Reverse Double-Dipping: When Data Dips You, Twice—Stimulus-Driven Information Leakage in Naturalistic Neuroimaging

Seung-Goo Kim[1]

***For correspondence:**
seung-goo.kim@ae.mpg.de (SGK)

[1] Research Group Neurocognition of Music and Language, Max Planck Institute for Empirical Aesthetics, Grüneburgweg 14, 60322 Frankfurt, Germany

Abstract This article elucidates a methodological pitfall of cross-validation for evaluating predictive models applied to naturalistic neuroimaging data—namely, 'reverse double-dipping' (RDD). In a broader context, this problem is also known as 'leakage in training examples', which is difficult to detect in practice. RDD can occur when predictive modeling is applied to data from a conventional neuroscientific design, characterized by a limited set of stimuli repeated across trials and/or participants. It results in spurious predictive performances due to overfitting to repeated signals, even in the presence of independent noise. Through comprehensive simulations and real-world examples following theoretical formulation, the article underscores how such information leakage can occur and how severely it could compromise the results and conclusions when it is combined with widely spread informal reverse inference. The article concludes with practical recommendations for researchers to avoid RDD in their experiment design and analysis.

Introduction

Recent advancement of 'naturalistic neuroimaging' has opened up new exciting developments in various fields of human neuroscience (***Sonkusare et al., 2019***; ***Nastase et al., 2020***; ***Hamilton and Huth, 2020***). The key idea of naturalistic neuroimaging is that experiments with naturalistic (i.e., real-world) stimuli are essential when investigating complex human behaviors. This latest reincarnation of ecological psychology (***Brunswik, 1943***; ***Gibson, 1978***) has gained wide popularity in psychology and human neuroscience. In particular, this idea has been widely adopted in domains where high-order cognitive and/or affective processes are involved, and a simple contrastive experimental approach can explain only little. For example, comparing brain responses to *music* vs. *non-music* to find neural correlates of "music perception" may be under an overly reduction-

ist assumption (i.e., "music-as-fixed-effect" fallacy; ***Kim, 2022***) that the human brain is governed by simple, interpretable rules that can extrapolate to explain complex behaviors (for more discussion, see ***Nastase et al., 2020***).

Currently, one of the most popular frameworks in analyzing naturalistic neuroimaging data is known as 'linearized encoding analysis' (for a comprehensive review of the encoding models in the music domain, see ***Kim, 2022***). This is a method that identifies how of a time-invariant system (e.g., the human brain or an artificial neural network [ANN]) transforms information from a stimulus to a response, by separating separates the whole transformation into non-linear mapping and linear mapping (***Wu et al., 2006***; ***Naselaris et al., 2011***). Often, the non-linear mapping ('linearization') is given by the researcher's hypothesis (i.e., *"the human language system utilizes information X and Y but not Z"*) and the linear mapping is found from the given linearization and human responses. This approach has been more widely used in the era of deep neural networks. For example, studies have shown the potentials of transfer learning of image classification models (***Allen et al., 2022***) or large-language models (***Caucheteux et al., 2023***) to explain human behaviors and neural activity.

Besides encoding analysis, model-free approaches have been popularized. For example, the synchrony of neural activity over time across multiple participants when watching an identical naturalistic stimulus (i.e., the same movie) has been suggested to quantify stimulus-driven effects (***Hasson et al., 2004***). Without explicit modeling of encoded information, the similarity in responses over multiple repetitions can attribute repeated responses to the repeated stimuli—no matter which information is being processed. The analysis of neural synchrony draws conclusions from strong correlations between the neural responses of different participants or trials while presenting identical stimuli that the certain (but unspecified) information in the stimulus evoked the time-locked neural responses.

More recently, data-driven segmentation algorithms such as hidden markov models have been proposed to model high-level concepts such as *event boundaries* in narratives (***Baldassano et al., 2018***). For model-free approaches, the repetition of identical stimuli is the key reference of the analysis. Thus, many naturalistic neuroscience experiments inspired by such approaches presented identical stimuli to multiple participants.

More broadly, stimulus repetition remains a cornerstone of experimental design in neuroscience. From the long tradition of event-related potential (ERP) experiments and the block-design in functional magnetic resonance imaging (fMRI), repetition has been the gold standard for cancelling out non-time-locked noise and isolating the time-locked signal.

However, in encoding analysis, time-locked responses to repeated stimuli, even with independent noise, could introduce **information leakage**, disabling regularization, ultimately leading to a false conclusion where irrelevant information is mistaken as relevant to human neural processes. This paper explains how such a fallacy could occur in the Theory section, identifies contributing factors based on simulations in the Simulation section, and demonstrates real-world cases in the Real Data section. Finally, implications for future analyses and experiments are discussed and practical recommendations are given in the Discussion section.

Theory

Types of information leakage

In essence, information leakage is a circular fallacy. Leakage in data mining and machine learning (or more generally, in predictive modeling) has been categorized into multiple cases: (i) *leaking features*, (ii) *leakage by design decisions*, (iii) *leakage in training examples* (***Kaufman et al., 2012***).

In data-mining competitions, the training data (a set of feature-target pairs) and the test (hold-out) data (an independent set of feature-target pairs) are separated by the organizer and only the training data and test features are provided to contenders. Contenders create their predictive models and submit their predictions on the test targets so that the organizer can compare their hold-out validation performances objectively, under the assumption that the curated data is highly representative of real-world problems.

In cross-validation, one individual researcher plays both the role of an *organizer* (i.e., partitioning data at hand into training and test sets) and that of a *contender* (i.e., modeling associations between feature and response in the training set and predicting responses from features in the test set). The *leaking features* could be information that is seemingly random but contains highly predictive information by accidental association (or overlooked failure of dissociation) that exist in both the training and test sets. Although in natural science no human organizer will create a dataset, mistaking confounding features as predictive or even causal is not unheard of (***Nuzzo, 2015***).

An once prevalent—but still not uncommon—case of the circular fallacy in neuroscience is a selective analysis based on the same data, which is widely known as *double-dipping* (***Kriegeskorte et al., 2009***). This can be understood as *leakage by design decisions*. Double-dipping can also occur in the context of cross-validation. For example, selecting features from the entire dataset prior to splitting (i.e., using information from both of a training set and a test set) would leak information (including noise) from the test set to the training set. In this case, it is you who dips the data into analysis twice.

The last kind, *leakage in training examples*, occurs due to an incidental similarity between training and test sets. This is also known as *twinning*: i.e., having one individual participant in a training set and their twin in a test set. If a researcher fails to recognize the true relationship between the twins, any random association stemming from their shared characteristics may spuriously appear predictive. In this case, *it is the data that dips you, twice*—thus, it can be called **reverse double-dipping** (RDD; ***Figure 1***). In the following, I will formally illustrate when/ RDD occurs, especially in the con-

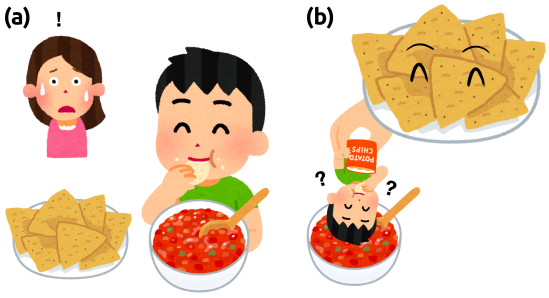


Figure 1. (a) In double-dipping, you dip an identical data point twice (the same stimulus and identical noise). **(b)** In reverse double-dipping, the dataset dips you twice, unbeknownst to you, with non-identical data points (the same stimulus but independent noise).

text of systems neuroscience.

Finite impulse response model

A finite impulse response (FIR) model is commonly used to identify a time-invariant linear system. Please see ***Table 1*** for the definitions of variables and notations used in this article. Let us consider an FIR model:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^{T\times 1}$ is a response vector over T time points of a single response unit (e.g., a channel or a voxel), $\mathbf{X} \in \mathbb{R}^{T\times FD}$ is a nonzero Toeplitz design matrix of F features[1] with D delays such that $\|\mathbf{x}_{(i)}^{\mathsf{T}}\mathbf{x}_{(i)}\| > 0$ for all $i \in \{1, \cdots, FD\}$, $\mathbf{x}_{(i)}$ is the i-th column of $\mathbf{X}$, $\|\cdot\|$ denotes the l_2-norm, which leads to $\|\mathbf{X}^{\mathsf{T}}\mathbf{X}\|_{\mathrm{F}} > 0$ where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. $\mathbf{b} \in \mathbb{R}^{FD\times 1}$ is an unknown weight vector (often called a temporal response function; or more generally a transfer function), $\mathbf{e} \in \mathbb{R}^{T\times 1}$ is a zero-mean, unit-variance Gaussian noise vector $\mathbf{e} \sim \mathcal{N}_T(\mathbf{0}, \mathbf{I}_T)$ where $\mathbf{I}_T \in \mathbb{R}^{T\times T}$ is an identity matrix. For convenience, we further assume that we standardize predictors and response variables prior to analysis so that their sample means are zero and sample variances are one.

