}$, we first calculate the $\hat{\boldsymbol{L}}$ that best explains the data by optimizing the marginal log likelihood:

$$\begin{aligned}\hat{\boldsymbol{L}} &= \arg\max \log \boldsymbol{p}(\mathbf{Y}|\mathbf{X}, \mathbf{X_0}, \boldsymbol{L}) \\ &= \arg\max \sum_{\boldsymbol{i}}^{\boldsymbol{nV}} \log \int\int\int \boldsymbol{d\beta_i d\beta_{0i} d\theta_i p(Y_i|\mathbf{X}, \mathbf{X_0}, \beta_i, \beta_{0i}, \theta_i) p(\beta_i|\boldsymbol{L}, \theta_i) p(\beta_{0i}) p(\theta_i)}\end{aligned} \tag{6}$$

And then obtain the estimated covariance matrix

$$\hat{\boldsymbol{U}} = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^{\boldsymbol{T}} \tag{7}$$

Once $\hat{\boldsymbol{U}}$ is estimated (after the iterative fitting procedure for $\boldsymbol{L}$ and $\boldsymbol{X_0}$), $\hat{\boldsymbol{U}}$ is converted to a correlation matrix to yield BRSA's estimation of the similarity structure.

BRSA recovers simulated similarity structure

To test the performance of BRSA in a case where the ground-truth covariance structure is known, we embedded structure into resting state fMRI data. Signals were simulated by first sampling response amplitudes according to a hypothetical covariance structure for the "16-state" task conditions (Fig **3A**), and then weighting the design matrix of the task in Fig **1A** by the simulated response amplitudes. The simulated signals were then added to resting state fMRI data. In this way, the "noise" in the test data reflected the

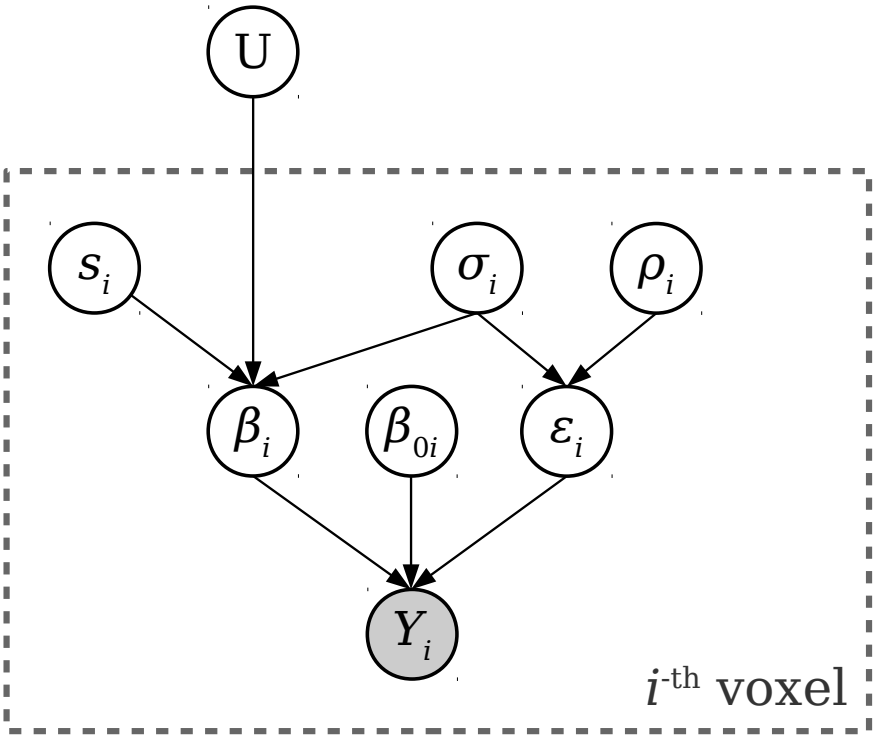

Figure 2. Generative model of Bayesian RSA. The covariance structure $\boldsymbol{U}$ shared across all voxels is treated as a hyper-parameter of the unknown response amplitude $\boldsymbol{\beta}$. For voxel $\boldsymbol{i}$, the BOLD time series $\boldsymbol{Y_i}$ are the only observable data. We assume $\boldsymbol{Y_i}$ is generated by task-related activity amplitudes $\boldsymbol{\beta_i}$ (the $\boldsymbol{i}$-th column of $\boldsymbol{\beta}$), intrinsic fluctuation amplitudes $\boldsymbol{\beta_{0i}}$ and spatially independent noise $\boldsymbol{\epsilon_i}$: $\boldsymbol{Y_i = X\beta_i + X_0\beta_{0i} + \epsilon_i}$, where $\mathbf{X}$ is the design matrix and $\mathbf{X_0}$ is the set of time courses of intrinsic fluctuations. $\boldsymbol{\epsilon_i}$ is modeled as an AR(1) process with autocorrelation coefficient $\boldsymbol{\rho_i}$ and noise standard deviation $\boldsymbol{\sigma_i}$. $\boldsymbol{\beta_i}$ depends on the voxel's pseudo-SNR $\boldsymbol{s_i}$ and noise level $\boldsymbol{\sigma_i}$ in addtion to $\boldsymbol{U}$: $\boldsymbol{\beta_i \sim N(0, (s_i\sigma_i)^2 U)}$. By marginalizing over $\boldsymbol{\beta_i}$, $\boldsymbol{\beta_{0i}}$, $\boldsymbol{\sigma_i}$, $\boldsymbol{\rho_i}$ and $\boldsymbol{s_i}$ for each voxel, we can obtain the likelihood function $\boldsymbol{p(Y_i|X, X_0, U)}$ and search for $\boldsymbol{U}$ which maximizes the total log likelihood $\boldsymbol{\log p(Y|X, X_0, U) = \sum_i^{n_V} \log p(Y_i|X, X_0, U)}$ of the observed data $\mathbf{Y}$ for all $\boldsymbol{n_V}$ voxels. The optimal $\hat{\boldsymbol{U}}$ can be converted to a correlation matrix, representing the estimated similarity between patterns.

spatial and temporal structure of realistic fMRI noise. To make the estimation task even more challenging, we simulated a situation in which within the ROI (Fig **3B**; we took the lateral occipital cortex as an ROI in this simulation, as an example) only a small set of voxels respond to the task conditions (Fig **3C**). This is to reflect the fact that SNR often varies across voxels and that an ROI is often pre-selected based on anatomical criteria or independent functional localizer, which do not guarantee that all the selected voxels will have task-related activity.

Fig **3E** shows the average covariance structure and similarity matrix estimated by BRSA. The corresponding results estimated based on $\hat{\boldsymbol{\beta}}$ in standard RSA are shown in Fig **3F**. This comparison clearly demonstrates that at low SNR and with a small amount of data, BRSA can recover the simulated covariance structure of task-related signals, while standard RSA is overwhelmed by the bias structure (eq. 5). It has been suggested that cross-run RSA, that is, similarity calculated between patterns estimated from separate scanning runs, can also reduce bias [14, 15, 17]. As shown in Fig **3G**, indeed the true covariance and similarity structure can be recovered better by this approach as compared to within-run RSA (Fig **3F**). However, this approach leads to faster degradation of results as SNR decreases, as demonstrated by the lowest two SNR levels in the simulation. The peak height of task-triggered response is often in the range of 0.1-0.5% in cognitive studies [30] while the noise level is often a few percents, which means the SNRs expected in real studies are likely in the lower range in our simulation, except when studying primary sensory stimulation. Furthermore, the inner products or correlation between noises in patterns estimated from separate runs can be positive or negative by chance. When the noise is large enough, even the correlation between pattern estimates in different runs corresponding to the same task conditions may become negative (as observed in Fig **3G**). This makes it difficult to associate results of cross-run RSA with a notion of pattern "similarity" because one would not expect patterns for a task condition to be anti-correlated across runs. Fig **3H** summarizes the average correlation between the off-diagonal elements of the estimated similarity matrix and those of the simulated similarity matrix. At high SNR, cross-run RSA's performance is similar to that of BRSA, and they both outperform within-run RSA. But BRSA performs the best at low SNR.

We also tested cross-run RSA with the estimated patterns spatially whitened using the procedure of [17]. Surprisingly, spatial whitening hurts similarity estimation. This might be because the spatial correlation structure of the simulated signal is different from that of the noise. Whitening based on the spatial correlation structure of noise would re-mix signals between different voxels to the extent of changing its similarity structure. Practically, it is difficult to estimate the spatial correlation of true signal patterns, because their estimates are always contaminated by noise.

Added bonus: inferring pseudo-SNR map

Although the voxel-specific parameters $\boldsymbol{\theta}$ are marginalized during fitting of the model, we can obtain their posterior distribution and estimate their posterior means. The estimated pseudo-SNR $\hat{\boldsymbol{s}}$ is of particular interest, as it informs us of where the estimated representational structure is more strongly supported in the ROI chosen by the researcher. As shown in Fig **3D**, the estimated pseudo-SNR map highly resembles the actual map of SNR in our simulated data in Fig **3C**, up to a scaling factor.

Estimating shared representational similarity across participants

As mentioned above, BRSA can be extended to jointly fit the data of a group of participants, thus identifying the shared representational similarity structure that best explains the data of all participants. This is achieved by searching for a single $\boldsymbol{U}$ that

maximizes the joint probability of observing all participants' data (Group Bayesian RSA ;GBRSA). The rationale or GBRSA is that it searches for the representational structure that best explains all the data. Using all the data to constrain the estimation of $\boldsymbol{U}$ reduces the variance of estimation for individual participants, an inspiration from hyper-alignment [20] and shared response model [21]. Fig **3H** shows that the similarity structure recovered by GBRSA has slightly higher correlation with the true similarity structure than the average similarity structure estimated by other methods, across most of the SNR levels and amounts of data. Cross-run RSA performs better only at the highest simulated SNR. However, low average SNR is common in many brain areas and this is where (G)BRSA offers more power for detecting the true but weak similarity structure.

Controlling for over-fitting: model selection by cross-validation on left-out data

Although Fig **2** shows that BRSA reduces bias, it does not eliminate it completely. This may be due to over-fitting to noise. Because it is unlikely that the time course of intrinsic fluctuation $\boldsymbol{X_0}$ and the design matrix $\boldsymbol{X}$ are perfectly orthogonal, part of the intrinsic fluctuation cannot be distinguished from task-related activity. Therefore, the structure of $\boldsymbol{\beta_0}$, the modulation of intrinsic fluctuation, could also influence the estimated $\hat{\boldsymbol{U}}$ when SNR is low.

For instance, in Fig **3E**, at the lowest SNR and least amount of data (top left subplot), the true similarity structure is almost undetectable using BRSA. Is this due to large variance in the estimates, or is it because BRSA is still biased, but to a lesser degree than standard RSA? If the result is still biased, then averaging results across subjects will not remove the bias, and the deviation of the average estimated similarity structure from the true similarity structure should not approach 0. To test this, we simulated many more subjects by preserving the spatial patterns of intrinsic fluctuation and the auto-regressive properties of the voxel-specific noise in the data used in Fig **3**, and generating intrinsic fluctuations that maintain the amplitudes of power spectrum in the frequency domain. To expose the limit of the performance of BRSA, we focused on the lower range of SNR and simulated only one run of data per "subject". Fig **4A** shows the quality of the average estimated similarity matrix with increasing number of simulated subjects. The average similarity matrices estimated by BRSA do not approach the true similarity matrix indefinitely as the number of subjects increase. Instead, their correlation saturates to a value smaller than 1. This indicates that the result of BRSA is still weakly biased, with the bias depending on the SNR. It is possible that as the SNR approaches 0, the estimated $\hat{\boldsymbol{U}}$ is gradually dominated by the impact of the part of $\boldsymbol{X_0}$ not orthogonal to $\boldsymbol{X}$. We reason that this is partly because the algorithm [31] we used to estimate the number of components in $\boldsymbol{X_0}$ is a relatively conservative method. In particular, in this simulation, the number of components of simulated intrinsic fluctuations were 20 ± 4, while the number of components estimated from these simulated data by the algorithm were 13 ± 3. However, empirically this algorithm [31] yields more stable and reasonable estimation than other methods we have tested [32]. It should be noted that BRSA still performs much better than standard RSA, for which the correlation between the estimated similarity matrix and the true similarity matrix never passed 0.1 in these simulations (not shown).

The expected bias structure when spatial noise correlation exists is difficult to derive. We used $\boldsymbol{(X^T X)^{-1}}$ as a proxy to evaluate the residual bias in the estimated similarity using BRSA. As expected, when the SNR approached zero, the model over-fit to the noise and the bias structure increasingly dominated the estimated structure despite increasing the number of simulated participants (Fig **4B**). This observation calls for an

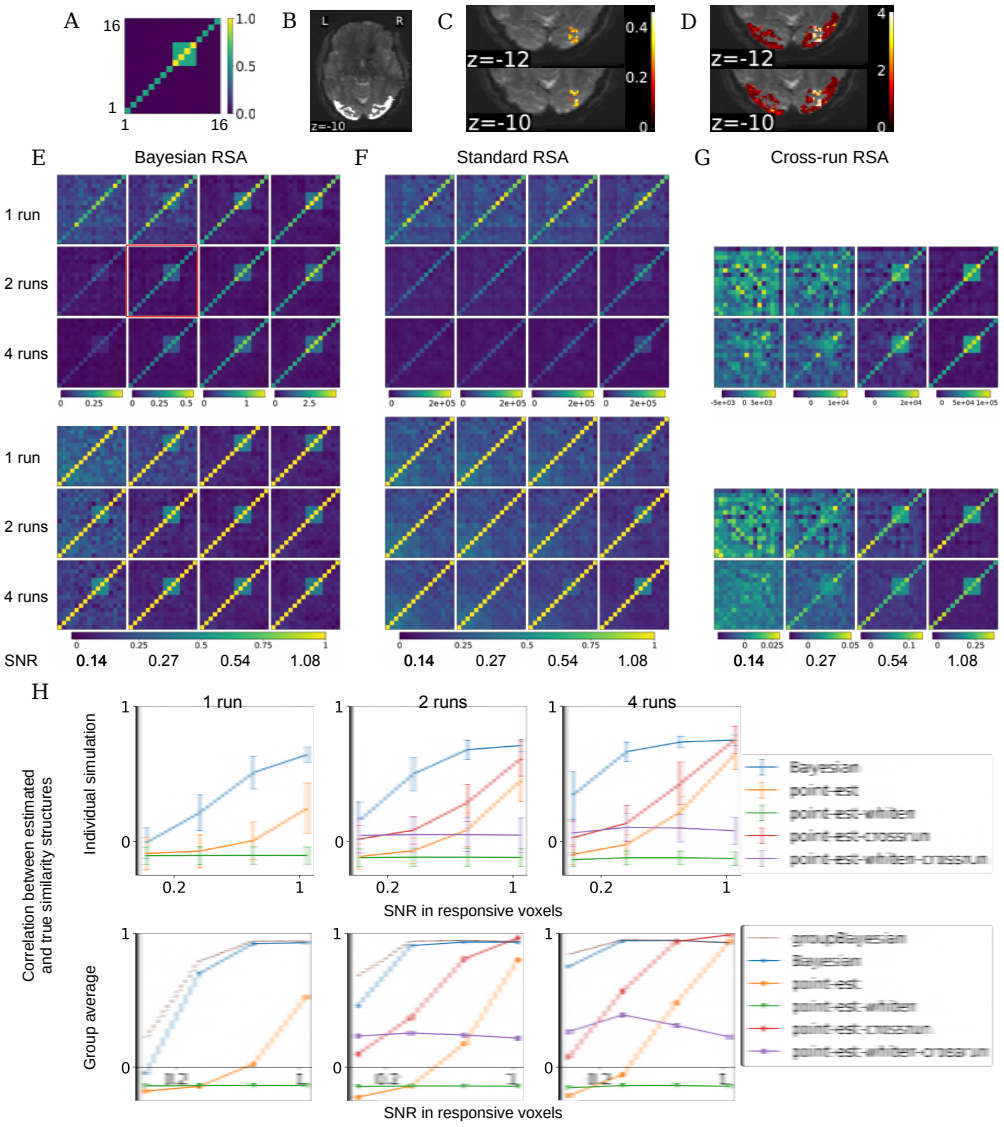

Figure 3. Performance of BRSA on simulated data. (A) The true covariance structure $\boldsymbol{U}$ from which the activity patterns were drawn. **(B)** We use lateral occipital cortex (bright region) as an example ROI and resting state fMRI data from the Human Connectome Project as noise. **(C)** We multiplied the design matrix of the task in Fig 1A with the simulated activity pattern and then added this "signal" to voxels that in a cubical region of the ROI. The colors show the actual SNR of the added signal for one example simulated brain, corresponding to the plot circumvented by a red square in E. **(D)** The pseudo-SNR map estimated by BRSA for the data with a true SNR map shown in C. The scale does not match the scale of true SNR, but the spatial pattern of SNR is recovered. **(E)** Average covariance matrix (top) and similarity matrix (bottom) estimated by BRSA in the cubic area in C, across different SNR levels (columns) and different numbers of runs (rows). **(F)** The corresponding result obtained by standard RSA based on activity patterns estimated within runs. **(G)** The corresponding result of RSA based on cross-correlating patterns estimated from separate runs.