Here, let us assume that we have access to the true weights, which are nonzero (i.e., $\|\mathbf{b}\| > 0$). With this model, we can generate a training dataset and a test dataset using the identical weights $\mathbf{b}$ but independent predictors and noise:

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{e}_i = \mathbf{s}_i + \mathbf{e}_i, \tag{2}$$

where $(\cdot)_i$ denotes the i-th independent partition in cross-validation with $i = 1$ for a training set and $i = 2$ for a test set. The true signal is denoted as $\mathbf{s}_i \equiv \mathbf{X}_i\mathbf{b} \in \mathbb{R}^{T\times 1}$.

To avoid overfitting to noise, regularization such as l_2-norm penalty (i.e., ridge penalty) is often used.

$$\hat{\mathbf{b}} = \left(\mathbf{X}_1^{\mathsf{T}}\mathbf{X}_1 + \mathbf{L}\right)^{-1}\mathbf{X}_1^{\mathsf{T}}\mathbf{y}_1, \tag{3}$$

where $\mathbf{L} \in \mathbb{R}^{FD\times FD}$ is a Tikhonov regularization matrix. In the usual case of ridge regression (i.e., a single penalty applied to all predictors), $\mathbf{L} = \lambda\mathbf{I}_{FD}$. In the general case of multi-penalty ridge (***Hoerl and Kennard, 1970***), predictor-delay-wise penalties can be defined: $\mathbf{L} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{FD})$. For now, we assume a single penalty applies to all predictors, except for the intercept, which remains unregularized.

In practice, the hyperparameter (i.e., λ) is typically optimized using the third, independent partition of data[2] (***Hastie et al., 2009***) as:

$$\hat{\mathbf{b}} = \left(\mathbf{X}_1^{\mathsf{T}}\mathbf{X}_1 + \mathcal{L}(\{\mathbf{X}_1, \mathbf{y}_1\}; \{\mathbf{X}_3, \mathbf{y}_3\})\right)^{-1}\mathbf{X}_1^{\mathsf{T}}\mathbf{y}_1, \tag{4}$$

where $\mathcal{L}(\cdot;\cdot)$ is an optimizer that finds the *optimal* regularization matrix $\mathbf{L}^*$ minimizing prediction error for given pairs of design and response variables $\{\mathbf{X}_i, \mathbf{y}_i\}$ and $\{\mathbf{X}_j, \mathbf{y}_j\}$. With this, a regularized

[1] While *feature* and *predictor* are often interchangeably used, in this article a feature refers to a variable that describes the characteristics of interest of the input, while a predictor refers to a feature with a specific delay (i.e., each column of a design matrix). That is, F features and D delays make $P = FD$ predictors.

[2] An independent set for optimization is known as a *validation set* in the statistical learning literature (***Hastie et al., 2009***). However, in some machine learning literature (***Varoquaux, 2018***), this set is referred to as a *test set*, and the test set is referred to as a *validation set*. To mitigate the confusion, this set for optimization is referred to as an *optimization set*.

prediction for $\mathbf{y}_2$ based on $\mathbf{X}$ is:

$$\hat{\mathbf{y}}_{2,\mathbf{X}} = \mathbf{X}_2\hat{\mathbf{b}} = \mathbf{X}_2\left(\mathbf{X}_1^\mathsf{T}\mathbf{X}_1 + \mathbf{L}^*\right)^{-1}\mathbf{X}_1^\mathsf{T}\mathbf{y}_1 = \mathbf{P}_\mathbf{X}\mathbf{y}_1, \tag{5}$$

where $\mathbf{P}_\mathbf{X} \in \mathbb{R}^{T\times T}$ is a regularized projection matrix based on $\mathbf{X}$. Because of the standardization ($\|\mathbf{y}\| = 1$) and the nonnegativity of the denominator, the expected value of a prediction accuracy metric (e.g., Pearson correlation) over random noise can be seen as proportional to the expected inner product of the prediction and the response as:

$$\mathbb{E}\left[\text{corr}\left(\hat{\mathbf{y}}_{2,\mathbf{X}}, \mathbf{y}_2\right)\right] = \mathbb{E}\left[\frac{\left(\hat{\mathbf{y}}_{2,\mathbf{X}}\right)^\mathsf{T}\mathbf{y}_2}{\|\hat{\mathbf{y}}_{2,\mathbf{X}}\|\|\mathbf{y}_2\|}\right] \propto \mathbb{E}\left[\left(\hat{\mathbf{y}}_{2,\mathbf{X}}\right)^\mathsf{T}\mathbf{y}_2\right]. \tag{6}$$

This can be further expanded as:

$$\begin{aligned}\mathbb{E}\left[\left(\hat{\mathbf{y}}_{2,\mathbf{X}}\right)^\mathsf{T}\mathbf{y}_2\right] &= \mathbb{E}\left[\mathbf{y}_1^\mathsf{T}\mathbf{P}_\mathbf{X}^\mathsf{T}\mathbf{y}_2\right] \\ &= \mathbb{E}\left[(\mathbf{s}_1 + \mathbf{e}_1)^\mathsf{T}\mathbf{P}_\mathbf{X}^\mathsf{T}(\mathbf{s}_2 + \mathbf{e}_2)\right] \\ &= \mathbf{s}_1^\mathsf{T}\mathbf{P}_\mathbf{X}^\mathsf{T}\mathbf{s}_2 + \mathbb{E}\left[\mathbf{e}_1^\mathsf{T}\right]\mathbf{P}_\mathbf{X}^\mathsf{T}\mathbf{s}_2 + \mathbf{s}_1^\mathsf{T}\mathbf{P}_\mathbf{X}^\mathsf{T}\mathbb{E}\left[\mathbf{e}_2\right] + \mathbb{E}\left[\mathbf{e}_1^\mathsf{T}\right]\mathbf{P}_\mathbf{X}^\mathsf{T}\mathbf{e}_2,\end{aligned} \tag{7}$$

where only the first term remains nonzero since $\mathbb{E}[\mathbf{e}] = \mathbf{0}$. That is,

$$\mathbb{E}\left[\text{corr}\left(\hat{\mathbf{y}}_{2,\mathbf{X}}, \mathbf{y}_2\right)\right] \propto \mathbf{s}_1^\mathsf{T}\mathbf{P}_\mathbf{X}^\mathsf{T}\mathbf{s}_2. \tag{8}$$

With a sufficiently strong signal $\|\mathbf{b}\| \gg 0$, the optimal regularization approaches zero: $\mathbf{L}^* \approx \mathbf{0}$ (***Hastie et al., 2009***), which allows for approximating the projection matrix as:

$$\mathbf{P}_\mathbf{X} = \mathbf{X}_2\left(\mathbf{X}_1^\mathsf{T}\mathbf{X}_1 + \mathbf{L}^*\right)^{-1}\mathbf{X}_1^\mathsf{T} \approx \mathbf{X}_2\left(\mathbf{X}_1^\mathsf{T}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^\mathsf{T}. \tag{9}$$

Consequently, ***Equation 8*** can be approximated as:

$$\mathbb{E}\left[\text{corr}\left(\hat{\mathbf{y}}_{2,\mathbf{X}}, \mathbf{y}_2\right)\right] \propto \mathbf{s}_1^\mathsf{T}\mathbf{P}_\mathbf{X}^\mathsf{T}\mathbf{s}_2 \overset{(9)}{\approx} \left(\mathbf{X}_1\mathbf{b}\right)^\mathsf{T}\left\{\mathbf{X}_2\left(\mathbf{X}_1^\mathsf{T}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^\mathsf{T}\right\}^\mathsf{T}\left(\mathbf{X}_2\mathbf{b}\right). \tag{10}$$

Due to the symmetry of the inverse covariance $\left\{\left(\mathbf{X}_i^\mathsf{T}\mathbf{X}_i\right)^{-1}\right\}^\mathsf{T} = \left\{\left(\mathbf{X}_i^\mathsf{T}\mathbf{X}_i\right)^\mathsf{T}\right\}^{-1} = \left(\mathbf{X}_i^\mathsf{T}\mathbf{X}_i\right)^{-1}$ and the nonzero assumption $\|\mathbf{X}_i^\mathsf{T}\mathbf{X}_i\|_\text{F} > 0$, this can be further simplified as:

$$\begin{aligned}\mathbb{E}\left[\text{corr}\left(\hat{\mathbf{y}}_{2,\mathbf{X}}, \mathbf{y}_2\right)\right] \propto \mathbf{s}_1^\mathsf{T}\mathbf{P}_\mathbf{X}^\mathsf{T}\mathbf{s}_2 &\approx \mathbf{b}^\mathsf{T}\mathbf{X}_1^\mathsf{T}\left[\mathbf{X}_1\left\{\left(\mathbf{X}_1^\mathsf{T}\mathbf{X}_1\right)^{-1}\right\}^\mathsf{T}\mathbf{X}_2^\mathsf{T}\right]\mathbf{X}_2\mathbf{b} \\ &= \mathbf{b}^\mathsf{T}\left(\mathbf{X}_1^\mathsf{T}\mathbf{X}_1\right)\left(\mathbf{X}_1^\mathsf{T}\mathbf{X}_1\right)^{-1}\mathbf{X}_2^\mathsf{T}\mathbf{X}_2\mathbf{b} \\ &= \mathbf{b}^\mathsf{T}\mathbf{X}_2^\mathsf{T}\mathbf{X}_2\mathbf{b} = \|\mathbf{X}_2\mathbf{b}\| = \|\mathbf{s}_2\| > 0.\end{aligned} \tag{11}$$

That is, with a sufficiently strong signal, the prediction accuracy is expected to be positive.

Red Team: null prediction

In the above, we assumed that we have access to the true predictors $\mathbf{X}$ and true weights $\mathbf{b}$ unlike many real-world scenarios. Please note that the predictors in the current setting are not objectively observable conditions (e.g., the presence or absence of sounds) but a selected set of features of the stimulus that are hypothesized to be relevant to the neural responses by the researchers (e.g., spectrotemporal modulation). In practice, while we clearly know which stimulus we presented, defining predictors requires knowledge of which information is encoded in the human brain, which

is ultimately unknown and often is the very aim of the study (e.g., *"Is information X encoded in the brain region A or not?"*).

Now, let us consider an imaginary scenario where an independent group of researchers ("Red Team" as in adversarial testing) tries to demonstrate that our analysis method is vulnerable to any null predictors. That is, the red team wants to show that our analysis method can yield significant results even with random predictors, not only the predictors of our choice. They receive our data $\mathbf{y}$ only, but not predictors $\mathbf{X}$. For simplicity, let us assume that we also inform the Red Team about the true delays. In this case, a reasonable course of action for the Red Team could be to generate a set of random vectors $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \sigma\mathbf{I})$ and to delay them to create a Toeplitz matrix $\mathbf{U}_i \in \mathbb{R}^{T \times FD}$ to make their prediction for $\mathbf{y}_2$ as:

$$\hat{\mathbf{y}}_{2,\mathbf{U}} = \mathbf{U}_2\hat{\mathbf{b}}_0 = \mathbf{U}_2\left(\mathbf{U}_1^\intercal\mathbf{U}_1 + \mathcal{L}(\{\mathbf{U}_1, \mathbf{y}_1\}; \{\mathbf{U}_3, \mathbf{y}_3\})\right)^{-1}\mathbf{U}_1^\intercal\mathbf{y}_1 = \mathbf{P}_\mathbf{U}\mathbf{y}_1. \tag{12}$$

Note that actual predictors that researchers would use based on theories, prior evidence, and intuitions should be much more informative than the Red Team's random numbers. That is, $\mathbf{U}$ is expected to perform worse than any reasonable predictors. Put differently, if the Red Team somehow *magically* makes *significant* predictions using the null predictors $\mathbf{U}$, it indicates there is something critically flawed in our analysis.

Since we assume that the Red Team generates their null predictors $\mathbf{U}$ independently from the true predictors $\mathbf{X}$, a valid optimization process such as cross-validation (***Hastie et al., 2009***) should lead to a strong regularization of non-informative predictors (i.e., $\mathbf{U}$). This simplifies the projection matrix to:

$$\mathbf{P}_\mathbf{U} = \mathbf{U}_2\left(\mathbf{U}_1^\intercal\mathbf{U}_1 + \lambda^*\mathbf{I}\right)^{-1}\mathbf{U}_1^\intercal \overset{\lambda^* \gg 0}{\approx} \mathbf{U}_2(\lambda^*\mathbf{I})^{-1}\mathbf{U}_1^\intercal = \frac{1}{\lambda^*}\mathbf{U}_2\mathbf{U}_1^\intercal. \tag{13}$$

Thus, given the $\mathbf{U}_i$, the expected value of the prediction accuracy over random noise is given as:

$$\begin{aligned}\mathbb{E}\left[\text{corr}\left(\hat{\mathbf{y}}_{2,\mathbf{U}}, \mathbf{y}_2\right)\right] &\propto \mathbf{s}_1^\intercal\mathbf{P}_\mathbf{U}^\intercal\mathbf{s}_2 \\ &\overset{(13)}{\approx} (\mathbf{X}_1\mathbf{b})^\intercal\left(\frac{1}{\lambda^*}\mathbf{U}_2\mathbf{U}_1^\intercal\right)^\intercal(\mathbf{X}_2\mathbf{b}) \\ &= \frac{1}{\lambda^*}\mathbf{b}^\intercal\mathbf{X}_1^\intercal\mathbf{U}_1\mathbf{U}_2^\intercal\mathbf{X}_2\mathbf{b}.\end{aligned} \tag{14}$$

which converges to zero as λ^* approaches infinity:

$$\lim_{\lambda^* \to \infty}\mathbb{E}\left[\text{corr}\left(\hat{\mathbf{y}}_{2,\mathbf{U}}, \mathbf{y}_2\right)\right] = 0. \tag{15}$$

That is, the expected prediction accuracy of the Red Team is null.

Repetition of stimulus

So far, we assumed that the three partitions (i.e., training, optimization, and test sets) are independent of each other, with independent stimuli and independent noise, but only sharing the identical weights. However, presenting multiple repetitions of an identical, short stimulus—from tens to thousands of times—has been one of the most classical techniques in neuroscience to cancel out random noise in the data and reveal time-locked neural responses that are consistently evoked by the stimulus (e.g., event-related potential, time-locked BOLD response). More recently, for investigating the representation of naturalistic stimuli, a design to present an identical set of stimuli to

multiple participants has been popularized in order to reveal stimulus-driven responses in terms of inter-subject correlation (***Hasson et al., 2004***) or to find a common functional coordinate via hyperalignment (***Haxby et al., 2020***).