Figure 3. (H) Top: average correlation (mean ± std) between the off-diagonal elements of the estimated and true similarity matrices, for each method, across SNR levels (x-axis) and amounts of data (separate plots). Bottom: The correlation between the average estimated similarity matrix of each method (for GBRSA, this is the single similarity matrix estimated) and the true similarity matrix. "point-est": methods based on point estimates of activity patterns; "-crossrun": similarity based on cross-correlation between runs; "-whiten": patterns were spatially whitened (similarity matrix not shown because the true structure could barely be seen)

evaluation procedure to detect over-fitting in applications to real data, when the ground truth of the similarity structure is unknown.

One approach to assess whether a BRSA model has over-fit the noise is cross-validation. In addition to estimating $\boldsymbol{U}$, the model can also estimate the posterior mean of all other parameters, including the neural patterns $\boldsymbol{\beta}$ of task-related activity, $\boldsymbol{\beta_0}$ of intrinsic fluctuation, noise variances $\boldsymbol{\sigma^2}$ and auto-correlation coefficients $\boldsymbol{\rho}$. For a left-out testing data set, the design matrix $\boldsymbol{X_{test}}$ is known given the task design. Together with the parameters estimated from the training data as well as the estimated variance and auto-correlation properties of the intrinsic fluctuation in the training data, we can calculate the log predictive probability of observing the test data. The unknown intrinsic fluctuation in the test data can be marginalized by assuming their statistical property stays unchanged from training data to test data. The predictive probability can then be contrasted against the cross-validated predictive probability provided by a null model separately fitted to the training data. The null model would have all the same assumptions as the full BRSA model, except that it would not assume any task-related activity captured by $\boldsymbol{X}$. When BRSA over-fits the data, the estimated spatial pattern $\hat{\boldsymbol{\beta}}$ would not reflect the true response pattern to the task and is unlikely to be modulated by the time course in $\boldsymbol{X_{test}}$. Thus the full model would predict signals that do not occur in the test data, and yield a lower predictive probability than the null model. The result of the full BRSA model on training data can therefore be accepted if the log predictive probability by the full model is higher than that of the null model significantly more often than chance.

Over-fitting might also arise when the assumed design matrix $\boldsymbol{X}$ does not correctly reflect task-related activity. When there is a sufficient amount of data but the design matrix does not reflect the true activity, the estimated covariance matrix $\hat{\boldsymbol{U}}$ in BRSA would approach zero, as would the posterior estimates of $\hat{\boldsymbol{\beta}}$. In this case as well, the full model would be indistinguishable from the null model.

We tested the effectiveness of relying on cross-validation to reject over-fitted results using the same simulation procedure as in Fig **3**, and repeated this simulation 36 times, each time with newly simulated signals and data from a new group of participants in HCP [33] as "noise". Fig **5A** shows the rate of correct acceptance when both training and test data have signals. We counted each simulation in which the cross-validation score (log predictive probability) of the full BRSA model was significantly higher than the score of the null model (based on a one-sided student's t-test at a threshold of $\boldsymbol{\alpha}$=0.05) as one incidence of correct acceptance. When the SNR is high (above 0.14), warranting reliable estimation of the similarity structures as indicated in Fig **3H**, the cross-validation procedure selected the full model significantly more often than chance (all p<7e-7, binomial test). At the lowest SNR (0.14) and with only 1 run of training data, the full model was never selected (p<3e-11), consistent with a poor estimation of the similarity matrix in Fig **3E**. As the amount of training data increased to 2 runs (even without changing the SNR), the rate of accepting the full model increased, although with the lowest SNR it was still not significantly different from chance (p=0.6), while the estimated similarity matrix was also noisy but started to be visually

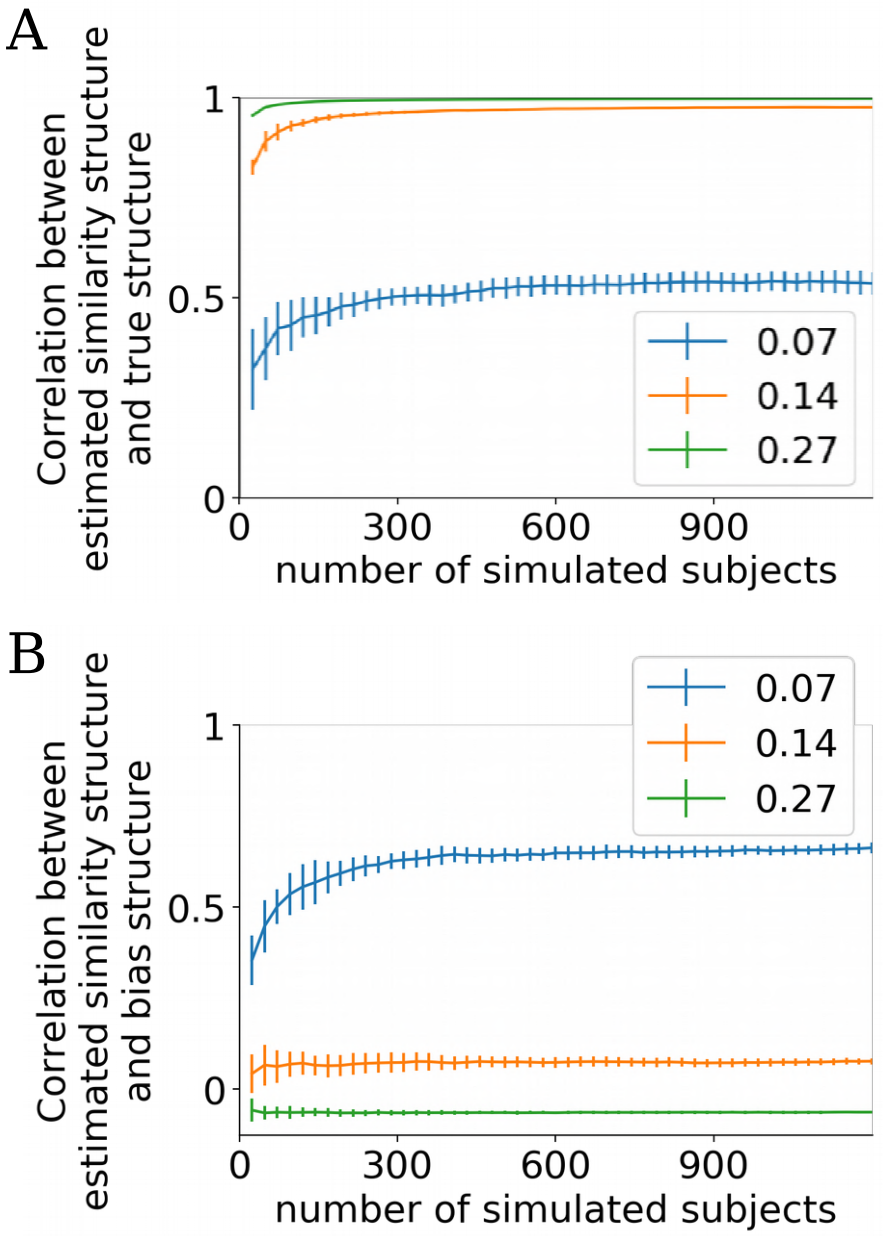

Figure 4. Limited performance of BRSA at very low SNR and small amount of data. (A) The average correlation between the off-diagonal elements of the estimated and true similarity matrices (mean ± std) as the number of simulated subjects increases. Each simulated subject had one run of data. Legend shows average SNR in task-responsive voxels. Half of the voxels do not include any signal related to the design matrix. The correlation reaches asymptotic levels slightly below 1 with increasing numbers of participants except when the SNR is extremely low (0.07), indicating that the bias is not fully eliminated.**(B)** The average correlation between the estimated similarity matrix and the expected bias structure assuming white noise. The estimated similarity structure is most dominated by the bias structure at the lowest SNR simulated (0.07). The negative correlation at the highest SNR reflects the weak negative correlation between the true similarity structure and expected bias structure (-0.055)

detectable in Fig **3E**. This indicates that the cross-validation procedure is relatively conservative. Fig **5B** shows the difference between the cross-validation scores of the full and null models as the amount of training data doubled in one group of simulated subjects, as an illustration of typical results. Dots to the left of the dashed line represent subjects for whom the full model explained the data better than the null model to a greater extent when two runs were used, as compared to one run of data only. The means and standard deviations of the t-statistics across simulated groups for all simulation configurations are displayed in Fig **5C**. The differences in cross-validation scores between full and null models are displayed in Fig **5D**.

The cross-validation procedure also helps avoid false acceptance when activity patterns are not consistently reproducible across runs. To illustrate this, we simulated the case when signals are only added to the training data but not to the test data. Now, the full model was always rejected across the simulated SNR and amounts of data (not shown). Finally, when neither training data nor testing data included signal, the cross-validation procedure also correctly rejected the full model in all cases. Fig **5E** and **5F** illustrate the difference between cross-validation scores of full and null models for the two simulations, respectively.

Extension: decoding task-related activity from new data

BRSA has a relatively rich model for the data: it attempts to model both the task-related signal and intrinsic fluctuation, and to capture voxel-specific SNR and noise properties. In addition to cross-validation, this also enables decoding of signals related to task conditions from new data. Similarly to the procedure of calculating cross-validated log likelihood, but without pre-assuming a design matrix for the test data, we can calculate the posterior mean of $\hat{\boldsymbol{X}}_{\boldsymbol{test}}$ and $\hat{\boldsymbol{X}}_{\boldsymbol{0test}}$ in the testing data. Fig **6A** shows the decoded design matrix $\hat{\boldsymbol{X}}_{\boldsymbol{test}}$ for one task condition (condition 6 in Fig **1B** and **6B**) and one participant, using one run of training data with the second-highest SNR. Although our method decodes some spurious high responses when there is no event of this task condition, overall the result captures many of the true responses in the design matrix. The average correlation between the decoded design matrix and the true design matrix is displayed in Fig **6B**. High values on the diagonal elements indicate that overall, the decoder based on BRSA can recover the task-related signals well. The structure of the off-diagonal elements appears highly similar to those of the correlation structure between corresponding columns in the original design matrix($r=0.82$, $p<1e\text{-}30$). This means that the signals corresponding to task conditions which often occur closely in time in training data are more likely be confused when they are decoded from testing data. Indeed, at the time of mistakenly decoded high response around the 90th TR in Fig **6A**, there is a true event of the first task condition ((Fo)Fy) in the design matrix. The decoder confused the response to the first condition as response to the sixth condition. The events of these two conditions did in fact often co-occur in the training data, therefore their overlap in the design matrix makes it difficult to distinguish which event triggered the response in the training data, and reduces the accuracy of posterior estimates of their activity patterns, causing further confusion at the stage of decoding. We suspect that such confusion is not limited to decoding based on BRSA, but should be a general limitation of multi-variate pattern analysis of fMRI data: due to the slow smooth BOLD response, the more often the events of two task conditions occur closely in time in the training data, the more difficult it becomes for the classifier to discern their patterns.

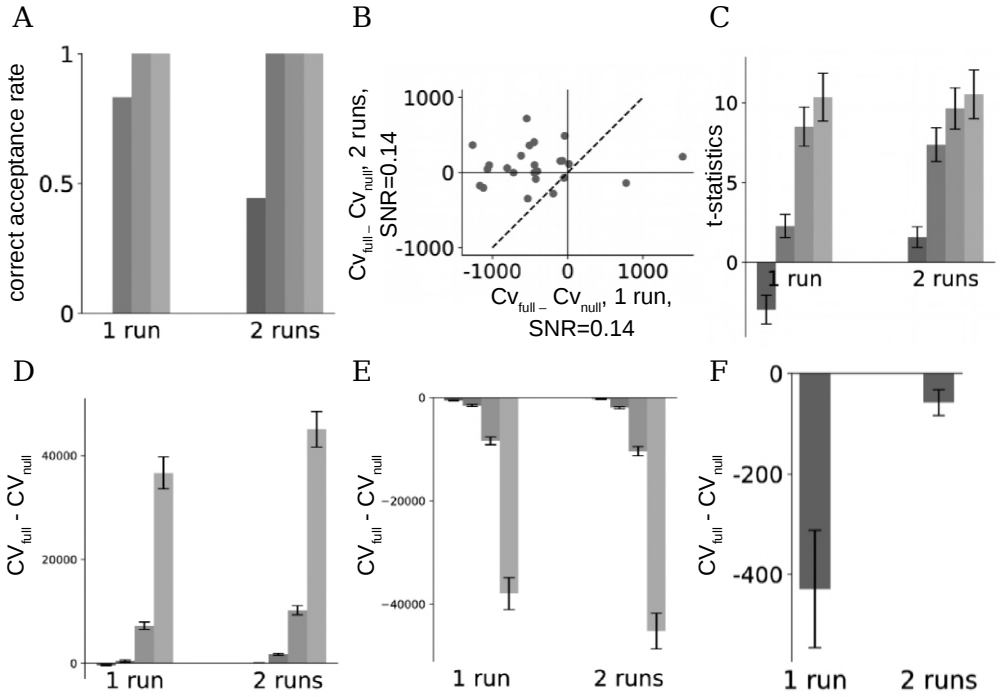

Figure 5. Cross-validation reduces the chance of false positive results. The full BRSA model and a null model that assumes no task-related activity were fit to 1 or 2 runs of simulated data of a group of subjects with different SNRs, as in Fig 3. Student's t-test was performed on the difference between the cross-validation scores of the full model and null model on a left-out run of data to determine whether the full model should be accepted. This procedure was repeated 36 times on different groups of simulated data. **(A)** Signals were added to both training and test data. The frequencies with which the full models were accepted based on the t-test (correct acceptance) are displayed. Darker bars correspond to low SNRs and lighter bars correspond to higher SNRs in Fig 3. **(B)** The difference between cross-validation scores of the full and null models for 1 or 2 runs of training data, at the lowest SNR (0.14). Dashed line: $\boldsymbol{x} = \boldsymbol{y}$. Points to the left of the dashed line show more evidence for the full model when two runs of data were used, as compared to one run. The chance of accepting the full model increases when there are more data to fit the BRSA model. **(C)** Mean $\pm$ std of the t-statistics of the difference between cross-validation scores of the full and null models across simulated groups, for the corresponding amounts of data and SNR in A. **(D)** Mean $\pm$ std of the difference between the cross-validation scores of the full models and the null models across simulated groups in A. **(E)** Mean $\pm$ std of the difference between the cross-validation scores when only the training data but not test data have signals. **(F)** Mean $\pm$ std of the difference between the cross-validation scores when neither the training data nor the test data have signals. In all cases in E and F the statistical test correctly rejected the full model.