A problem occurs when the encoding analysis is naïvely applied to such data. To illustrate the point, let us consider an ideal design to reveal such a time-locked response, where the underlying signal is identical across partitions but the noise is independent: $\mathbf{s}_1 = \mathbf{s}_2 = \mathbf{s}_3$ but $\mathbf{e}_1 \neq \mathbf{e}_2 \neq \mathbf{e}_3$. Then, our projection matrix with $\lambda^* \approx 0$ will be:

$$\mathbf{P}_\mathbf{X} = \mathbf{X}_1 \left(\mathbf{X}_1^\mathsf{T}\mathbf{X}_1 + \mathbf{L}^*\right)^{-1} \mathbf{X}_1^\mathsf{T} \approx \mathbf{X}_1 \left(\mathbf{X}_1^\mathsf{T}\mathbf{X}_1\right)^{-1} \mathbf{X}_1^\mathsf{T}, \tag{16}$$

when $\mathbf{X}_1^\mathsf{T}\mathbf{X}_1$ is invertible. This simplifies the expected prediction accuracy to:

$$\begin{aligned} \mathbb{E}\left[\mathrm{corr}\left(\hat{\mathbf{y}}_{2,\mathbf{X}}, \mathbf{y}_2\right)\right] \propto \mathbf{s}_1^\mathsf{T}\mathbf{P}_\mathbf{X}^\mathsf{T}\mathbf{s}_1 &\overset{(16)}{\approx} (\mathbf{X}_1\mathbf{b})^\mathsf{T}\mathbf{X}_1 \left(\mathbf{X}_1^\mathsf{T}\mathbf{X}_1\right)^{-1} \mathbf{X}_1^\mathsf{T}(\mathbf{X}_1\mathbf{b}) \\ &= \mathbf{b}^\mathsf{T}\mathbf{X}_1^\mathsf{T}\mathbf{X}_1 \left(\mathbf{X}_1^\mathsf{T}\mathbf{X}_1\right)^{-1} \mathbf{X}_1^\mathsf{T}\mathbf{X}_1\mathbf{b} \\ &= \mathbf{b}^\mathsf{T}\mathbf{X}_1^\mathsf{T}\mathbf{X}_1\mathbf{b} = \|\mathbf{s}_1\| > 0. \end{aligned} \tag{17}$$

Equivalently, ***Equation 17*** being positive can be shown based on that the covariance matrix $\mathbf{X}_1^\mathsf{T}\mathbf{X}_1$ is symmetric and that $\mathbf{X}_1$ is a rectangular matrix with independent columns, which means $\mathbf{X}_1^\mathsf{T}\mathbf{X}_1$ is positive definite. By definition, $\mathbf{b}^\mathsf{T}\mathbf{X}_1^\mathsf{T}\mathbf{X}_1\mathbf{b} > 0$ for $\mathbf{b} \neq \mathbf{0}$. This leads to a rather unsurprising conclusion—with a sufficiently strong signal, the expected prediction accuracy will be positive, also with the repeated signals[3].

In the case of the Red Team, however, a repeated strong signal (i.e., $\|\mathbf{s}_1\| = \|\mathbf{s}_2\| = \|\mathbf{s}_3\| \gg 0$), even unbeknownst to the Red Team, can alter the optimization process, disabling proper regularization (see ***Appendix 1***). When unregularized (i.e., $\mathbf{L}^* \approx \mathbf{0}$), the projection matrix can be approximated as:

$$\mathbf{P}_\mathbf{U} = \mathbf{U}_1 \left(\mathbf{U}_1^\mathsf{T}\mathbf{U}_1 + \mathbf{L}^*\right)^{-1} \mathbf{U}_1^\mathsf{T} \approx \mathbf{U}_1 \left(\mathbf{U}_1^\mathsf{T}\mathbf{U}_1\right)^{-1} \mathbf{U}_1^\mathsf{T}, \tag{18}$$

Thus, similarly to ***Equation 17***, the expected null prediction accuracy of the Red Team can be approximated as:

$$\mathbb{E}\left[\mathrm{corr}\left(\hat{\mathbf{y}}_{2,\mathbf{U}}, \mathbf{y}_2\right)\right] \propto \mathbf{s}_1^\mathsf{T}\mathbf{P}_\mathbf{U}^\mathsf{T}\mathbf{s}_1 \overset{(18)}{\approx} \mathbf{s}_1^\mathsf{T}\mathbf{U}_1 \left(\mathbf{U}_1^\mathsf{T}\mathbf{U}_1\right)^{-1} \mathbf{U}_1^\mathsf{T}\mathbf{s}_1. \tag{19}$$

$\mathbf{U}_1^\mathsf{T}\mathbf{U}_1$ is positive definite due to its symmetry and independence of the columns in $\mathbf{U}_1$. Since the inverse operation preserves the signs of eigenvalues of a square matrix, its inversion $\left(\mathbf{U}_1^\mathsf{T}\mathbf{U}_1\right)^{-1}$ is also positive definite. A substitution ($\mathbf{d} \equiv \mathbf{U}_1^\mathsf{T}\mathbf{s}_1 \neq \mathbf{0}$) can clarify that ***Equation 19*** is positive by definition:

$$\mathbf{s}_1^\mathsf{T}\mathbf{P}_\mathbf{U}^\mathsf{T}\mathbf{s}_1 \approx \mathbf{d}^\mathsf{T} \left(\mathbf{U}_1^\mathsf{T}\mathbf{U}_1\right)^{-1} \mathbf{d} > 0. \tag{20}$$

This may surprise some readers; however, ***Equation 20*** implies that the null prediction of the Red Team is expected to be greater than zero. Depending on the signal-to-noise ratio, this may result in Type-I (false positive) errors. This is due to the circular fallacy introduced by the *leakage in training examples*, i.e., RDD.

In short, the mechanism of RDD can be summarized in the following steps:

[3] Notablly, here the expected prediction accuracy is proportional to the square of the signal in the training set, rather than the test set (***Equation 9***). This may affect generalization performance in hold-out validation (where the performance is evaluated only once on a separate test set), but in cross-validation the performance will be averaged out across sets. Either way, repetition of stimulus across sets leads to a spurious inflation of performance estimation.

1. The repeated signal disables the regularization of a seemingly valid optimization process even with independent noise.
2. As the regularization hyperparameter approaches zero, the projection matrix becomes positive definite, which was null when properly regularized.
3. Because of the repeated signal, the expected prediction accuracy over random noise is proportional to a bilinear form involving a nonzero vector and a positive-definite square matrix, i.e., $\mathbf{d}^{\mathrm{T}}\mathbf{M}\mathbf{d} > 0$.

Toy example

For a graphical illustration, a simple case of small-scale simulation (i.e., a toy example) is shown in ***Figure 2*** and ***Figure 3***. See Simulation methods for details of how the simulations were created. In this example, two features with a high temporal autocorrelation were generated without repetitions (***Figure 2***a) and with repetitions (***Figure 3***a). The Red Team created their own null features (***Figure 2***m). Thus, all coefficients were highly regularized for the Red Team (***Figure 2***r; pink). Consequently, the Red Team's predictions were mostly flat (***Figure 2***o,p; dotted lines) and expectedly, the Red Team's prediction accuracies were around $r = 0$ (***Figure 2***l; pink). However, with the stimulus repeated across sets (i.e., identical signals in the training and test sets), the regularization for the Red Team was much smaller, almost close to the true features (***Figure 3***r; pink vs. lime green) and the Red Team's prediction accuracies were around $r = 0.5$ (***Figure 3***l; pink). That is, due to the repetition of the stimulus, even with independent noise (***Figure 3***e,k), the prediction accuracies based on the null features were falsely inflated.

Having shown how RDD could occur, it is important to note that multiple assumptions were made to arrive here. In practice, various factors—such as signal strength, temporal and spatial correlation structures, feature collinearity, similarity across partitions, and the flexibility of the FIR model—might interact to inflate Type-I errors.

Simulation

In this section, I highlight major factors that worsen the Type-I error due to RDD. To keep the number of combinations manageable, univariate models ($F = 1$, $V = 1$) were first considered. Then, while iteratively pruning out irrelevant factors, models with multivariate features (F) and multivariate responses (V) were considered. Methodological details of the simulation are described in the Simulation methods section. The parameters of simulation are summarized in ***Table 2***.

Univariate-feature, univariate-response

The first batch of simulations was restricted to a univariate feature (the number of features $F = 1$) and a univariate response (the number of variates $V = 1$). The explored parameter levels were: $D \in \{1, 3, 5, 7, 9, 11\}$, $S \in \{-10, 0, 10\}$, $\phi_X \in \{0, 0.5, 1\}$, $\phi_U \in \{0, 0.5, 1\}$, $\phi_E \in \{0, 0.5, 1\}$, $\phi_B \in \{0, 0.5, 1\}$, $\texttt{IsRep} \in \{0, 1\}$. With these parameter levels, total 2,916 combinations were created, each sampled 1,000 times.

To illustrate the most distinctive effects, prediction accuracies averaged across 1,000 random sampling are shown in ***Figure 4***. Expectedly, the SNR increased the prediction accuracy based on

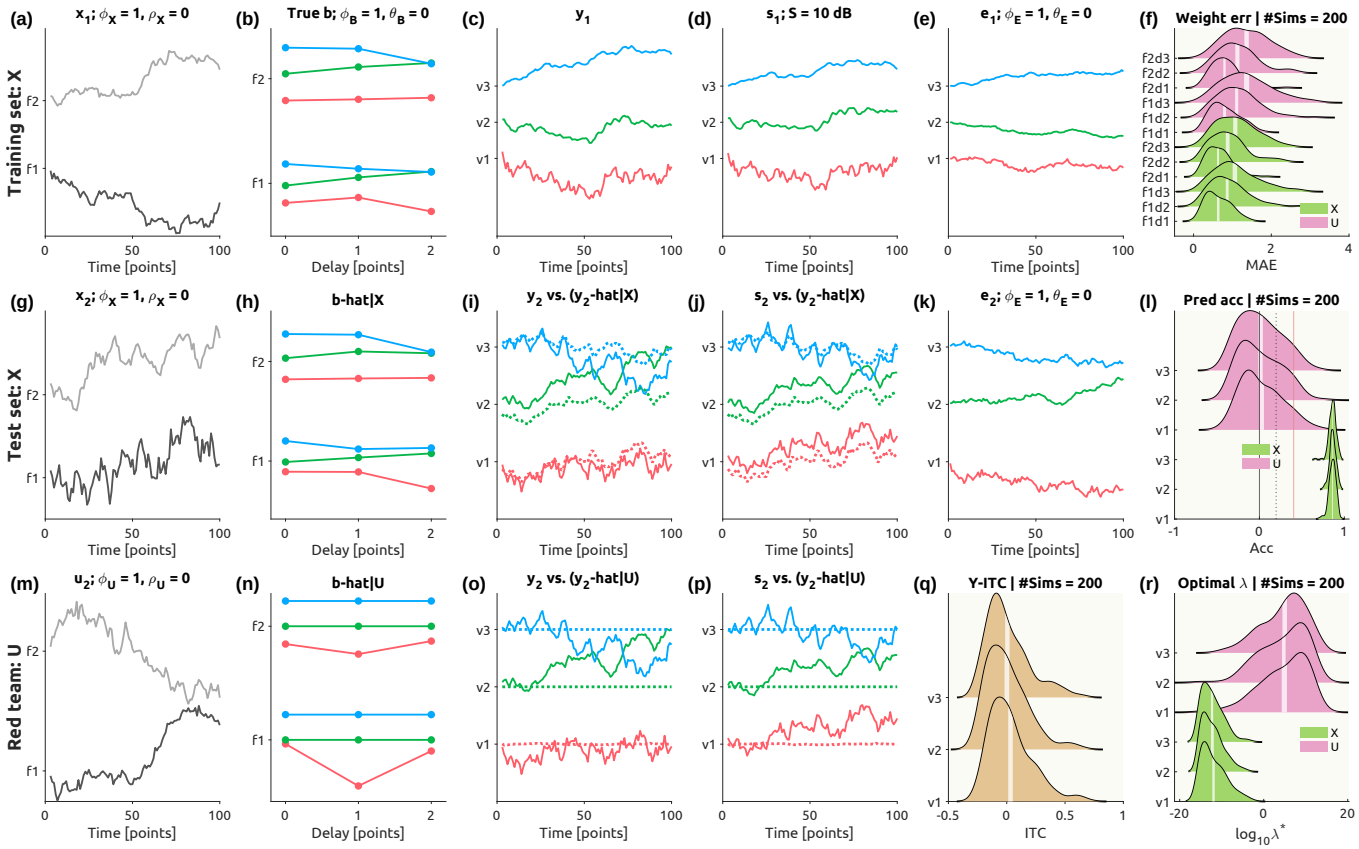


Figure 2. A toy example without stimulus repetition. Apart from the right-most panels in pale beige **(f, l, r)**, the panels in the first row **(a–e)** correspond to the training set with the true features $\mathbf{X}_1$, the panels in the second row **(g–k)** correspond to the test set with the true features $\mathbf{X}_2$, and the panels in the third row **(m–p)** correspond to the Red Team's null features $\mathbf{U}_2$. The panels in the first column **(a, g, m)** show the true features $\mathbf{x}$ or null features $\mathbf{u}$ in gray scale. The panels in the second column show the true weights $\mathbf{b}$ **(b)** or estimation based on the true features **(h)** or null features **(n)** where the RGB color represents one of three response variates. The panels in the third column **(c, i, o)** show the response $\mathbf{y}$ (solid lines) and prediction based on true or null features $\mathbf{y}$-hat (dashed lines), and the panels in fourth column **(d, j, p)** show the true signal $\mathbf{s}$ (solid lines) and the prediction (dashed lines). The two panels in the fifth column **(e, k)** show true noise $\mathbf{e}$. In the pale beige panels **(f, l, q, r)**, distributions from 200 simulations are shown in ridgeline plots with the 95% confidence interval of the mean shown in white strips: **(f)** mean absolute error (MAE) of weight estimation based on the true features ($\mathbf{X}$; lime green) or null features ($\mathbf{U}$; pink), **(l)** prediction accuracies in Pearson's correlation coefficients with a black vertical line for an absolute zero, a gray vertical line for an uncorrected $P < 0.05$ (assuming independent time points), and a red vertical line for a Bonferroni-corrected $P < 0.05$ adjusted for the number of variates. **(q)** inter-trial correlation (ITC) of the responses (brown), **(r)** exponents of the geometrically averaged optimal λ's. Parameters to generate this simulation set are: $T = 100, V = 3, F = 2, D = 3, S = 1\text{dB}, \phi_X = 1, \rho_X = 0, \phi_B = 1, \theta = 0, \phi_E = 1, \theta_E = 0, \texttt{IsRep} = 0$. See ***Table 2*** for the explanation of parameters.

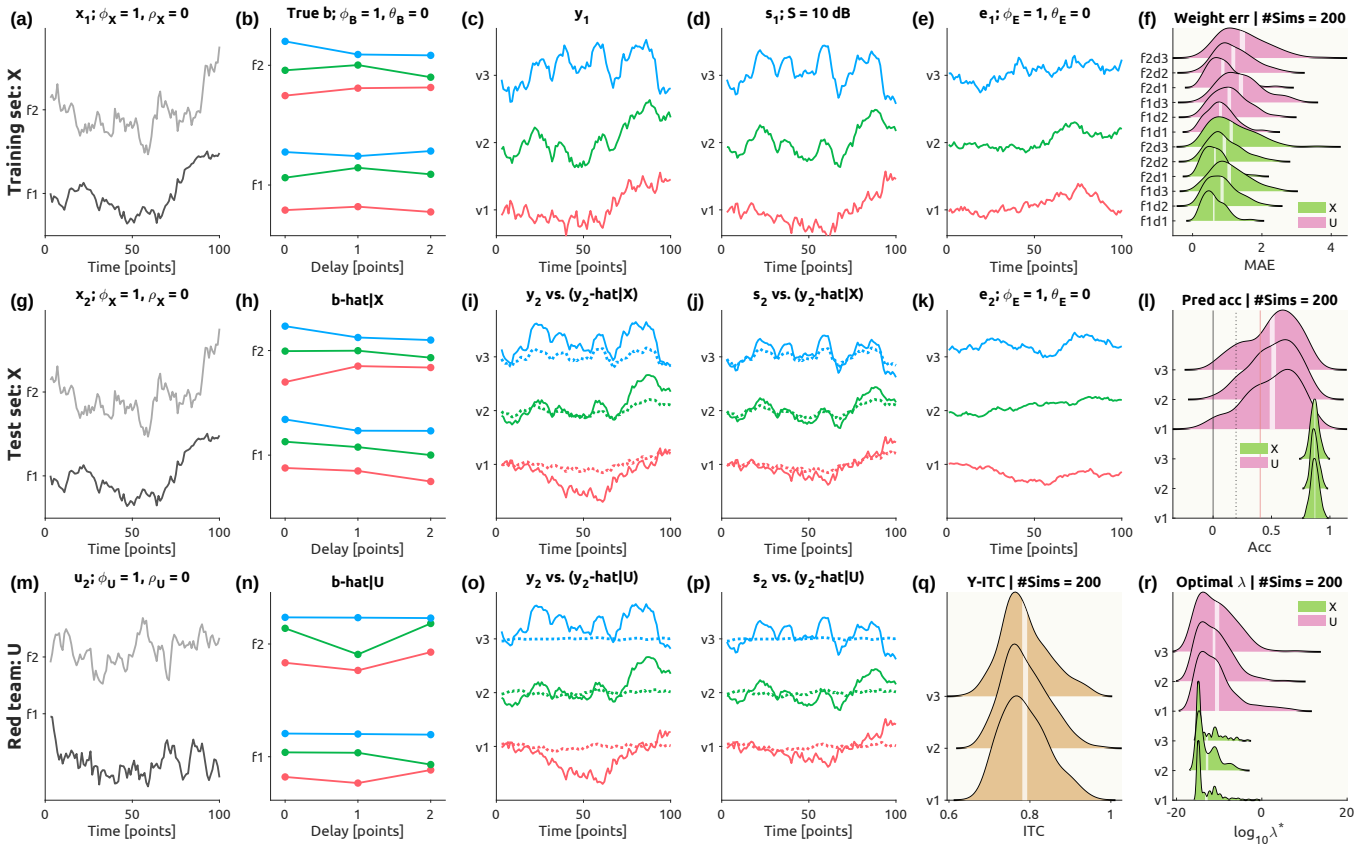


Figure 3. A toy example with stimulus repetition. The visualization scheme is identical to ***Figure 2***. Parameters to generate this simulation set are identical to those in ***Figure 2***, except `IsRep` = 1.