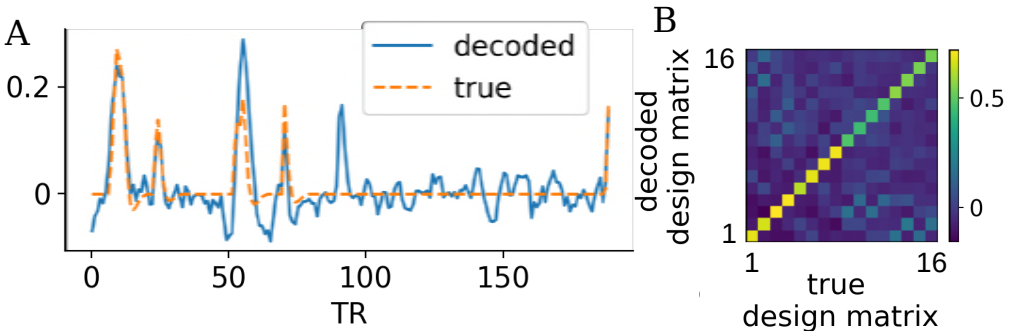

Figure 6. Decoding capabilities of the BRSA method. (A) Decoded task-related activity of the sixth condition from one simulated subject in one run of test data, and the true design matrix of that condition in the test data. The simulated data with the second highest SNR in Fig 3 were used. BRSA model was fitted to one run of training data. **(B)** Average correlation between the decoded signals for each task condition (rows) and the time courses for each condition in the design matrix used to simulate the test data (columns).

Discussion

In this paper, we demonstrated that bias can arise in the result of representational similarity analysis, a popular method in many recent fMRI studies. By analytically deriving the source of the bias with simplifying assumptions, we showed that it is determined by both the timing structure of the experiment design and the correlation structure of the noise in the data. Traditional RSA is based on point estimates of neural activation patterns which unavoidably include high amounts of noise. The task design and noise property induce covariance structure in the noise of the pattern estimates. This structure in turn biases the covariance structure of these point estimates, and a bias persists in the similarity matrix. Such bias is especially severe when the SNR is low and when the order of the task conditions cannot be fully counterbalanced.

To reduce this bias, we proposed a Bayesian framework that interprets the representational structure as reflecting the shared covariance structure of activity levels across voxels. Our BRSA method estimates this covariance structure directly from data, bypassing the structured noise in the point estimates of activity levels, and explicitly modeling the spatial and temporal structure of the noise. This is different from many other methods that attempt to correct the bias after it has been introduced.

In addition to inferring the representational similarity structure, our method also infers activation patterns (as an alternative to the traditional GLM), SNR for different voxels, and even the "design matrix" for data recorded without knowledge of the underlying conditions. The inferred activation patterns are regularized not only by the SNR, but also by the learned similarity structure. The inference of an unknown "design matrix" allows one to uncover uninstructed task conditions (e.g., in free thought) using the full Bayesian machinery and all available data.

In a realistic simulation using real fMRI data as background noise, we showed that BRSA generally outperforms standard RSA and cross-run RSA, especially when SNR is low and when the amount of data is limited, making out method a good candidate in scenarios of low SNR and difficult-to-balance tasks. Because temporal and spatial correlation also exist in the noise of data from other neural recording modalities, the method can also be applied to other types of data when the bias in standard RSA is of

concern. To detect overfitting to noise, the difference between the cross-validated score of the full model of BRSA and a null model can serve as the basis for model selection. We further extend the model to allow for estimating the shared representational structure across a group of participants.

The bias demonstrated in this paper does not necessarily question the validity of all previous results generated by RSA. However, it does call for more caution when applying RSA to higher-level brain areas for which SNR in fMRI is typically low, and when the order between events of different task conditions cannot be fully counterbalanced. This is especially the case with decision making tasks that involve learning or structured sequential decisions, in which events cannot be randomly shuffled. Even when the order of task conditions can be randomized, it may not be perfectly counter-balanced. Thus, a small deviation of the bias structure from a diagonal matrix may still exists. If the same random sequence is used for all participants, the tiny bias can persist in the results of all participant and become a confound. Therefore, it is also important to use different task sequences across participants.

Prior to the proposal of our method, similarity measures calculated between patterns estimated from separate scanning runs (cross-run RSA) was proposed to overcome the bias [15, 17, 34]. The inner product between noise pattern estimates from separate runs is theoretically unbiased. However, at low SNR, cross-run RSA suffers from large noise, sometimes generating results where noisy pattern estimates of the same condition from different runs appear anti-correlated. In addition, even though the cross-run covariance matrix is not biased, the magnitude of cross-run correlation is under-estimated because the computation requires division by the standard deviation of the estimated patterns, which is in turn inflated by the high amount of noise carried in the estimated patterns. In our simulation, cross-run RSA appears to slightly outperform BRSA at very high SNR but the results of both methods are already very close to the true similarity structure in this case. On the other hand, at very low SNR, cross-run RSA fails to reveal the true similarity structure, while BRSA does. However, cross-run RSA may be a more conservative approach given that the cross-run covariance matrix (and Mahalonobis distance [17] is unbiased. It is difficult to predict whether BRSA or cross-run RSA are more suitable for any specific study and brain area of interest. Nonetheless, based on our results, both approaches should always be favored over traditional within-run RSA based on pattern estimates.

It is surprising that spatial whitening, that is often recommended [17], in fact hurts the result of standard RSA and cross-run RSA in our simulation. This may be because, in our simulation, the spatial correlation of noise is not the same as the spatial correlation in the simulated signal. While whitening reduces the correlation between noise in the estimated $\hat{\boldsymbol{\beta}}$ of different voxels, it may cause undesired remixing of true signals between voxels. As discussed before, in practice, it is difficult to know whether the intrinsic fluctuation and task-evoked signals share the same spatial correlation structure, because we do not know the ground truth of signals in real data. The cost and benefit of spatial whitening on standard and cross-validated RSA therefore awaits more studies. Instead of performing spatial whitening, BRSA estimates a few time series $\boldsymbol{X_0}$ that best explain the correlation of noise between voxels and marginalizes their modulations in each voxel. Without remixing signals across voxels, it still captures spatial noise correlation.

In our study, we did not directly compare BRSA to cross-validated Mahalanobis distance [17] because they are foundamentally different measures: BRSA aims to estimate the correlation between patterns, which is close to the cosine angle between two patterns vectors [35, 36]; in contrast, Mahalanobis distance aims to measure the distance between patterns. Nonetheless, given the theoretical soundness of the cross-validated Mahalanobis distance, it could also be a good alternative to BRSA when

there are multiple runs in a task.

Our BRSA method is closely related to the PCM [16, 37]. A major difference is that PCM models the point estimates $\hat{\boldsymbol{\beta}}$ after GLM analysis while BRSA models fMRI data $\mathbf{Y}$ directly. The original PCM [16] in fact considered the contribution of the noise in pattern estimates to the similarity matrix, but assumed that the noise in $\hat{\boldsymbol{\beta}}$ is i.i.d across task conditions. This means that the bias in the covariance matrix was assumed to be restricted to a diagonal matrix. We showed here that when the order of task conditions cannot be fully counter-balanced, such as in the example in Fig **1**, this assumption is violated and the bias cannot be accounted for by methods such as PCM.

If one knew the covariance structure of the noise $\boldsymbol{\Sigma_\epsilon}$, then the diagonal component of the noise covariance structure assumed in PCM [16] could be replaced by the bias term $\mathbf{(X^T X)^{-1} X^T \Sigma_\epsilon X (X^T X)^{-1}}$ to adapt PCM to estimate the covariance structure $\hat{\boldsymbol{U}}$ that best explains $\hat{\boldsymbol{\beta}}$ [38] if spatial noise correlation is not considered. However, as shown in Fig **1G**, different voxels can have a wide range of different autocorrelation coefficients. Assuming a single $\boldsymbol{\Sigma_\epsilon}$ for all voxels may be over-simplified. In addition, PCM assumes all voxels within one ROI have equal SNR. However, typically only a small fraction of voxels exhibits high SNR [27]. Therefore, it is useful to model the noise property and SNR of each voxel individually.

In addition to these differences, BRSA explicitly models spatial noise correlation. It also comes with the ability to select between a full model and null model based on cross-validated log likelihood, and the method can be applied to fMRI decoding. PCM can additionally evaluate the likelihood of a few fixed candidate representational structures given by different computational models. It can also estimate the additive contributions of several candidate pattern covariance structures to the observed covariance structure. These options are not yet available in the current implementation of BRSA. Combining the strength of PCM and BRSA is an interesting future direction.

Many aspects of flexibility may be incorporated to BRSA. For example, the success of the analysis hinges on the assumption that the HRF used in the design matrix correctly reflects the true hemodynamics in the ROI, but it has been found that HRF in fact vary across people and across brain regions [39, 40]. Jointly fitting the shape of the HRF and the representational structure may improve the estimation. In addition, it is possible that even if the landscape of activity patterns for a task condition stays the same, the global amplitude of the response pattern may vary across trials due to repetition supression [41–43] and attention [44, 45]. Such modulation may not be predictable by response time or stimulus duration. Allowing global amplitude modulation of patterns associated with a task condition to vary across trials might capture such variability and increase the power of the method.

Our simulations revealed that BRSA is not entirely unbiased, that is, results cannot be improved indefinitely by adding more subjects. We hypothesize that the residual bias is due to the underestimation of the number of components necessary to capture the spatial correlation introduced by intrinsic fluctuation. Development of a proper but less conservative algorithm for estimating the number of components suitable for BRSA may improve its performance.

Comparing the cross-validation score of the full model and a null model is one approach to detect overfitting. One interesting finding is that when the design matrix does not explain the real brain response (Fig **5C** where signal was not added to either training or test data), and when there is a sufficient amount of training data, the full model becomes indistinguishable from the null model. Even though such cross-validation does not select the null model significantly more often than the full model, not finding the opposite is sufficient to warn the researcher not to trust the resulting similarity matrix as reflecting the true structure. When this happens, it is advisable to focus on taking measures to improve the design of study. Ultimately, task

designs that are not fully counterbalanced and low SNR in fMRI data are two critical factors that cause bias in traditional RSA and impact the power of detecting similarity structure. Carefully designing tasks that fully balance the task conditions, randomizing the sequence of a task across participants, and increasing the number of measurements, are our recommended approaches in the first place. In the analysis phase of the project, one can then use BRSA.

Materials and methods

Generative model of Bayesian RSA

Our generative model of fMRI data follows the general assumption of GLM. In addition, we model spatial noise correlation by a few time series $\boldsymbol{X_0}$ shared across all voxels. The contribution of $\boldsymbol{X_0}$ to the $\boldsymbol{i^{th}}$ voxel is $\boldsymbol{\beta_{0i}}$. Thus, for voxel $\boldsymbol{i}$, we assume that

$$\boldsymbol{Y_i = X\beta_i + X_0\beta_{0i} + \epsilon_i} \tag{8}$$

$\boldsymbol{Y_i}$ is the time series of voxel $\boldsymbol{i}$. $\boldsymbol{X}$ is the design matrix shared by all voxels. $\boldsymbol{\beta_i}$ is the response amplitudes of the voxel $\boldsymbol{i}$ to the task conditions. $\boldsymbol{\epsilon_i}$ is the residual noise in voxel $\boldsymbol{i}$ which cannot be explained by either $\boldsymbol{X}$ or $\boldsymbol{X_0}$. We assume that $\boldsymbol{\epsilon}$ is spatially independent across voxels, and all the correlation in noise between voxels are captured by the shared intrinsic fluctuation $\boldsymbol{X_0}$.

We use an AR(1) process to model $\boldsymbol{\epsilon_i}$: for the $\boldsymbol{i^{th}}$ voxel, we denote the noise at time $\boldsymbol{t > 0}$ as $\boldsymbol{\epsilon_{t,i}}$, and assume

$$\boldsymbol{\epsilon_{t,i} = \rho_i\epsilon_{t-1,i} + \eta_{t,i}, \quad \eta_{t,i} \sim N(0, \sigma_i^2)} \tag{9}$$

where $\boldsymbol{\sigma_i^2}$ is the variance of the "innovation" noise at each time point and $\boldsymbol{\rho_i}$ is the autoregressive coefficient for the $\boldsymbol{i^{th}}$ voxel.

We assume that the covariance of the multivariate Gaussian distribution from which the activity amplitudes $\boldsymbol{\beta_i}$ are generated has a scaling factor that depends on its pseudo-SNR $\boldsymbol{s_i}$:

$$\boldsymbol{\beta_i \sim N(0, (s_i\sigma_i)^2 U).} \tag{10}$$

This is to reflect the fact that not all voxels in an ROI respond to tasks.