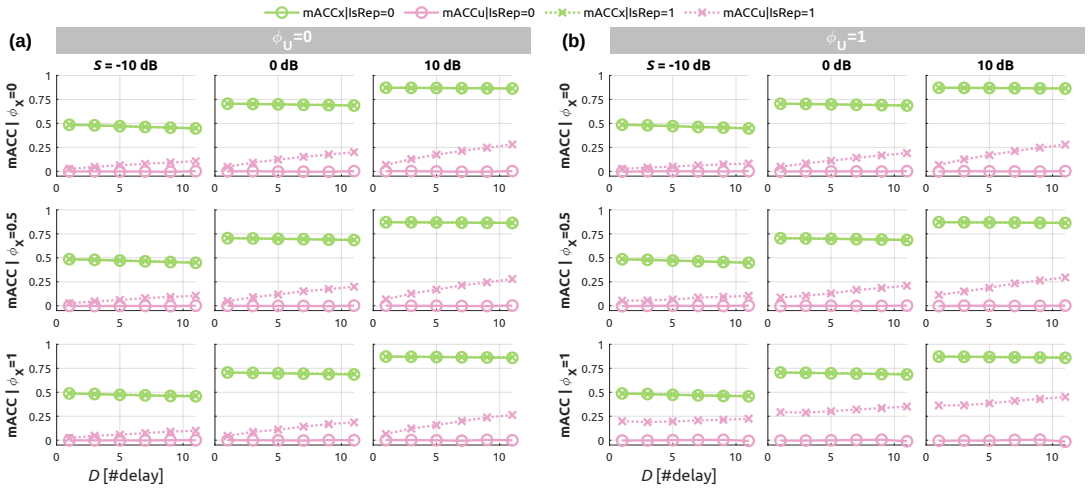


Figure 4. Simulations of univariate-feature, univariate-response models. Mean prediction accuracies are plotted over the number of delays D when the temporal autocorrelation **(a)** $\phi_U = 0$ and **(b)** $\phi_U = 1$. Each marker corresponds to the averaged Pearson correlation coefficients of 1000 simulations. Predictions based on true features ($\mathbf{X}$) are shown in lime green and null features ($\mathbf{U}$) in pink. Open circles with solid lines indicate accuracies when the true signals were not repeated. Crosses with dashed lines represent cases where the true signals were repeated. Each column corresponds to a specified SNR level. $\phi_X = 0$ in the top row and $\phi_X = 1$ in the bottom row. Other parameters were as follows: $K = 1000$, $\phi_B = 0$, $\phi_E = 0$. Other combinations of parameters showed generally similar patterns. See ***Table 3*** and ***Table 4*** for effect sizes.

the true features (green markers). The prediction based on the null features remained zero when the stimuli were independent across the cross-validation (CV) partitions (pink circles). However, when the stimuli were identical across the partitions, the null prediction accuracy increased over the number of delays and the SNR (pink crosses). In particular, the effect of temporal autocorrelation ϕ_X when $\phi_U = 1$ substantially increased the null prediction with the repeated stimuli.

To comprehensively assess the possible effects, a full factorial model on mean prediction accuracies was fitted with all seven variables as in Wilkinson notation:

$$r \sim 1 + D * S * \phi_X * \phi_U * \phi_E * \phi_B * \texttt{IsRep}, \tag{21}$$

where r is the mean prediction accuracy averaged for each set of 1000 simulations either based on true or null predictors, 1 represents an intercept, and $*$ denotes factor crossing as in $a * b = a + b + a : b$ with $:$ denoting an interaction. This yielded linear models with 128 terms from the intercept to the seven-way interaction, 2,916 observations, and adjusted $R^2 = 0.985$ for $\mathbf{X}$ and 0.957 for $\mathbf{U}$, which are reasonable given that the accuracy metrics are averaged within each combination of parameters. Due to the large number of observations, P-values were not highly selective ($P_{\text{Bonferroni}} < 0.0001$ for many of contrasts; 9 for $\mathbf{X}$, 27 for $\mathbf{U}$ out of 127). Thus, on top of $P_{\text{Bonferroni}} < 0.0001$, only effects with moderate effect sizes ($\eta_p^2 \geq 0.16$) were considered for further discussion (***Table 3***, ***Table 4***).

As expected, the SNR consistently increased the prediction accuracy ($\eta_p^2[S] = 0.985$ for $\mathbf{X}$; $\eta_p^2[S] = 0.660$ for $\mathbf{U}$). Most importantly, the interaction of the stimulus repetition ($\texttt{IsRep}$) with the signal strength (SNR, S), and the autocorrelation of the true and null features (ϕ_X, ϕ_U) were markedly found on the prediction accuracy based on the null features ($\max \eta_p^2 = 0.661$, ***Table 4***). Put differently, the RDD effect (i.e., a false inflation of the null prediction accuracy by the repetition of stimulus

across CV partitions) was most pronounced when the true underlying signal was strong and the true features were temporally autocorrelated. This finding is consistent with ***Equation 17*** above, showing that the RDD effect depends on the nonzero strength of the underlying signal. The clearer the signal, the more likely the RDD effect will occur when analyzed in a circular CV design.

In addition, the RDD effect strongly interacted with the complexity of the underlying (and fitted) models (i.e., $\texttt{IsRep} : D$; ***Table 4***). While in a correct CV design ($\texttt{IsRep} = 0$), the model complexity did not increase the prediction accuracy as this was regularized by independent optimization and evaluation. However, in a circular CV design ($\texttt{IsRep} = 1$), the model complexity increased the prediction accuracy (hence the RDD effect) since the model was fitted to the identical signal across CV partitions. Naturally, this effect further interacted with the strength of the signal ($\texttt{IsRep} : D : S$; ***Table 4***).

The effect of the autocorrelation of the weight time series ϕ_B was found to be negligible for both $\mathbf{X}$ and $\mathbf{U}$ ($\eta_p^2 < 0.009$). Guided by these results, we fixed the autocorrelation of the weight time series to zero ($\phi_B = 0$) in the following simulations.

Multivariate-feature, univariate-response

Since most often we are interested in multivariate features (e.g., motion energies, spectrograms, deep ANN embeddings), it is of interest how the dimensionality and multicollinearity of features influence RDD artifact.

The explored parameter levels were: $D \in \{1, 5, 9\}$, $F \in \{5, 10, 15, 20\}$, $S \in \{-10, 0, 10\}$, $\phi_X \in \{0, 0.5, 1\}$, $\phi_U \in \{0, 0.5, 1\}$, $\phi_E \in \{0, 0.5, 1\}$, $\phi_B = 0$, $\rho_X \in \{0, 0.5, 1\}$, $\rho_U \in \{0, 0.5, 1\}$, $\texttt{IsRep} \in \{0, 1\}$. With these levels, the full combinations amounted to 17,496, for each of which, once more, 1,000 random samplings were carried out.

Figure 5 displays the simulated effects. The number of features F reduced the true models' prediction accuracies without repetitions (green solid lines), and more so with more complex response functions (e.g., $D = 9$). In contrast, increasing the number of features led to higher prediction accuracies in null models with stimulus repetition (pink dashed lines). Moreover, at a higher SNR (e.g., 10 dB), the null prediction accuracies (pink dashed lines) were even higher than the true prediction accuracies (green solid lines) with many independent predictors (e.g., 9 delays × ≥15 features when $\rho_X = 0$; ***Figure 5***a). This suggests that, in a bad combination, the null prediction accuracy can go beyond the noise ceiling (i.e., green solid lines; assuming we know the true predictors), which is the plausibly highest prediction accuracy bounded by the noise level of the signal. This crossing of the true and null prediction accuracies was attenuated when features were highly correlated ($\rho_X = 1$; ***Figure 5***b), which reduced the effective degrees of freedom.

To quantify the observed effects, once again, a full factorial model on mean prediction accuracies was fitted with all nine variables:

$$r \sim 1 + D * F * S * \phi_X * \phi_U * \phi_E * \rho_X * \rho_U * \texttt{IsRep}. \tag{22}$$

A full factorial model with 17,496 observations and 511 terms resulted in adjusted $R^2 = 0.975$ for $\mathbf{X}$ and 0.923 for $\mathbf{U}$ (***Table 5***, ***Table 6***). A strong effect of the number of features F was found for $\mathbf{X}$ ($\eta_p^2 = 0.276$) but only an intermediate effect for $\mathbf{U}$ ($\eta_p^2 = 0.102$). Additionally, its interaction with

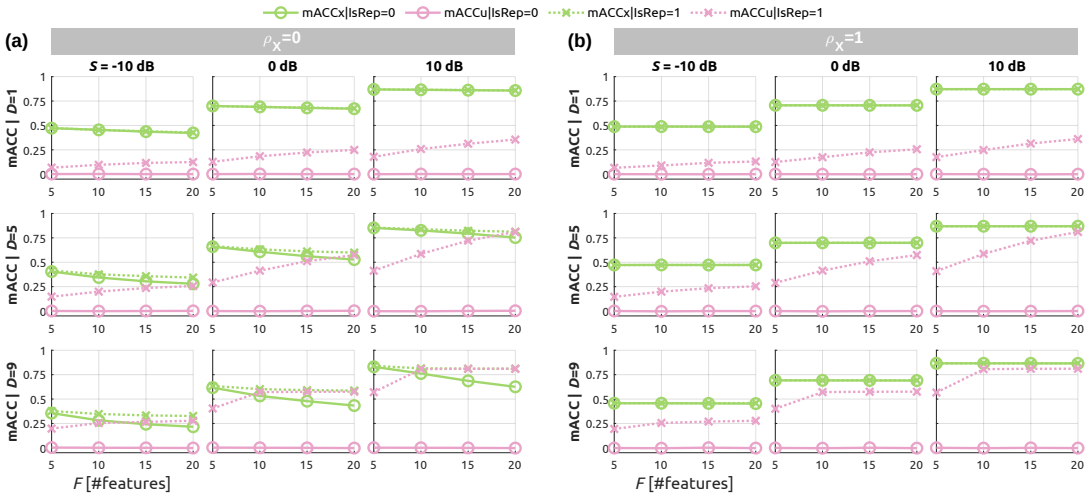


Figure 5. Simulations of multivariate-feature, univariate-response models. Mean prediction accuracies are plotted over the number of features F when the multicolinearity **(a)** $\rho_X = 0$ and **(b)** $\rho_X = 1$. Marker styles and colors match to those in ***Figure 4***. Each column represents a specified SNR level. The number of delays is $D = 1$ in the top rows and $D = 9$ in the bottom rows. Other parameters were as follows: $K = 1000$, $\phi_U = 0$, $\phi_B = 0$, $\phi_E = 0$, $\rho_U = 0$. See ***Table 5*** and ***Table 6*** for effect sizes.

Figure 5—figure supplement 1. Additional cases with $\rho_U = 0$ and $\rho_U = 1$ when $\rho_X = 0$

stimulus repetition (`IsRep`) for the null features was only intermediate $\mathbf{U}$ ($\eta_p^2 = 0.101$), suggesting the dimension of the features alone was not a strong factor for the RDD effect.

However, the interaction between the stimulus repetition and the multicollinearity of the null features (ρ_U : `IsRep`) was strong ($\eta_p^2 = 0.442$; ***Table 6***), where the RDD effect was more pronounced when the null features exhibited lower autocorrelation (***Figure 5—figure Supplement 1***). That is, when the null features have greater effective degrees of freedom (i.e., lower autocorrelation), the model could more flexibly fit the repeated signal, inflating the RDD effect.

In addition, an interesting finding was that the true prediction accuracy decreased as the number of features F and the number of delays D increased. This was due to the limited number of samples ($T = 100$) as compared to the high dimensionality of the feature space, which made the linear model ill-posed. While regularization makes fitting feasible, the inherent limitation exist. Put differently, these results suggest that a sufficient number of samples is required to faithfully estimate the transfer function of the high-dimensional feature space.

Multivariate-feature, multivariate-response

Finally, it was tested whether the dimensionality and spatial autocorrelation of responses (variates) affect the RDD artifact. First, it is worth noting that the encoding model is typically a univariate-response model (e.g., “voxel-wise” or “channel-wise”). The assumption of spatially (i.e., across response units) independent noise is similar to that of the classical general linear model (***Friston et al., 1994***), where a diagonal covariance structure across lattice sampling grids (e.g., voxels) is assumed[4]. Likewise, the encoding model is independently optimized for each response unit in the

[4]Therefore, the term ‘massive-univariate’ would be more fitting than ‘multivariate’.

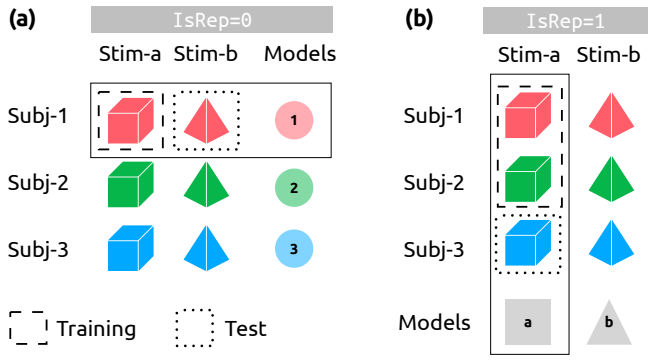


Figure 6. Schema of two cross-validation designs. For simplicity, let us say we have three subjects $\{1, 2, 3\}$ (depicted in red, green, blue) and two stimuli $\{a, b\}$ (depicted as a cube and a tetrahedron). Note that we assume any repetitions of stimuli within a subject are averaged prior to the cross-validation. **(a)** `IsRep=0`. Subject-specific models (pale colored circles). Each model (e.g., pale red circle in solid rectangle) is trained with one stimulus (dashed rectangle) and tested on another stimulus (dotted rectangle). **(b)** `IsRep=1`. Stimulus-specific models (gray square and triangle). Each model (e.g., gray square in sold rectangle) is trained with two subjects (dashed rectangle) and tested on the other subject (dotted rectangle). The optimization set, which could be in the training set of the outer loop, is not marked for simplicity.

current paper[5]. Thus, while the spatial dependency may affect the family-wise error rates of multilpe testing (since many correction methods exploit neighbouring supports), it is unexpected that the spatial dependency directly alters the unit-wise optimization and following weight estimation. However, for completeness, simulations were carried out with following parameters: $D \in \{1, 7, 11\}$, $F \in \{5, 10, 20\}$, $V \in \{5, 10, 20\}$, $S \in \{-10, 0, 10\}$, $\phi_X \in \{0, 0.5, 1\}$, $\phi_U \in \{0, 0.5, 1\}$, $\phi_E \in \{0, 0.5, 1\}$, $\phi_B = 0$, $\theta_B \in \{0, 0.5, 1\}$, $\theta_E \in \{0, 0.5, 1\}$, $\rho_X \in \{0, 0.5, 1\}$, $\rho_U \in \{0, 0.5, 1\}$, `IsRep` $\in \{0, 1\}$. With these parameter levels, there were 354,287 total combinations, as before, each sampled 1,000 times. A full factorial model, incorporating all twelve variables, was fitted using 354,287 observations and 4,096 terms. As expected, the spatial dependency θ_B and θ_E neither showed any marked main effects nor interactions ($\eta_p^2 < 0.160$; ***Table 7***, ***Table 8***).