We further use Cholesky decomposition to parametrize the covariance structure $\boldsymbol{U}$: $\boldsymbol{U = LL^T}$, where $\boldsymbol{L}$ is a lower triangular matrix. Thus, $\boldsymbol{\beta_i}$ can be written as $\boldsymbol{\beta_i = s_i\sigma_i L\alpha_i}$, where $\boldsymbol{\alpha_i \sim N(0, I)}$. This change of parameter allows for estimating $\mathbf{U}$ of lower rank (if the researcher has sufficient reason to make such a guess) by setting $\boldsymbol{L}$ as lower-triangular matrix with a few rightmost-columns truncated. With an improper uniform prior for $\boldsymbol{\beta_{0i}}$, and temporarily assuming $\boldsymbol{X_0}$ is given, we have the unmarginalized likelihood for each voxel $\boldsymbol{i}$:

$$\begin{aligned}
&\boldsymbol{p(Y_i,\beta_i,\beta_{0i}|X, X_0, L, \sigma_i, \rho_i, s_i)}\\
&\quad\boldsymbol{=p(Y_i|\beta_i,\beta_{0i}, X, X_0, \sigma_i, \rho_i)p(\beta_i|L, \sigma_i, s_i)p(\beta_{0i})}\\
&\quad\boldsymbol{=p(Y_i|s_i\sigma_i L\alpha_i,\beta_{0i}, X, X_0, \sigma_i, \rho_i)p(\alpha_i)p(\beta_{0i})}\\
&\quad\boldsymbol{\propto p(Y_i|s_i\sigma_i L\alpha_i,\beta_{0i}, X, X_0, \sigma_i, \rho_i)p(\alpha_i)}\\
&\quad\boldsymbol{= \exp[-\frac{1}{2}(Y_i - s_i\sigma_i XL\alpha_i - X_0\beta_{0i})^T\Sigma_{\epsilon_i}^{-1}(Y_i - s_i\sigma_i XL\alpha_i - X_0\beta_{0i})]}\\
&\qquad\boldsymbol{\cdot (2\pi)^{-\frac{n_T}{2}}|\Sigma_{\epsilon_i}^{-1}|^{\frac{1}{2}}(2\pi)^{-\frac{k}{2}}\exp[-\frac{1}{2}\alpha_i^T\alpha_i]}
\end{aligned} \tag{11}$$

where $\boldsymbol{k \leq n_C}$ is the rank of $\boldsymbol{L}$.

In contrast to the full model, our null model assumes

$$\begin{aligned}
&p(\boldsymbol{Y_i},\boldsymbol{\beta_{0i}}|\boldsymbol{X_0},\boldsymbol{\sigma_i},\boldsymbol{\rho_i}) \\
&\quad=p(\boldsymbol{Y_i}|\boldsymbol{\beta_{0i}},\boldsymbol{X_0},\boldsymbol{\sigma_i},\boldsymbol{\rho_i})p(\boldsymbol{\beta_{0i}}) \\
&\quad\propto p(\boldsymbol{Y_i}|\boldsymbol{\beta_{0i}},\boldsymbol{X_0},\boldsymbol{\sigma_i},\boldsymbol{\rho_i}) \\
&\quad=\exp[-\frac{1}{2}(\boldsymbol{Y_i}-\boldsymbol{X_0}\boldsymbol{\beta_{0i}})^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon_i}}^{-1}(\boldsymbol{Y_i}-\boldsymbol{X_0}\boldsymbol{\beta_{0i}})] \\
&\qquad\cdot(2\pi)^{-\frac{n_T}{2}}|\boldsymbol{\Sigma}_{\boldsymbol{\epsilon_i}}^{-1}|^{\frac{1}{2}}
\end{aligned} \tag{12}$$

For data within one run, $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon_i}}^{-1}$, the inverse matrix of the covariance of $\boldsymbol{\epsilon_i}$, is a banded symmetric matrix which can be written as $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon_i}}^{-1}=\frac{1}{\boldsymbol{\sigma_i^2}}(\boldsymbol{I}-\boldsymbol{\rho_i}\boldsymbol{F}+\boldsymbol{\rho_i^2}\boldsymbol{D})$, where $\boldsymbol{F}$ is 1 only at the superdiagonal and subdiagonal elements and 0 everywhere else, and $\boldsymbol{D}$ is 1 on all diagonal elements except for the first and last one, and 0 elsewhere. For abbreviation, we can denote $\boldsymbol{A_i}=\boldsymbol{A}(\boldsymbol{\rho_i})=\boldsymbol{I}-\boldsymbol{\rho_i}\boldsymbol{F}+\boldsymbol{\rho_i^2}\boldsymbol{D}$ which is a function of $\boldsymbol{\rho_i}$. $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon_i}}^{-1}$ can be factorized as $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon_i}}^{-1}=\frac{1}{\boldsymbol{\sigma_i^2}}\boldsymbol{A_i}$. When $\boldsymbol{Y_i}$ includes concatenated time series across several runs, $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon_i}}^{-1}$ is a block diagonal matrix with each block diagonal elements corresponding to one run, constructed in the same way.

To derive the log likelihood of $\boldsymbol{L}$ for data of all voxels in the ROI, we need to marginalizing all other unknown parameters. Below, we marginalize them step by step.

By marginalizing $\boldsymbol{\beta_{0i}}$, we have

$$\begin{aligned}
&p(\boldsymbol{Y_i},\boldsymbol{\beta_i}|\boldsymbol{X},\boldsymbol{X_0},\boldsymbol{L},\boldsymbol{\sigma_i},\boldsymbol{\rho_i},\boldsymbol{s_i}) \\
&\quad\propto\int p(\boldsymbol{Y_i}|\boldsymbol{s_i}\boldsymbol{\sigma_i}\boldsymbol{L}\boldsymbol{\alpha_i},\boldsymbol{\beta_{0i}},\boldsymbol{X},\boldsymbol{X_0},\boldsymbol{\sigma_i},\boldsymbol{\rho_i})p(\boldsymbol{\alpha_i})d\boldsymbol{\beta_{0i}} \\
&\quad=(2\pi)^{-\frac{n_T+k-n_0}{2}}|\boldsymbol{\Sigma}_{\boldsymbol{\epsilon_i}}^{-1}|^{\frac{1}{2}}|\boldsymbol{X_0}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon_i}}^{-1}\boldsymbol{X_0}|^{-\frac{1}{2}}\exp[-\frac{1}{2}\boldsymbol{\alpha_i}^T\boldsymbol{\alpha_i}] \\
&\qquad\cdot\exp[-\frac{1}{\boldsymbol{2\sigma_i^2}}(\boldsymbol{Y_i}-\boldsymbol{s_i}\boldsymbol{\sigma_i}\boldsymbol{X}\boldsymbol{L}\boldsymbol{\alpha_i})^T\boldsymbol{A_i^*}(\boldsymbol{Y_i}-\boldsymbol{s_i}\boldsymbol{\sigma_i}\boldsymbol{X}\boldsymbol{L}\boldsymbol{\alpha_i})]
\end{aligned} \tag{13}$$

$\boldsymbol{n_0}$ is the number of components in $\boldsymbol{X_0}$. In the equation above, we denoted $\boldsymbol{A_i^*}=\boldsymbol{\sigma_i^2}(\boldsymbol{\Sigma}_{\boldsymbol{\epsilon_i}}^{-1}-\boldsymbol{\Sigma}_{\boldsymbol{\epsilon_i}}^{-1}\boldsymbol{X_0}(\boldsymbol{X_0}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon_i}}^{-1}\boldsymbol{X_0})^{-1}\boldsymbol{X_0}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon_i}}^{-1})=\boldsymbol{A_i}-\boldsymbol{A_i}\boldsymbol{X_0}(\boldsymbol{X_0}^T\boldsymbol{A_i}\boldsymbol{X_0})^{-1}\boldsymbol{X_0}^T\boldsymbol{A_i}$.

By further marginalizing $\boldsymbol{\alpha_i}$ which is equivalent to marginalizing $\boldsymbol{\beta_i}$, we get

$$\begin{aligned}
&p(\boldsymbol{Y_i}|\boldsymbol{X},\boldsymbol{X_0},\boldsymbol{L},\boldsymbol{\sigma_i},\boldsymbol{\rho_i},\boldsymbol{s_i}) \\
&\quad=\int p(\boldsymbol{Y_i}|\boldsymbol{s_i}\boldsymbol{\sigma_i}\boldsymbol{L}\boldsymbol{\alpha_i},\boldsymbol{X},\boldsymbol{X_0},\boldsymbol{\sigma_i},\boldsymbol{\rho_i})p(\boldsymbol{\alpha_i})d\boldsymbol{\alpha_i} \\
&\quad\propto(2\pi)^{-\frac{n_T-n_0}{2}}|\boldsymbol{\Sigma}_{\boldsymbol{\epsilon_i}}^{-1}|^{\frac{1}{2}}|\boldsymbol{X_0}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon_i}}^{-1}\boldsymbol{X_0}|^{-\frac{1}{2}}|\boldsymbol{\Lambda_i^*}|^{\frac{1}{2}} \\
&\qquad\cdot\exp[-\frac{1}{2}(\frac{1}{\boldsymbol{\sigma_i^2}}\boldsymbol{Y_i}^T\boldsymbol{A_i^*}\boldsymbol{Y_i}-\boldsymbol{\mu_i^{*T}}\boldsymbol{\Lambda_i^{*-1}}\boldsymbol{\mu_i^*})]
\end{aligned} \tag{14}$$

where $\boldsymbol{\Lambda_i^*}=(\boldsymbol{I}+\boldsymbol{s_i^2}\boldsymbol{L}^T\boldsymbol{X}^T\boldsymbol{A_i^*}\boldsymbol{X}\boldsymbol{L})^{-1}$ and $\boldsymbol{\mu_i^{*T}}=\frac{\boldsymbol{s_i}}{\boldsymbol{\sigma_i}}\boldsymbol{\Lambda_i^*}\boldsymbol{L}^T\boldsymbol{X}^T\boldsymbol{A_i^*}\boldsymbol{Y_i}$ are the variance and mean of the posterior distribution of $\boldsymbol{\alpha_i}$, respectively.

All the steps of marginalization above utilize the property of multivariate Gaussian distribution. Next we marginalize the noise variance $\boldsymbol{\sigma_i^2}$. We assume an improper uniform distribution of $\boldsymbol{\sigma_i^2}$ in $\mathbb{R}^+$. It is also possible to assume a conjugate prior for $\boldsymbol{\sigma_i^2}$. Given that data of at least hundreds of time points are obtained in each run to provide enough constraint to $\boldsymbol{\sigma_i^2}$, our choice does not appear to cause problem. To isolate $\boldsymbol{\sigma_i^2}$, using the property of Cholesky decomposition of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon_i}}^{-1}$, the above equation can be

written as

$$\begin{aligned} &p(\boldsymbol{Y_i}|\boldsymbol{X}, \boldsymbol{X_0}, \boldsymbol{L}, \boldsymbol{\sigma_i}, \boldsymbol{\rho_i}, \boldsymbol{s_i}) \\ &\propto (2\pi)^{-\frac{n_T-n_0}{2}} \boldsymbol{\sigma_i^2}^{-\frac{n_T-n_0}{2}} (1-\boldsymbol{\rho_i^2})^{\frac{n_r}{2}} |\boldsymbol{X_0^T A_i X_0}|^{-\frac{1}{2}} |\boldsymbol{\Lambda_i^*}|^{\frac{1}{2}} \\ &\cdot \exp[\frac{1}{2\boldsymbol{\sigma_i^2}} (\boldsymbol{s_i^2 Y_i^T A_i^* X L \Lambda_i^* L^T X^T A_i^* Y_i - Y_i^T A_i^* Y_i})] \end{aligned} \tag{15}$$

This form is proportional to an inverse-Gamma distribution of $\boldsymbol{\sigma_i^2}$. $\boldsymbol{n_r}$ is the number of runs in the data. Therefore, we can analytically marginalize $\boldsymbol{\sigma_i^2}$ and obtain

$$\begin{aligned} &p(\boldsymbol{Y_i}|\boldsymbol{X}, \boldsymbol{X_0}, \boldsymbol{L}, \boldsymbol{\rho_i}, \boldsymbol{s_i}) \\ &= \int p(\boldsymbol{Y_i}|\boldsymbol{X}, \boldsymbol{X_0}, \boldsymbol{L}, \boldsymbol{\sigma_i}, \boldsymbol{\rho_i}, \boldsymbol{s_i}) p(\boldsymbol{\sigma_i^2}) d\boldsymbol{\sigma_i^2} \\ &\propto (2\pi)^{-\frac{n_T-n_0}{2}} (1-\boldsymbol{\rho_i^2})^{\frac{n_r}{2}} |\boldsymbol{X_0^T A_i X_0}|^{-\frac{1}{2}} |\boldsymbol{\Lambda_i^*}|^{\frac{1}{2}} \Gamma(\frac{\boldsymbol{n_T-n_0}}{2}-1) \\ &\cdot [\frac{\boldsymbol{Y_i^T A_i^* Y_i - s_i^2 Y_i^T A_i^* X L \Lambda_i^* L^T X^T A_i^* Y_i}}{2}]^{1-\frac{n_T-n_0}{2}} \end{aligned} \tag{16}$$

We did not find ways to further analytically marginalize $\boldsymbol{s_i}$ or $\boldsymbol{\rho_i}$. But we can numerically marginalize them by weighted sum of 16 at $\boldsymbol{n_l \times n_m}$ discrete grids $\boldsymbol{\{\rho_{il}, s_{im}\}}$ $(\boldsymbol{0 < l < n_l}, \boldsymbol{0 < m < n_m})$ with each grid representing one area of the parameter space of $\boldsymbol{(\rho, s)}$.

$$\begin{aligned} &p(\boldsymbol{Y_i}|\boldsymbol{X}, \boldsymbol{X_0}, \boldsymbol{L}) \\ &\approx \sum_{l=1}^{n_l} \sum_{m=1}^{n_m} p(\boldsymbol{Y_i}|\boldsymbol{X}, \boldsymbol{X_0}, \boldsymbol{L}, \boldsymbol{\rho_{il}}, \boldsymbol{s_{im}}) w(\boldsymbol{\rho_{il}}, \boldsymbol{s_{im}}) \\ &\propto \sum_{l=1}^{n_l} \sum_{m=1}^{n_m} (2\pi)^{-\frac{n_T-n_0}{2}} (1-\boldsymbol{\rho_{il}^2})^{\frac{n_r}{2}} |\boldsymbol{X_0^T A_{il} X_0}|^{-\frac{1}{2}} |\boldsymbol{\Lambda_{ilm}^*}|^{\frac{1}{2}} \Gamma(\frac{\boldsymbol{n_T-n_0}}{2}-1) \\ &\cdot [\frac{\boldsymbol{Y_i^T A_{il}^* Y_i - s_{im}^2 Y_i^T A_{il}^* X L \Lambda_{ilm}^* L^T X^T A_{il}^* Y_i}}{2}]^{1-\frac{n_T-n_0}{2}} w(\boldsymbol{\rho_{il}}, \boldsymbol{s_{im}}) \end{aligned} \tag{17}$$

The weights $\boldsymbol{w(\rho_{il}, s_{im})}$ are the prior probabilities of the two parameters in the area represented by $\boldsymbol{\{\rho_{il}, s_{im}\}}$. We assume uniform prior of $\boldsymbol{\rho}$ in (-1,1). All the simulations in this paper used a negative exponential distribution as prior for $\boldsymbol{s}$. The grids $\boldsymbol{s_{im}}$ are each chosen at the centers of mass of the prior distribution in the bins they represent in $\boldsymbol{(0, +\infty)}$. All bins equally divide the area under the curve of the prior distribution for $\boldsymbol{s}$. Alternative forms of priors such as uniform in (0, 1) and truncated log normal distribution are also implemented in the tool.

Because we made the assumption that $\boldsymbol{\epsilon_i}$ is independent across voxels. The log likelihood for all data is the sum of the log likelihood for each voxel.

$$\log p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{X_0}, \boldsymbol{L}) = \sum_{i=1}^{n_V} \log p(\boldsymbol{Y_i}|\boldsymbol{X}, \boldsymbol{X_0}, \boldsymbol{L}). \tag{18}$$

For the null model, the likelihood for each voxel after marginalizing $\boldsymbol{\beta_{0i}}$ and $\boldsymbol{\sigma_i^2}$ can be similarly derived,

$$\begin{aligned} &p(\boldsymbol{Y_i}|\boldsymbol{X_0}, \boldsymbol{\rho_i}) \\ &\propto (2\pi)^{-\frac{n_T-n_0}{2}} (1-\boldsymbol{\rho_i^2})^{\frac{n_r}{2}} |\boldsymbol{X_0^T A_i X_0}|^{-\frac{1}{2}} \\ &\cdot \Gamma(\frac{\boldsymbol{n_T-n_0}}{2}-1) [\frac{\boldsymbol{Y_i^T A_i^* Y_i}}{2}]^{1-\frac{n_T-n_0}{2}} \end{aligned} \tag{19}$$

and the total log likelihood can be calculated similarly by numerically marginalizing $\boldsymbol{\rho_i}$ and summing the log likelihood for all voxels.

Model fitting procedure

To fit the model, we need the derivative of the total log likelihood with respect to $\boldsymbol{L}$. It can be derived that conditional on any grid of parameter pairs $\{\boldsymbol{\rho_{il}}, \boldsymbol{s_{im}}\}$, the derivative of the log likelihood for voxel $\boldsymbol{i}$ against each lower-triangular element of $\boldsymbol{L}$ is the corresponding lower-triangular element of the matrix

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{L}} \log & p(\boldsymbol{Y_i}|\boldsymbol{X}, \boldsymbol{X_0}, \boldsymbol{L}, \boldsymbol{\rho_{il}}, \boldsymbol{s_{im}}) \\ &= -\boldsymbol{s_{im}^2}\boldsymbol{X}^T\boldsymbol{A_{il}^*}\boldsymbol{X}\boldsymbol{L}\boldsymbol{\Lambda_{ilm}^*} \\ &\quad + \frac{\boldsymbol{s_{im}^2}(\boldsymbol{n_T} - \boldsymbol{n_0} - \boldsymbol{2})}{\boldsymbol{Y_i}^T\boldsymbol{A_{il}^*}\boldsymbol{Y_i} - \boldsymbol{s_{im}^2}\boldsymbol{Y_i}^T\boldsymbol{A_{il}^*}\boldsymbol{X}\boldsymbol{L}\boldsymbol{\Lambda_{ilm}^*}\boldsymbol{L}^T\boldsymbol{X}^T\boldsymbol{A_{il}^*}\boldsymbol{Y_i}} \\ &\quad \cdot (\boldsymbol{I} - \boldsymbol{s_{im}^2}\boldsymbol{X}^T\boldsymbol{A_{il}^*}\boldsymbol{X}\boldsymbol{L}\boldsymbol{\Lambda_{ilm}^*}\boldsymbol{L}^T)\boldsymbol{X}^T\boldsymbol{A_{il}^*}\boldsymbol{Y_i}\boldsymbol{Y_i}^T\boldsymbol{A_{il}^*}\boldsymbol{X}\boldsymbol{L}\boldsymbol{\Lambda_{ilm}^*}\end{aligned} \tag{20}$$

where $\boldsymbol{A_{il}^*}$ and $\boldsymbol{\Lambda_{ilm}^*}$ are $\boldsymbol{A_i^*}$ and $\boldsymbol{\Lambda_i^*}$ evaluated at $\{\boldsymbol{\rho_{il}}, \boldsymbol{s_{im}}\}$. The derivative of the total log likelihood against $\boldsymbol{L}$ after marginalizing over all grids $\{\boldsymbol{\rho_{il}}, \boldsymbol{s_{im}}\}$ of all voxels is

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{L}} \log & p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{X_0}, \boldsymbol{L}) \\ &= \sum_{i=1}^{n_V}\sum_{l=1}^{n_l}\sum_{m=1}^{n_m} p(\boldsymbol{\rho_{il}}, \boldsymbol{s_{im}}|\boldsymbol{Y_i}, \boldsymbol{X}, \boldsymbol{X_0}, \boldsymbol{L})\frac{\partial}{\partial \boldsymbol{L}} \log p(\boldsymbol{Y_i}|\boldsymbol{X}, \boldsymbol{X_0}, \boldsymbol{L}, \boldsymbol{\rho_{il}}, \boldsymbol{s_{im}})\end{aligned} \tag{21}$$

$\boldsymbol{p(\rho_{il}, s_{im}|Y_i, X, X_0, L)}$ is the posterior probability of $\{\boldsymbol{\rho_{il}}, \boldsymbol{s_{im}}\}$ conditional on a given $\boldsymbol{L}$. It can be obtained by normalizing $\boldsymbol{p(Y_i|X, X_0, L, \rho_{il}, s_{im})w(\rho_{il}, s_{im})}$ after calculating 16.

With the derivative 21, the total log likelihood 18 can be maximized using gradient-based method such as Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm to search for the optimal $\boldsymbol{L}$ [46–49].

However, the derivations above have made the assumption that $\boldsymbol{X_0}$ is given, while it is not. The requirement for $\boldsymbol{X_0}$ should be to appropriately capture the correlation of noise across voxels without overfitting. Therefore, at the starting of the model fitting, regular regression of $\boldsymbol{Y}$ against $\boldsymbol{X}$ and any nuisance regressors such as head motion and constant baseline is performed. Then the algorithm by Gavish and Donoho [31] is used to select the optimal number of components $\boldsymbol{n_0}$ to choose $\boldsymbol{X_0}$ from the eigenvectors of the residual of regression. Because regular regression does not shrink the magnitudes of $\boldsymbol{\beta}$, their magnitudes can only be over-estimated. $\boldsymbol{n_0}$ thus has no risk of being over-estimated. This $\boldsymbol{n_0}$ is then fixed throughout the model fitting. Next, the first $\boldsymbol{n_0}$ principal components of the residual of regression are set as $\hat{\boldsymbol{X}}_\mathbf{0}$ to allow for calculating the marginal log likelihood in 21 and gradient ascent with BFGS. A sufficient steps of iterations are performed to optimize $\boldsymbol{L}$. Then $\hat{\boldsymbol{\beta}}_{\text{post}}$, the posterior expectations of $\boldsymbol{\beta}$, are calculated with the current $\hat{\boldsymbol{L}}$ and with $\boldsymbol{s}$, $\boldsymbol{\rho}$, $\boldsymbol{\sigma}$ being marginalized. $\hat{\boldsymbol{X}}_\mathbf{0}$ is subsequently recalculated using PCA from the residuals after subtracting $\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{post}}$ from $\boldsymbol{Y}$. The alternation between optimizing $\boldsymbol{L}$ and re-estimating $\hat{\boldsymbol{X}}_\mathbf{0}$ is repeated until convergence.

Once we obtain $\hat{\boldsymbol{L}}$, the estimate of $\boldsymbol{L}$, the estimate of the covariance structure is $\hat{\boldsymbol{U}} = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^{\boldsymbol{T}}$. Converting it into a correlation matrix yields the similarity matrix by BRSA. Even though $\hat{\boldsymbol{X}}_\mathbf{0}$ is estimated from data based on posterior estimation of $\boldsymbol{\beta}$ repeatedly during fitting, $\boldsymbol{L}$ is still optimized for the log likelihood with all other unknown variables marginalized. Thus the estimated $\hat{\boldsymbol{U}}$ is an empirical prior of $\boldsymbol{\beta}$

estimated from data. This is the reason we consider our model as an empirical Bayesian method.

Many subcomponents of the expressions in these equations do not depend on $\boldsymbol{L}$ and thus can be pre-computed before optimizing for $\boldsymbol{L}$. The fixed grids of $(\boldsymbol{\rho}, \boldsymbol{s})$ further make several subcomponents shared across voxels when evaluating 16. These all reduce the amount of computation needed.

The fitting of the null model is similar to that of the full model except that there is no $\boldsymbol{L}$ to be optimized.

Model selection and decoding

Once a model has been fitted to some data from a participant or a group of participants, we can estimate the posterior mean of $\boldsymbol{\rho}$, $\boldsymbol{s}$, $\boldsymbol{\sigma^2}$, $\boldsymbol{\beta}$ and $\boldsymbol{\beta_0}$, conditional on the empirical prior $\hat{\boldsymbol{U}}$ (essentially $\hat{\boldsymbol{L}}$), data $\boldsymbol{Y}$, design matrix $\boldsymbol{X}$ and estimated intrinsic fluctuations $\hat{\boldsymbol{X}}_{\boldsymbol{0}}$. Below, we derive their formula and the procedure in which they are used for calculating cross-validated log likelihood of new data and decoding task-related signal $\hat{\boldsymbol{X}}_{\text{test}}$ and $\hat{\boldsymbol{X}}_{\boldsymbol{0}\text{test}}$ from new data in the context of fMRI decoding.

The posterior mean of these variables are

$$\begin{aligned}\hat{\boldsymbol{\sigma^2}}_{\boldsymbol{i}(\text{post})} &= \sum_{l=1}^{n_l}\sum_{m=1}^{n_m} p(\boldsymbol{\rho}_{\boldsymbol{il}}, \boldsymbol{s}_{\boldsymbol{im}}|\boldsymbol{Y}_{\boldsymbol{i}}, \boldsymbol{X}, \hat{\boldsymbol{X}}_{\boldsymbol{0}}, \hat{\boldsymbol{L}}) \int \boldsymbol{\sigma}_{\boldsymbol{i}}^{\boldsymbol{2}} p(\boldsymbol{\sigma}_{\boldsymbol{i}}^{\boldsymbol{2}}|\boldsymbol{Y}_{\boldsymbol{i}}, \boldsymbol{L}, \boldsymbol{X}, \hat{\boldsymbol{X}}_{\boldsymbol{0}}, \boldsymbol{\rho}_{\boldsymbol{il}}, \boldsymbol{s}_{\boldsymbol{im}}) d\boldsymbol{\sigma}_{\boldsymbol{i}} \\ &= \sum_{l=1}^{n_l}\sum_{m=1}^{n_m} p(\boldsymbol{\rho}_{\boldsymbol{il}}, \boldsymbol{s}_{\boldsymbol{im}}|\boldsymbol{Y}_{\boldsymbol{i}}, \boldsymbol{X}, \hat{\boldsymbol{X}}_{\boldsymbol{0}}, \hat{\boldsymbol{L}}) \frac{\boldsymbol{1}}{\boldsymbol{n_T} - \boldsymbol{n_0} - \boldsymbol{4}} \\ &\quad \cdot \boldsymbol{Y}_{\boldsymbol{i}}^{\boldsymbol{T}} \boldsymbol{A}_{\boldsymbol{il}}^{*} \boldsymbol{Y}_{\boldsymbol{i}} - \boldsymbol{s}_{\boldsymbol{im}}^{\boldsymbol{2}} \boldsymbol{Y}_{\boldsymbol{i}}^{\boldsymbol{T}} \boldsymbol{A}_{\boldsymbol{il}}^{*} \boldsymbol{X} \hat{\boldsymbol{L}} \boldsymbol{\Lambda}_{\boldsymbol{ilm}}^{*} \hat{\boldsymbol{L}}^{\boldsymbol{T}} \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{A}_{\boldsymbol{il}}^{*} \boldsymbol{Y}_{\boldsymbol{i}}\end{aligned} \tag{22}$$

$$\hat{\boldsymbol{s}}_{\boldsymbol{i}(\text{post})} = \sum_{l=1}^{n_l}\sum_{m=1}^{n_m} p(\boldsymbol{\rho}_{\boldsymbol{il}}, \boldsymbol{s}_{\boldsymbol{im}}|\boldsymbol{Y}_{\boldsymbol{i}}, \boldsymbol{X}, \hat{\boldsymbol{X}}_{\boldsymbol{0}}, \hat{\boldsymbol{L}}) \boldsymbol{s}_{\boldsymbol{im}} \tag{23}$$

$$\hat{\boldsymbol{\rho}}_{\boldsymbol{i}(\text{post})} = \sum_{l=1}^{n_l}\sum_{m=1}^{n_m} p(\boldsymbol{\rho}_{\boldsymbol{il}}, \boldsymbol{s}_{\boldsymbol{im}}|\boldsymbol{Y}_{\boldsymbol{i}}, \boldsymbol{X}, \hat{\boldsymbol{X}}_{\boldsymbol{0}}, \hat{\boldsymbol{L}}) \boldsymbol{\rho}_{\boldsymbol{il}} \tag{24}$$

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{i}(\text{post})} = \sum_{l=1}^{n_l}\sum_{m=1}^{n_m} p(\boldsymbol{\rho}_{\boldsymbol{il}}, \boldsymbol{s}_{\boldsymbol{im}}|\boldsymbol{Y}_{\boldsymbol{i}}, \boldsymbol{X}, \hat{\boldsymbol{X}}_{\boldsymbol{0}}, \hat{\boldsymbol{L}}) \boldsymbol{s}_{\boldsymbol{im}}^{\boldsymbol{2}} \hat{\boldsymbol{L}} \boldsymbol{\Lambda}_{\boldsymbol{ilm}} \hat{\boldsymbol{L}}^{\boldsymbol{T}} \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{A}_{\boldsymbol{il}}^{*} \boldsymbol{Y}_{\boldsymbol{i}} \tag{25}$$

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\boldsymbol{0i}(\text{post})} &= \sum_{l=1}^{n_l}\sum_{m=1}^{n_m} p(\boldsymbol{\rho}_{\boldsymbol{il}}, \boldsymbol{s}_{\boldsymbol{im}}|\boldsymbol{Y}_{\boldsymbol{i}}, \boldsymbol{X}, \hat{\boldsymbol{X}}_{\boldsymbol{0}}, \hat{\boldsymbol{L}}) \\ &\quad \cdot (\hat{\boldsymbol{X}}_{\boldsymbol{0}}^{\boldsymbol{T}} \boldsymbol{A}_{\boldsymbol{il}}^{*} \hat{\boldsymbol{X}}_{\boldsymbol{0}})^{-1} \hat{\boldsymbol{X}}_{\boldsymbol{0}}^{\boldsymbol{T}} \boldsymbol{A}_{\boldsymbol{il}}^{*} (\boldsymbol{Y}_{\boldsymbol{i}} - \boldsymbol{s}_{\boldsymbol{im}}^{\boldsymbol{2}} \boldsymbol{X} \hat{\boldsymbol{L}} \boldsymbol{\Lambda}_{\boldsymbol{ilm}} \hat{\boldsymbol{L}}^{\boldsymbol{T}} \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{A}_{\boldsymbol{il}}^{*} \boldsymbol{Y}_{\boldsymbol{i}})\end{aligned} \tag{26}$$

For null model, $\hat{\boldsymbol{\sigma^2}}_{\boldsymbol{i}(\text{post})}$, $\hat{\boldsymbol{\beta}}_{\boldsymbol{0i}(\text{post})}$ and $\hat{\boldsymbol{\rho}}_{\boldsymbol{i}(\text{post})}$ are of similar forms except that all terms including $\boldsymbol{s}_{\boldsymbol{im}}$ are removed and that $p(\boldsymbol{\rho}_{\boldsymbol{il}}, \boldsymbol{s}_{\boldsymbol{im}}|\boldsymbol{Y}_{\boldsymbol{i}}, \boldsymbol{X}, \hat{\boldsymbol{X}}_{\boldsymbol{0}}, \hat{\boldsymbol{L}})$ is replaced by $p(\boldsymbol{\rho}_{\boldsymbol{il}}|\boldsymbol{Y}_{\boldsymbol{i}}, \hat{\boldsymbol{X}}_{\boldsymbol{0}})$.

To calculate cross-validated log likelihood, we assume the posterior estimates above and the statistical properties of $\boldsymbol{X_0}$ stay unchanged in the testing data. We use zero-mean AR(1) process to describe the statistical properties of $\boldsymbol{X_0}$. The AR(1) parameters estimated from $\hat{\boldsymbol{X}}_{\boldsymbol{0}}$ serve as the parameters of the empirical prior for $\boldsymbol{X_0}$ in the testing data. When $\boldsymbol{X_0}$ at each time point $\boldsymbol{t}$ is treated as a random vector $\boldsymbol{X}_{\boldsymbol{0}}^{(\boldsymbol{t})}$, the AR(1) parameters of each component can be jointly written as the diagonal matrix $\boldsymbol{V}_{\boldsymbol{\Delta X_0}}$ for the variance of the innovation noise, and diagonal matrix $\boldsymbol{T}_{\boldsymbol{X_0}}$ for the auto-regressive coefficients, both of size $\boldsymbol{n_0} \times \boldsymbol{n_0}$.

For model selection purpose, design matrix $\boldsymbol{X}_{\text{test}}$ for the testing data should be generated in the same manner as they are for the training data by the researcher. For full BRSA model, $\boldsymbol{X}_{\text{test}}\hat{\boldsymbol{\beta}}_{\text{post}}$ is the predicted task-related signal in $\boldsymbol{Y}_{\text{test}}$. $\boldsymbol{Y}_{\text{res}} = \boldsymbol{Y}_{\text{test}} - \boldsymbol{X}_{\text{test}}\hat{\boldsymbol{\beta}}_{\text{post}}$ is the residual variation which cannot be explained by the design matrix and the posterior activity pattern $\hat{\boldsymbol{\beta}}_{\text{post}}$. Null model does not predict any task-related activity, so all $\boldsymbol{Y}_{\text{test}}$ constitutes residual variation $\boldsymbol{Y}_{\text{res}}$. In either the full model or the null model, the posterior estimate $\hat{\boldsymbol{\beta}}_{\boldsymbol{0i}(\text{post})}$ expresses their prediction about how voxels should be co-modulated by a fluctuation, while the fluctuation time course $\boldsymbol{X}_{\boldsymbol{0}\text{test}}$ is only predictable in terms of its variance and temporal autocorrelation expressed by $\boldsymbol{V}_{\boldsymbol{\Delta X_0}}$ and $\boldsymbol{T}_{\boldsymbol{X_0}}$. $\hat{\boldsymbol{\sigma^2}}_{\boldsymbol{i}(\text{post})}$ and $\hat{\boldsymbol{\rho}}_{\boldsymbol{i}(\text{post})}$ express the models' predictions about the variance and temporal dependency of the fluctuation in each voxel in addition to the co-fluctuation. With these parameters estimated from training data, both the full and null models can marginalize the unknown $\boldsymbol{X}_{\boldsymbol{0}\text{test}}$ and yield their corresponding predictive log likelihoods for the testing data $\boldsymbol{Y}_{\text{test}}$. These log likelihoods are the basis for selecting between the full and null models.

To calculate the log likelihood, we notice that the predictive model of $\boldsymbol{Y}_{\text{res}}$ in testing data by both models are dynamical system models in which $\boldsymbol{X}_{\boldsymbol{0}\text{test}}$ is the latent state and $\boldsymbol{Y}_{\text{res}}$ is the observed data. They are slightly different from the standard dynamical system model [50] in that not only the latent states, but also the noise, have temporal dependency [51]:

$$\boldsymbol{X}^{(\boldsymbol{t})}_{\boldsymbol{0}\text{test}} \sim \boldsymbol{N}(\boldsymbol{X}^{(\boldsymbol{t-1})}_{\boldsymbol{0}\text{test}}\boldsymbol{T}_{\boldsymbol{X_0}}, \boldsymbol{V}_{\boldsymbol{\Delta X_0}}) \tag{27}$$

$$\begin{aligned}\boldsymbol{Y}^{(\boldsymbol{t})}_{\text{res}} - \boldsymbol{X}^{(\boldsymbol{t})}_{\boldsymbol{0}\text{test}}\hat{\boldsymbol{\beta}}_{\boldsymbol{0}\text{post}} \\ \sim \boldsymbol{N}((\boldsymbol{Y}^{(\boldsymbol{t})}_{\text{res}} - \boldsymbol{X}^{(\boldsymbol{t})}_{\boldsymbol{0}\text{test}}\hat{\boldsymbol{\beta}}_{\boldsymbol{0}\text{post}})\text{Diag}(\hat{\boldsymbol{\rho}}_{\text{post}}), \text{Diag}(\hat{\boldsymbol{\sigma^2}}_{\text{post}}))\end{aligned} \tag{28}$$

Where $\text{Diag}(\hat{\boldsymbol{\rho}}_{\text{post}})$ and $\text{Diag}(\hat{\boldsymbol{\sigma^2}}_{\text{post}})$ are diagonal matrices with vectors $\hat{\boldsymbol{\rho}}_{\text{post}}$ and $\hat{\boldsymbol{\sigma^2}}_{\text{post}}$ being their diagonal elements, respectively.

Because a modified forward-backward algorithm from the standard approach [50] is needed to calculate the preditive log likelihood $\boldsymbol{p}(\boldsymbol{Y}_{\text{res}}|\hat{\boldsymbol{\beta}}_{\boldsymbol{0}\text{post}}, \boldsymbol{T}_{\boldsymbol{X_0}}, \boldsymbol{V}_{\boldsymbol{\Delta X_0}}, \text{Diag}(\hat{\boldsymbol{\rho}}_{\text{post}}), \text{Diag}(\hat{\boldsymbol{\sigma^2}}_{\text{post}}))$ and the posterior distribution of $\boldsymbol{X}_{\boldsymbol{0}\text{test}}$, we describe the procedure below.

Define

$$\hat{\boldsymbol{G}}(\boldsymbol{X}^{(\boldsymbol{t})}_{\boldsymbol{0}\text{test}}) = \boldsymbol{p}(\boldsymbol{X}^{(\boldsymbol{t})}_{\boldsymbol{0}\text{test}}|\boldsymbol{Y}^{(\boldsymbol{1})}_{\text{res}}, \cdots, \boldsymbol{Y}^{(\boldsymbol{t})}_{\text{res}}) \tag{29}$$

$$\hat{\boldsymbol{H}}(\boldsymbol{X}^{(\boldsymbol{t})}_{\boldsymbol{0}\text{test}}) = \frac{\boldsymbol{p}(\boldsymbol{Y}^{(\boldsymbol{t+1})}_{\text{res}}, \cdots, \boldsymbol{Y}^{(\boldsymbol{n_T})}_{\text{res}}|\boldsymbol{X}^{(\boldsymbol{t})}_{\boldsymbol{0}\text{test}}, \boldsymbol{Y}^{(\boldsymbol{t})}_{\text{res}})}{\boldsymbol{p}(\boldsymbol{Y}^{(\boldsymbol{t+1})}_{\text{res}}, \cdots, \boldsymbol{Y}^{(\boldsymbol{n_T})}_{\text{res}}|\boldsymbol{Y}^{(\boldsymbol{1})}_{\text{res}}, \cdots, \boldsymbol{Y}^{(\boldsymbol{t})}_{\text{res}})}, \text{ for } \boldsymbol{t} < \boldsymbol{n_T} \tag{30}$$

and

$$\begin{aligned}\boldsymbol{c_t} &= \boldsymbol{p}(\boldsymbol{Y}^{(\boldsymbol{t})}_{\text{res}}|\boldsymbol{Y}^{(\boldsymbol{1})}_{\text{res}}, \cdots, \boldsymbol{Y}^{(\boldsymbol{t-1})}_{\text{res}}), \text{ for } \boldsymbol{t} > \boldsymbol{0} \\ \boldsymbol{c_1} &= \boldsymbol{p}(\boldsymbol{Y}^{(\boldsymbol{1})}_{\text{res}})\end{aligned} \tag{31}$$

Therefore, the cross-validated log likelihood is

$$\log \boldsymbol{p}(\boldsymbol{Y}^{(\boldsymbol{1})}_{\text{res}}, \cdots, \boldsymbol{Y}^{(\boldsymbol{n_T})}_{\text{res}}|\hat{\boldsymbol{\beta}}_{\boldsymbol{0}\text{post}}, \boldsymbol{T}_{\boldsymbol{X_0}}, \boldsymbol{V}_{\boldsymbol{\Delta X_0}}, \hat{\boldsymbol{\sigma^2}}_{\text{post}}, \hat{\boldsymbol{\rho}}_{\text{post}}) = \sum_{\boldsymbol{t=1}}^{\boldsymbol{n_T}} \log \boldsymbol{c_t} \tag{32}$$

It can be derived that the posterior distribution of $\boldsymbol{X}^{(\boldsymbol{t})}_{\boldsymbol{0}\text{test}}$ is

$$\boldsymbol{\gamma}(\boldsymbol{X}^{(\boldsymbol{t})}_{\boldsymbol{0}\text{test}}) = \boldsymbol{p}(\boldsymbol{X}^{(\boldsymbol{t})}_{\boldsymbol{0}\text{test}}|\boldsymbol{Y}^{(\boldsymbol{1})}_{\text{res}}, \cdots, \boldsymbol{Y}^{(\boldsymbol{n_T})}_{\text{res}}) = \hat{\boldsymbol{G}}(\boldsymbol{X}^{(\boldsymbol{t})}_{\boldsymbol{0}\text{test}})\hat{\boldsymbol{H}}(\boldsymbol{X}^{(\boldsymbol{t})}_{\boldsymbol{0}\text{test}}) \tag{33}$$

Below, we denote the mean and covariance of $\hat{G}(X_{0\text{test}}^{(t)})$ as $\mu_{X_0}^{(t)}$ and $\Gamma_{X_0}^{(t)}$, and the mean and covariance of $\gamma(X_{0\text{test}}^{(t)})$ as $\tilde{\mu}_{X_0}^{(t)}$ and $\tilde{\Gamma}_{X_0}^{(t)}$.

$\mu_{X_0}^{(t)}$, $\Gamma_{X_0}^{(t)}$ and c_t can be calculated by the forward step. $\tilde{\mu}_{X_0}^{(t)}$ and $\tilde{\Gamma}_{X_0}^{(t)}$ can be calculated by the backward step. To perform model selection, only forward step is necessary.

To perform the forward step, we first note that for $t = 1$

$$X_{0\text{test}}^{(1)} \sim N(0, V_{\Delta X_0}(I - T_{X_0}^2)^{-1}) \tag{34}$$

and

$$Y_{\text{res}}^{(1)} \sim N(X_{0\text{test}}^{(1)}\hat{\beta}_{0\text{post}}, \text{Diag}(\hat{\sigma^2}_{\text{post}})(I - \text{Diag}(\hat{\rho^2}_{\text{post}}))^{-1}) \tag{35}$$

Denote $V_{X_0} = V_{\Delta X_0}(I - T_{X_0}^2)^{-1}$, we have

$$\begin{aligned} c_1\hat{G}(X_{0\text{test}}^{(1)}) &= p(X_{0\text{test}}^{(1)}|Y_{\text{res}}^{(1)})p(Y_{\text{res}}^{(1)}) \\ &= p(X_{0\text{test}}^{(1)})p(Y_{\text{res}}^{(1)}|X_{0\text{test}}^{(1)}) \\ &= (2\pi)^{-\frac{n_0}{2}}|V_{X_0}|^{-\frac{1}{2}}\exp[-\frac{1}{2}X_{0\text{test}}^{(1)}V_{X_0}^{-1}{X_{0\text{test}}^{(1)}}^T] \\ &\quad \cdot \exp[-\frac{1}{2}\sum_{i=1}^{n_V}\frac{(Y_{\text{res}}^{(1)} - X_{0\text{test}}^{(1)}\hat{\beta}_{0\text{post}})^2(1-\rho_{i(\text{post})}^2)}{\sigma_{i(\text{post})}^2}]\prod_{i=1}^{n_V}(\frac{1-\rho_{i(\text{post})}^2}{\sigma_{i(\text{post})}^2})^{\frac{1}{2}} \end{aligned} \tag{36}$$

$\hat{G}(X_{0\text{test}}^{(1)})$ is a multivariate normal distribution of $X_{0\text{test}}^{(1)}$, we can find its covariance and mean from 36:

$$\Gamma_{X_0}^{(1)} = [V_{X_0}^{-1} + \hat{\beta}_{0\text{post}}(I - \text{Diag}(\hat{\rho^2}_{\text{post}}))\text{Diag}(\hat{\sigma^2}_{\text{post}})^{-1}\hat{\beta}_{0\text{post}}^T]^{-1} \tag{37}$$

$$\mu_{X_0}^{(1)} = Y_{\text{res}}^{(1)}(I - \text{Diag}(\hat{\rho^2}_{\text{post}}))\text{Diag}(\hat{\sigma^2}_{\text{post}})^{-1}\hat{\beta}_{0\text{post}}^T\Gamma_{X_0}^{(1)} \tag{38}$$

Because $\hat{G}(X_{0\text{test}}^{(1)})$ is a normalized probability distribution, the components in 36 after factoring out the multivariate normal distribution $\hat{G}(X_{0\text{test}}^{(1)})$ is c_1:

$$\begin{aligned} c_1 &= (2\pi)^{-\frac{n_V}{2}}|V_{X_0}|^{-\frac{1}{2}}|\Gamma_{X_0}^{(1)}|^{\frac{1}{2}}\prod_{i=1}^{n_V}(\frac{\hat{\sigma^2}_{i(\text{post})}}{1-\hat{\rho^2}_{i(\text{post})}})^{-\frac{1}{2}} \\ &= \exp\{-\frac{1}{2}[Y_{\text{res}}^{(1)}(I - \text{Diag}(\hat{\rho^2}_{\text{post}}))\text{Diag}(\hat{\sigma^2}_{\text{post}})^{-1}{Y_{\text{res}}^{(1)}}^T \\ &\quad - \mu_{X_0}^{(1)}{\Gamma_{X_0}^{(1)}}^{-1}{\mu_{X_0}^{(1)}}^T]\} \end{aligned} \tag{39}$$

For any $t > 1$, the following relation holds:

$$\begin{aligned} &c_t\hat{G}(X_{0\text{test}}^{(t)}) \\ &= \int p(Y_{\text{res}}^{(t)}|X_{0\text{test}}^{(t)}, X_{0\text{test}}^{(t-1)}, Y_{\text{res}}^{(t-1)})p(X_{0\text{test}}^{(t)}|X_{0\text{test}}^{(t-1)})\hat{G}(X_{0\text{test}}^{(t-1)})dX_{0\text{test}}^{(t-1)} \end{aligned} \tag{40}$$

$p(Y_{\text{res}}^{(t)}|X_{0\text{test}}^{(t)}, X_{0\text{test}}^{(t-1)}, Y_{\text{res}}^{(t-1)})$ is defined by 28. $p(X_{0\text{test}}^{(t)}|X_{0\text{test}}^{(t-1)})$ is defined by 27. Mean and covariance of $\hat{G}(X_{0\text{test}}^{(t-1)})$ are calculated by the previous step for $t - 1$. Therefore, by marginalizing $X_{0\text{test}}^{(t-1)}$, we obtain

$$\Gamma_{X_0}^{(t)} = (K_2 - J(K_1 + {\Gamma_{X_0}^{(t-1)}}^{-1})J^T)^{-1} \tag{41}$$

and

$$\boldsymbol{\mu}_{\boldsymbol{X_0}}^{(t)} = [\boldsymbol{\Delta Y}_{\text{res}}^{(t)}\text{Diag}(\hat{\boldsymbol{\sigma}^2}_{\text{post}})^{-1}\hat{\boldsymbol{\beta}}_{\boldsymbol{0}\text{post}}^{\boldsymbol{T}} + (\boldsymbol{\mu}_{\boldsymbol{X_0}}^{(t-1)}\boldsymbol{\Gamma}_{\boldsymbol{X_0}}^{(t-1)^{-1}} - \boldsymbol{\Delta Y}_{\text{res}}^{(t)}\text{Diag}(\hat{\boldsymbol{\sigma}^2}_{\text{post}})^{-1}\text{Diag}(\hat{\boldsymbol{\rho}}_{\text{post}})\hat{\boldsymbol{\beta}}_{\boldsymbol{0}\text{post}}^{\boldsymbol{T}})(\boldsymbol{K_1} + \boldsymbol{\Gamma}_{\boldsymbol{X_0}}^{(t-1)})^{-1}\boldsymbol{J}^{\boldsymbol{T}}]\boldsymbol{\Gamma}_{\boldsymbol{X_0}}^{(t)} \tag{42}$$

where $\boldsymbol{\Delta Y}_{\text{res}}^{(t)} = \boldsymbol{Y}_{\text{res}}^{(t)} - \boldsymbol{Y}_{\text{res}}^{(t-1)}\text{Diag}(\hat{\boldsymbol{\rho}}_{\text{post}})$.
$\boldsymbol{J} = \boldsymbol{V}_{\boldsymbol{\Delta X_0}}^{-1}\boldsymbol{T}_{\boldsymbol{X_0}}^{\boldsymbol{T}} + \hat{\boldsymbol{\beta}}_{\boldsymbol{0}\text{post}}\text{Diag}(\hat{\boldsymbol{\sigma}^2}_{\text{post}})^{-1}\text{Diag}(\hat{\boldsymbol{\rho}}_{\text{post}})\hat{\boldsymbol{\beta}}_{\boldsymbol{0}\text{post}}^{\boldsymbol{T}}$,
$\boldsymbol{K_1} = \boldsymbol{T}_{\boldsymbol{X_0}}\boldsymbol{V}_{\boldsymbol{\Delta X_0}}^{-1}\boldsymbol{T}_{\boldsymbol{X_0}}^{\boldsymbol{T}} + \hat{\boldsymbol{\beta}}_{\boldsymbol{0}\text{post}}\text{Diag}(\hat{\boldsymbol{\sigma}^2}_{\text{post}})^{-1}\text{Diag}(\hat{\boldsymbol{\rho}}_{\text{post}})^{\boldsymbol{2}}\hat{\boldsymbol{\beta}}_{\boldsymbol{0}\text{post}}^{\boldsymbol{T}}$ and
$\boldsymbol{K_2} = \boldsymbol{V}_{\boldsymbol{\Delta X_0}}^{-1} + \hat{\boldsymbol{\beta}}_{\boldsymbol{0}\text{post}}\text{Diag}(\hat{\boldsymbol{\sigma}^2}_{\text{post}})^{-1}\hat{\boldsymbol{\beta}}_{\boldsymbol{0}\text{post}}^{\boldsymbol{T}}$. Note that $\boldsymbol{J}$, $\boldsymbol{K_1}$, $\boldsymbol{K_2}$ are all constants.

Similarly to 39, after factoring out $\hat{\boldsymbol{G}}(\boldsymbol{X}_{\boldsymbol{0}\text{test}}^{(t)})$, we obtain

$$\begin{aligned}\boldsymbol{c_t} = (\boldsymbol{2\pi})^{-\frac{n_V}{2}}|\boldsymbol{K_1} + \boldsymbol{\Gamma}_{\boldsymbol{X_0}}^{(t-1)}|^{-\frac{1}{2}}|\boldsymbol{V}_{\boldsymbol{\Delta X_0}}|^{-\frac{1}{2}}|\boldsymbol{\Gamma}_{\boldsymbol{X_0}}^{(t-1)}|^{-\frac{1}{2}}|\boldsymbol{\Gamma}_{\boldsymbol{X_0}}^{(t)}|^{\frac{1}{2}}\prod_{\boldsymbol{i=1}}^{\boldsymbol{n_V}}\boldsymbol{\sigma}_{\boldsymbol{i}\text{post}}^{\boldsymbol{-1}} \\ \cdot\exp[-\frac{\boldsymbol{1}}{\boldsymbol{2}}\boldsymbol{\mu}_{\boldsymbol{X_0}}^{(t-1)}\boldsymbol{\Gamma}_{\boldsymbol{X_0}}^{(t-1)^{-1}}\boldsymbol{\mu}_{\boldsymbol{X_0}}^{(t-1)^{T}} + \frac{\boldsymbol{1}}{\boldsymbol{2}}\boldsymbol{\mu}_{\boldsymbol{X_0}}^{(t)}\boldsymbol{\Gamma}_{\boldsymbol{X_0}}^{(t)^{-1}}\boldsymbol{\mu}_{\boldsymbol{X_0}}^{(t)^{T}} \\ -\frac{\boldsymbol{1}}{\boldsymbol{2}}\boldsymbol{\Delta Y}_{\text{res}}^{(t)}\text{Diag}(\hat{\boldsymbol{\sigma}^2}_{\text{post}})^{-1}\boldsymbol{\Delta Y}_{\text{res}}^{(t)^{T}} + \frac{\boldsymbol{1}}{\boldsymbol{2}}(\boldsymbol{\mu}_{\boldsymbol{X_0}}^{(t-1)}\boldsymbol{\Gamma}_{\boldsymbol{X_0}}^{(t-1)^{-1}} \\ -\boldsymbol{\Delta Y}_{\text{res}}^{(t)}\text{Diag}(\hat{\boldsymbol{\sigma}^2}_{\text{post}})^{-1}\text{Diag}(\hat{\boldsymbol{\rho}}_{\text{post}})\hat{\boldsymbol{\beta}}_{\boldsymbol{0}\text{post}}^{\boldsymbol{T}})(\boldsymbol{K_1} + \boldsymbol{\Gamma}_{\boldsymbol{X_0}}^{(t-1)})^{-1} \\ (\boldsymbol{\mu}_{\boldsymbol{X_0}}^{(t-1)}\boldsymbol{\Gamma}_{\boldsymbol{X_0}}^{(t-1)^{-1}} - \boldsymbol{\Delta Y}_{\text{res}}^{(t)}\text{Diag}(\hat{\boldsymbol{\sigma}^2}_{\text{post}})^{-1}\text{Diag}(\hat{\boldsymbol{\rho}}_{\text{post}})\hat{\boldsymbol{\beta}}_{\boldsymbol{0}\text{post}}^{\boldsymbol{T}})^{T}]\end{aligned} \tag{43}$$

By calculating 41, 42 and 43 recursively with $\boldsymbol{t}$ incremented by 1 until $\boldsymbol{n_T}$, the predictive log likelihood 32 of both the full and null models can be calculated to serve as the basis of model selection.

To calculate the mean and variance of the posterior distribution $\boldsymbol{\gamma}(\boldsymbol{X_{0\text{test}}}^{(t)})$ of $\boldsymbol{X}_{\boldsymbol{0}\text{test}}$, backward step is needed. We denote its mean as $\hat{\boldsymbol{\mu}}_{\boldsymbol{X_0}}^{(t)}$, and covariance as $\hat{\boldsymbol{\Gamma}}_{\boldsymbol{X_0}}^{(t)}$.

For any $\boldsymbol{t} < \boldsymbol{n_T}$, it can be derived that

$$\boldsymbol{c_{t+1}}\hat{\boldsymbol{H}}(\boldsymbol{X}_{\boldsymbol{0}\text{test}}^{(t)}) = \int \hat{\boldsymbol{H}}(\boldsymbol{X}_{\boldsymbol{0}\text{test}}^{(t+1)})\boldsymbol{p}(\boldsymbol{X}_{\boldsymbol{0}\text{test}}^{(t+1)}|\boldsymbol{X}_{\boldsymbol{0}\text{test}}^{(t)})\boldsymbol{p}(\boldsymbol{Y}_{\text{res}}^{(t+1)}|\boldsymbol{X}_{\boldsymbol{0}\text{test}}^{(t)}, \boldsymbol{X}_{\boldsymbol{0}\text{test}}^{(t+1)}, \boldsymbol{Y}_{\text{res}}^{(t)})\boldsymbol{dX}_{\boldsymbol{0}\text{test}}^{(t+1)} \tag{44}$$

By plugging in 33, we get

$$\boldsymbol{\gamma}(\boldsymbol{X_{0\text{test}}}^{(t)}) = \frac{\hat{\boldsymbol{G}}(\boldsymbol{X}_{\boldsymbol{0}\text{test}}^{(t)})}{\boldsymbol{c_{t+1}}} \cdot \int \frac{\boldsymbol{\gamma}(\boldsymbol{X_{0\text{test}}}^{(t+1)})}{\hat{\boldsymbol{G}}(\boldsymbol{X}_{\boldsymbol{0}\text{test}}^{(t+1)})}\boldsymbol{p}(\boldsymbol{X}_{\boldsymbol{0}\text{test}}^{(t+1)}|\boldsymbol{X}_{\boldsymbol{0}\text{test}}^{(t)})\boldsymbol{p}(\boldsymbol{Y}_{\text{res}}^{(t+1)}|\boldsymbol{X}_{\boldsymbol{0}\text{test}}^{(t)}, \boldsymbol{X}_{\boldsymbol{0}\text{test}}^{(t+1)}, \boldsymbol{Y}_{\text{res}}^{(t)})\boldsymbol{dX}_{\boldsymbol{0}\text{test}}^{(t+1)} \tag{45}$$

After the marginalization in 45 and observing the terms related to $\boldsymbol{X_{0\text{test}}}^{(t)}$, we get the recursive relations

$$\hat{\boldsymbol{\Gamma}}_{\boldsymbol{X_0}}^{(t)} = (\boldsymbol{\Gamma}_{\boldsymbol{X_0}}^{(t)^{-1}} + \boldsymbol{K_1} - \boldsymbol{J}^{\boldsymbol{T}}(\hat{\boldsymbol{\Gamma}}_{\boldsymbol{X_0}}^{(t+1)^{-1}} - \boldsymbol{\Gamma}_{\boldsymbol{X_0}}^{(t+1)^{-1}} + \boldsymbol{K_2})^{-1}\boldsymbol{J})^{-1} \tag{46}$$

and

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{\boldsymbol{X_0}}^{(t)} = [\boldsymbol{\mu}_{\boldsymbol{X_0}}^{(t)}\boldsymbol{\Gamma}_{\boldsymbol{X_0}}^{(t)^{-1}} - \boldsymbol{\Delta Y}_{\text{res}}^{(t+1)}\text{Diag}(\hat{\boldsymbol{\sigma}^2}_{\text{post}})^{-1}\text{Diag}(\hat{\boldsymbol{\rho}}_{\text{post}})\hat{\boldsymbol{\beta}}_{\boldsymbol{0}\text{post}}^{\boldsymbol{T}} \\ + (\hat{\boldsymbol{\mu}}_{\boldsymbol{X_0}}^{(t+1)}\hat{\boldsymbol{\Gamma}}_{\boldsymbol{X_0}}^{(t+1)^{-1}} - \boldsymbol{\mu}_{\boldsymbol{X_0}}^{(t+1)}\boldsymbol{\Gamma}_{\boldsymbol{X_0}}^{(t+1)^{-1}} + \boldsymbol{\Delta Y}_{\text{res}}^{(t+1)}\text{Diag}(\hat{\boldsymbol{\sigma}^2}_{\text{post}})^{-1}\hat{\boldsymbol{\beta}}_{\boldsymbol{0}\text{post}}^{\boldsymbol{T}}) \\ (\hat{\boldsymbol{\Gamma}}_{\boldsymbol{X_0}}^{(t+1)^{-1}} - \boldsymbol{\Gamma}_{\boldsymbol{X_0}}^{(t+1)^{-1}} + \boldsymbol{K_2})^{-1}\boldsymbol{J}]\hat{\boldsymbol{\Gamma}}_{\boldsymbol{X_0}}^{(t)}\end{aligned} \tag{47}$$

Note that $\gamma(\boldsymbol{X_{0\text{test}}}^{(\boldsymbol{n_T})}) = \hat{\boldsymbol{G}}(\boldsymbol{X}_{\boldsymbol{0}\text{test}}^{(\boldsymbol{n_T})})$, therefore $\hat{\boldsymbol{\mu}}_{\boldsymbol{X_0}}^{(\boldsymbol{n_T})} = \boldsymbol{\mu}_{\boldsymbol{X_0}}^{(\boldsymbol{n_T})}$ and $\hat{\boldsymbol{\Gamma}}_{\boldsymbol{X_0}}^{(\boldsymbol{n_T})} = \boldsymbol{\Gamma}_{\boldsymbol{X_0}}^{(\boldsymbol{n_T})}$. By recursively calculating 46 and 47 with $\boldsymbol{t}$ decremented by 1 from $\boldsymbol{n_T - 1}$ until 1, the posterior distribution of $\boldsymbol{X}_{\boldsymbol{0}\text{test}}^{(\boldsymbol{t})}$ given all the testing data can be calculated.

For decoding purpose, we need to obtain not only the posterior mean of intrinsic fluctuations $\boldsymbol{X}_{\boldsymbol{0}\text{test}}^{(\boldsymbol{t})}$, but also the task-related activity $\boldsymbol{X}_{\text{test}}^{(\boldsymbol{t})}$. Therefore, we do not subtract a predicted signal $\boldsymbol{X}_{\text{test}}\hat{\boldsymbol{\beta}}_{\text{post}}$ based on a hypothetical design matrix from testing data $\boldsymbol{Y}_{\text{test}}$. We perform the forward-backward algorithm on $\boldsymbol{Y}_{\text{test}}$ directly. By replacing $\hat{\boldsymbol{\beta}}_{\boldsymbol{0}\text{post}}$ in the equations from 27 to 47 with $[\hat{\boldsymbol{\beta}}_{\text{post}}^{\boldsymbol{T}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{0}\text{post}}^{\boldsymbol{T}}]^{\boldsymbol{T}}$ and other related terms accordingly, the posterior mean of both $\boldsymbol{X}_{\text{test}}^{(\boldsymbol{t})}$ and $\boldsymbol{X}_{\boldsymbol{0}\text{test}}^{(\boldsymbol{t})}$ can be decoded just as $\boldsymbol{X}_{\boldsymbol{0}\text{test}}^{(\boldsymbol{t})}$ is decoded in 47.

Data processing and analysis

Data used in Fig **1B** are from the experiments of Schuck et al. [22], following the same preprocessing procedure as the original study. The fMRI data were acquired at TR=2.4s. Data of 24 participants were used. Their design matrices were used for all the following analyses and simulations. Data in Fig **1E,G** and Fig **3** were preprocessed data obtained from Human Connectome Project (HCP) [33]. The first 24 participants who have completed all 3T protocols and whose data were acquired in quarter 8 of the HCP acquiring period without image quality issues were selected for analysis in Fig 3. Data from 864 participants without image quality issues in HCP were used in the analysis in Fig 5. Each participants in the HCP data have 2 runs of resting state data with posterior-anterior phase encoding direction and 2 runs with anterior-posterior phase encoding direction. Time series were resampled at the same TR as the design matrix before further analysis.

$\hat{\boldsymbol{\beta}}$ point estimates in Fig **1** were obtained with AFNI's *3ddeconvolve* [52]. The design matrices were set up by convolving the stereotypical double-Gamma HRF in SPM [53] with event time courses composed with impulses lasting for the duration of the participants' reaction time. AR(1) coefficients in Fig **1G** were estimated after upsampling the fMRI time series in the HCP data to the TR in Schuck et al. [22] and linear detrending. Upsampling is to reflect the lower temporal resolution more typically employed in task-related fMRI studies.

In the experiments of Fig **3**, lateral occipital cortex was chosen as the ROI, which included $\boldsymbol{4804 \pm 29}$ (mean $\pm$ standard deviation) voxels. Task related signals were only added to voxels within a bounding box of which the coordinates satisfy $\boldsymbol{25 < x < 35}$, $\boldsymbol{-95 < y < -5}$ and $\boldsymbol{-15 < z < 5}$. $\boldsymbol{189.0 \pm 0.2}$ voxels fell within this bounding box. $\boldsymbol{\beta}$ were simulated according to the covariance matrix in Fig **3A** and scaled by one values in 1, 2, 4, 8. To evaluate the performance of the recovered correlation structure by different methods, the correlation between the off-diagonal elements of the recovered similarity matrix from data of each simulated participant was correlated with those elements of the ideal similarity matrix to yield the top panel of Fig **3H**. The top panel reflects the correlation of individual results. The bottom panel reflects the correlation of average results over simulated participants.

In order to make fair comparison with BRSA which considers temporal auto-correlation in noise, all the point estimates of $\hat{\boldsymbol{\beta}}$ in Fig **3** were performed with restricted maximum likelihood estimation. AR(1) parameters of each voxel were estimated after initial regular regression. The AR(1) parameters were used to re-compute the temporal noise covariance matrices for each voxel and $\hat{\boldsymbol{\beta}}$ were calculated again assuming these noise covariance matrices. When spatial whitening of $\hat{\boldsymbol{\beta}}$ were performed, it followed the procedure of Diedrichsen et al. [17]. Point estimates of the

spatial covariance of noise were first calculated from residuals of regression. These are not full rank matrices due to large numbers of voxels. The off-diagonal elements were further shrunk by weighting the point estimate of the spatial noise covariance structure by 0.4 and a diagonal covariance matrix with the same diagonal elements as the point estimate covariance matrix by 0.6.

To simulate the fMRI noise in Fig **4**, we first estimated the number of principal components to describe the spatial noise correlation in the 24 resting state fMRI data from HCP databse using the algoritm of Gavish and Donoho [31]. The spatial patterns of these principal components were kept fixed as the modulation magnitude $\boldsymbol{\beta_0}$ by the intrinsic fluctuation. AR(1) parameters for each voxel's spatially indepndent noise were estimated from the residuals after subtrating these principal components. For each simulated subject, time courses of intrinsic flucutations were newly simulated by scrambling the phase of the Fourier transformation of the $\boldsymbol{X_0}$ estimated from the real data, thus preserving the amplitudes of their frequency spectrum. AR(1) noise were then added to each voxel with the same parameters as estimated from the real data. To speed up the simulation, only 200 random voxels from the ROI in Fig **3B** were kept for each participant in these simulations. Among them, 100 random voxels were added with simulated task-related signals. Thus, each simulated participant has different spatial patterns of $\boldsymbol{\beta_0}$ due to the random selection of voxels. 500 simulated datasets were generated based on the real data of each participant, for each of the three SNR levels. In total 36000 subjects were simulated. The simulated pool of subjects were sub-divided into bins with a fixed number of simulated subjects ranging from 24 to 1200. The mean and standard deviation of the correlation between the true similarity matrix and the average similarity matrix based on the subjects in each bin were calculated, and plotted in Fig **4A**.

All SNRs in Fig **3** and Fig **4** were calculated post hoc, using the standard deviation of the added signals in the bounding box region devided by the standard deviation of the noise in each voxel, and averaged across voxels and simulated subjects for each level of simulation.

Acknowledgments

This work was funded by the Intel Corporation and the John Templeton Foundation (15541). The opinions expressed are those of the authors and do not necessarily reflect the views of the John Templeton Foundation or of Intel Corporation. JWP was supported by grants from the McKnight Foundation, Simons Collaboration on the Global Brain (SCGB AWD1004351) and the NSF CAREER Award (IIS-1150186). Resting state fMRI data used in several experiments were obtained from the MGH-USC Human Connectome Project (HCP) database.

References

1. Ogawa S, Lee TM, Kay AR, Tank DW. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. Proceedings of the National Academy of Sciences. 1990;87(24):9868–9872.

2. Boynton GM, Engel SA, Glover GH, Heeger DJ. Linear systems analysis of functional magnetic resonance imaging in human V1. Journal of Neuroscience. 1996;16(13):4207–4221.

3. Jazzard P, Matthews P, Smith S. Functional magnetic resonance imaging: An introduction to methods. Oxford Univ. Press, 404pp; 2003.

4. Uğurbil K, Xu J, Auerbach EJ, Moeller S, Vu AT, Duarte-Carvajalino JM, et al. Pushing spatial and temporal resolution for functional and diffusion MRI in the Human Connectome Project. Neuroimage. 2013;80:80–104.

5. Cabeza R, Nyberg L. Imaging cognition II: An empirical review of 275 PET and fMRI studies. Journal of cognitive neuroscience. 2000;12(1):1–47.

6. Haxby JV. Multivariate pattern analysis of fMRI: the early beginnings. Neuroimage. 2012;62(2):852–855.

7. Haxby JV, Connolly AC, Guntupalli JS. Decoding neural representational spaces using multivariate pattern analysis. Annual review of neuroscience. 2014;37:435–456.

8. Kriegeskorte N, Mur M, Bandettini PA. Representational similarity analysis-connecting the branches of systems neuroscience. Frontiers in systems neuroscience. 2008;2:4.

9. Xue G, Dong Q, Chen C, Lu Z, Mumford JA, Poldrack RA. Greater neural pattern similarity across repetitions is associated with better memory. Science. 2010;330(6000):97–101.

10. Ritchey M, Wing EA, LaBar KS, Cabeza R. Neural similarity between encoding and retrieval is related to memory via hippocampal interactions. Cerebral Cortex. 2012; p. bhs258.

11. Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, et al. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron. 2008;60(6):1126–1141.

12. Connolly AC, Guntupalli JS, Gors J, Hanke M, Halchenko YO, Wu YC, et al. The representation of biological classes in the human brain. The Journal of Neuroscience. 2012;32(8):2608–2618.

13. Iordan MC, Greene MR, Beck DM, Fei-Fei L. Basic level category structure emerges gradually across human ventral visual cortex. Journal of cognitive neuroscience. 2015;.

14. Henriksson L, Khaligh-Razavi SM, Kay K, Kriegeskorte N. Visual representations are dominated by intrinsic fluctuations correlated between areas. NeuroImage. 2015;114:275–286.

15. Alink A, Walther A, Krugliak A, van den Bosch JJ, Kriegeskorte N. Mind the drift-improving sensitivity to fMRI pattern information by accounting for temporal pattern drift. bioRxiv. 2015; p. 032391.

16. Diedrichsen J, Ridgway GR, Friston KJ, Wiestler T. Comparing the similarity and spatial structure of neural representations: a pattern-component model. Neuroimage. 2011;55(4):1665–1678.

17. Diedrichsen J, Provost S, Zareamoghaddam H. On the distribution of cross-validated Mahalanobis distances. arXiv preprint arXiv:160701371. 2016;.

18. Friston KJ, Jezzard P, Turner R. Analysis of functional MRI time-series. Human brain mapping. 1994;1(2):153–171.

19. Cai MB, Schuck NW, Pillow JW, Niv Y. A Bayesian method for reducing bias in neural representational similarity analysis. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R, editors. Advances in Neural Information Processing Systems 29. Curran Associates, Inc.; 2016. p. 4951–4959. Available from: `https://goo.gl/JZv6mu`.

20. Haxby JV, Guntupalli JS, Connolly AC, Halchenko YO, Conroy BR, Gobbini MI, et al. A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron. 2011;72(2):404–416.

21. Chen PHC, Chen J, Yeshurun Y, Hasson U, Haxby J, Ramadge PJ. A Reduced-Dimension fMRI Shared Response Model. In: Advances in Neural Information Processing Systems; 2015. p. 460–468.

22. Schuck NW, Cai MB, Wilson RC, Niv Y. Human orbitofrontal cortex represents a cognitive map of state space. Neuron. 2016;91:1–11.

23. Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika. 1964;29(1):1–27.

24. Zarahn E, Aguirre GK, D'Esposito M. Empirical analyses of BOLD fMRI statistics. NeuroImage. 1997;5(3):179–197.

25. Woolrich MW, Ripley BD, Brady M, Smith SM. Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage. 2001;14(6):1370–1386.

26. Van Essen DC, Smith SM, Barch DM, Behrens TE, Yacoub E, Ugurbil K, et al. The WU-Minn human connectome project: an overview. Neuroimage. 2013;80:62–79.

27. Kay KN, Naselaris T, Prenger RJ, Gallant JL. Identifying natural images from human brain activity. Nature. 2008;452(7185):352–355.

28. Robbins H. An empirical Bayes approach to statistics. COLUMBIA UNIVERSITY New York City United States; 1956.

29. Kay KN, Rokem A, Winawer J, Dougherty RF, Wandell BA. GLMdenoise: a fast, automated technique for denoising task-based fMRI data. Frontiers in neuroscience. 2013;7.

30. Poldrack RA, Mumford JA, Nichols TE. Handbook of functional MRI data analysis. Cambridge University Press; 2011.

31. Gavish M, Donoho DL. The optimal hard threshold for singular values is $\mathbf{4}/\sqrt{\mathbf{3}}$. IEEE Transactions on Information Theory. 2014;60(8):5040–5053.

32. Minka TP. Automatic choice of dimensionality for PCA. In: Advances in neural information processing systems; 2001. p. 598–604.

33. Glasser MF, Sotiropoulos SN, Wilson JA, Coalson TS, Fischl B, Andersson JL, et al. The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage. 2013;80:105–124.

34. Walther A, Nili H, Ejaz N, Alink A, Kriegeskorte N, Diedrichsen J. Reliability of dissimilarity measures for multi-voxel pattern analysis. NeuroImage. 2015;.

35. Ramírez FM, Cichy RM, Allefeld C, Haynes JD. The neural code for face orientation in the human fusiform face area. Journal of Neuroscience. 2014;34(36):12155–12167.

36. Ramírez FM. Representational confusion: the plausible consequence of demeaning your data. bioRxiv. 2017; p. 195271.

37. Diedrichsen J, Yokoi A, Arbuckle S. Pattern Component Modeling: A Flexible Approach For Understanding The Representational Structure Of Brain Activity Patterns. bioRxiv. 2017;doi:10.1101/120584.

38. Diedrichsen J, Yokoi A, Arbuckle SA. Pattern component modeling: A flexible approach for understanding the representational structure of brain activity patterns. NeuroImage. 2017;.

39. Handwerker DA, Ollinger JM, D'Esposito M. Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. Neuroimage. 2004;21(4):1639–1651.

40. Aguirre GK, Zarahn E, D'esposito M. The variability of human, BOLD hemodynamic responses. Neuroimage. 1998;8(4):360–369.

41. Squire LR, Ojemann JG, Miezin FM, Petersen SE, Videen TO, Raichle ME. Activation of the hippocampus in normal humans: a functional anatomical study of memory. Proceedings of the National Academy of Sciences. 1992;89(5):1837–1841.

42. Desimone R. Neural mechanisms for visual memory and their role in attention. Proceedings of the National Academy of Sciences. 1996;93(24):13494–13499.

43. Henson R, Shallice T, Dolan R. Neuroimaging evidence for dissociable forms of repetition priming. Science. 2000;287(5456):1269–1272.

44. O'Craven KM, Rosen BR, Kwong KK, Treisman A, Savoy RL. Voluntary attention modulates fMRI activity in human MT–MST. Neuron. 1997;18(4):591–598.

45. Wojciulik E, Kanwisher N, Driver J. Covert visual attention modulates face-specific activity in the human fusiform gyrus: fMRI study. Journal of Neurophysiology. 1998;79(3):1574–1578.

46. Broyden C. A new double-rank minimisation algorithm. Preliminary report. In: Notices of the American Mathematical Society. vol. 16. AMER MATHEMATICAL SOC 201 CHARLES ST, PROVIDENCE, RI 02940-2213; 1969. p. 670.

47. Fletcher R. A new approach to variable metric algorithms. The computer journal. 1970;13(3):317–322.

48. Goldfarb D. A family of variable-metric methods derived by variational means. Mathematics of computation. 1970;24(109):23–26.

49. Shanno DF. Conditioning of quasi-Newton methods for function minimization. Mathematics of computation. 1970;24(111):647–656.

50. Bishop CM. Pattern recognition and machine learning. springer; 2006.

51. Ephraim Y, Malah D, Juang BH. On the application of hidden Markov models for enhancing noisy speech. IEEE Transactions on Acoustics, Speech, and Signal Processing. 1989;37(12):1846–1856.

52. Cox RW. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Computers and Biomedical research. 1996;29(3):162–173.

53. Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RS. Statistical parametric maps in functional imaging: a general linear approach. Human brain mapping. 1994;2(4):189–210.