Real Data

In this section, I present real-data examples where RDD spuriously inflated null prediction accuracies using open-access data where healthy participants listened to various musical excerpts while measuring neural activity (electroencephalography [EEG] or functional magnetic resonance imaging [fMRI]) or behavioral ratings (***Kaneshiro et al., 2020***; ***Sachs et al., 2020***).

Based on the well-established encoding of the acoustic energy in the human auditory system, true features were the audio envelopes extracted from the musical stimuli using a cochlear model (***Chi et al., 2005***). Null features were either (a) the phase-randomized envelope (preserving spectral magnitudes and autocorrelation structures) as the most realistic one, (b) the normal noise, and (c) the uniform noise as the least realistic one. Details of the real data and analysis implementation are provided in the Materials and Methods section.

Importantly, the data were analyzed with two competing cross-validation (CV) designs (***Figure 6***):

[5] In some EEG studies, hyperparameters were averaged across response units (i.e., EEG channels). This practice introduces spatial dependency that leads to a suboptimal regularization for individual response units.

1. $\texttt{IsRep} = 0$: No stimulus was repeated across CV sets (training, optimization, and test). That is, subject-specific models were fit with stimuli assigned to CV sets.
2. $\texttt{IsRep} = 1$: Identical stimuli, albeit with different noise realizations, were repeated in all three sets. Thus, stimulus-specific models were fit with participants partitioned into CV sets.

The magnitude of RDD artifact was estimated as the difference in null prediction accuracies between two CV schemes: $\mathrm{RDD} = \bar{r}_{\mathrm{stim}}(\mathbf{U};\, \texttt{IsRep} = 1) - \bar{r}_{\mathrm{subj}}(\mathbf{U};\, \texttt{IsRep} = 0)$ where r_{stim} is a prediction accuracy of a stimulus-specific model and r_{subj} is that of a subject-specific model. $\bar{(\cdot)}$ denotes averaging across null models ($M = 100$). That is, if there were no false inflation of prediction accuracy due to RDD, the null predictions with and without stimulus repetitions should be equal (i.e., $\mathcal{H}_0 : \mathbb{E}(\mathrm{RDD}) = 0$). Otherwise, the null prediction with stimulus repetitions is expected to be greater than the null prediction without repetitions ($\mathcal{H}_A : \mathbb{E}(\mathrm{RDD}) > 0$).

All data analyses were consistently done using an MATLAB package Linearized Encoding Analysis (LEA; https://github.com/seunggookim/lea).

Electroencephalography

Scalp electrical potential data were recorded in 48 healthy participants while listening to Western-style Indian pop music (i.e., Bollywood music; ***Kaneshiro et al., 2020***). Using this EEG dataset, linearized encoding analysis was performed with the audio envelope as a true feature and its phase-randomized signals as null features.

Without stimulus repetition ($\texttt{IsRep} = 0$), a clear fronto-central topography is shown in the prediction accuracy ($\max r(X;0) = 0.047$, ***Figure 7*****a**) as well as in the ridge hyperparameter ($\min \log_{10} \lambda(X;0) = 5.86$, ***Figure 7*****b**), reflecting the envelope encoding in the bilateral auditory cortices while listening to music. For this particular data, the estimated weights were stronger in the left that right fronto-central channels (***Figure 7*****c**). With the phase-randomized envelope, as expected, the null prediction accuracy was minimal ($\max r(U;0) = 0.006$, ***Figure 7*****d**; $\max \mathbb{E}\left[r(U;0)\right] = 0.001$, ***Figure 7*****g**) with all channels were highly regularized ($\min \log_{10} \lambda(U;0) = 10.98$, ***Figure 7*****e**; $\min \mathbb{E}\left[\log_{10} \lambda(U;0)\right] = 12.11$, ***Figure 7*****h**).

With stimulus repetition ($\texttt{IsRep} = 1$), true prediction accuracies were increased ($\max r(X;1) = 0.085$, ***Figure 7*****j**). This is because, unlike the simulation, our feature (i.e., the audio envelope) was not the sole information that the human EEG data encode.

However, most strikingly, the null prediction accuracies with stimulus repetition showed an almost identical topography to the actual encoding results (***Figure 7*****m,p**), with even higher values than the true prediction without repetition ($\max r(X;0) = 0.047$, $\max r(U;1) = 0.052$, $\max \mathbb{E}\left[r(U;1)\right] = 0.056$). Note that, by definition, the null feature (phase-randomized envelopes) should have not predicted anything in the EEG data. However, when the identical stimuli were repeated over CV partitions, the regularization was disabled—regardless of the given features—in channels where the stimulus-evoked response is strong (***Figure 7*****n,q**). Then, even random weights (***Figure 7*****o,r**; ***Figure 7—figure Supplement 9***; i.e., widely different from the true weights) could successfully predict the repeated signal. Because RDD artifact reflects the genuine biological signal that is repeatedly evoked by identical stimuli, the observed patterns of the prediction accuracy may appear indistinguishable from the true signal without close investigation of weights. This pattern of RDD was

observed, albeit weaker, even when the null features were unrealistic such as normal noise (***Figure 7—figure Supplement 1***) or uniform (***Figure 7—figure Supplement 2***). Moreover, the RDD effect was consistent across different sets of delays (***Figure 7—figure Supplement 3***, ***Figure 7—figure Supplement 4***, ***Figure 7—figure Supplement 5***, ***Figure 7—figure Supplement 6***, ***Figure 7—figure Supplement 7***, ***Figure 7—figure Supplement 8***).

Transfer function weights showed interesting patterns (***Figure 7—figure Supplement 9***, ***Figure 7—figure Supplement 10***, ***Figure 7—figure Supplement 11***). Spatially, the eigenvectors explaining the largest variance (i.e., PC1) of the phase-randomized envelope were similar to those of the true envelope while the eigenvectors of the uniform or normal noise without autocorrelation showed rather noisy (spatially high frequencies) patterns. Temporally, even though the uniform or normal noise features had no autocorrelation, the eigenvariate time series showed smooth patterns, reflecting the autocorrelation of the EEG data.

When comparing the null prediction accuracies between the CV schemes (e.g., ***Figure 7*****g** vs. ***Figure 7*****p**), the RDD effects were found significant in all channels and displayed a fronto-central topography that is highly plausible for the auditory cortical activity ($P_{FDR} < 0.01$; ***Figure 8***). It is noteworthy that the RDD effect is much stronger for the phase-randomized envelope, which preserves the autocorrelation structure of the stimulus. Also, the number of delays seems to further inflate the RDD effect as demonstrated in the simulation results (e.g., ***Figure 4***).

Functional magnetic resonance imaging

Blood-oxygen-level-dependent (BOLD) data were acquired in 39 healthy participants while listening to Western instrumental musical pieces that either evoke happiness or sadness as validated in independent listeners (***Sachs et al., 2020***). As done for the EEG dataset, linearized encoding analysis was performed with the fMRI data as responses, the audio envelope as a true feature, the phase-randomized envelope as a null feature, and delays from 3 to 9 seconds (***Figure 9***). Similarly to the EEG results, the phase-randomized envelope strikingly predicted the BOLD time series in the bilateral auditory cortices including the Heschl's gyrus and planum temporale (***Figure 9*****m,p**) while no consistent pattern in the transfer function weights was found over the phase randomizations (***Figure 9*****r**), clearly demonstrating the RDD effect. Once again, the anatomical location and the extent of the heightened null prediction accuracies precisely matched the true encoding results (***Figure 9*****a,j**), which would seem 'highly convincing' to many human neuroscientists.

When analyzed with different noise models and delays (***Figure 9—figure Supplement 1***, ***Figure 9—figure Supplement 2***, ***Figure 9—figure Supplement 3***, ***Figure 9—figure Supplement 4***, ***Figure 9—figure Supplement 5***, ***Figure 9—figure Supplement 6***, ***Figure 9—figure Supplement 7***, ***Figure 9—figure Supplement 8***), the similar RDD pattern was consistently observed (i.e., inflated prediction accuracies in the auditory cortices). The RDD effect was statistically significant in not only the bilateral superior temporal gyri but the medial occipital cortices and the inferior frontal cortices, where acoustic energy is not expected to be encoded (***Figure 10***). Consistently with the EEG results, the RDD effect was stronger for the phase-randomized envelope than the normal or uniform noise as well as for longer delay points than shorter ones.

The transfer function weights (***Figure 9—figure Supplement 9***, ***Figure 9—figure Supplement 10***,

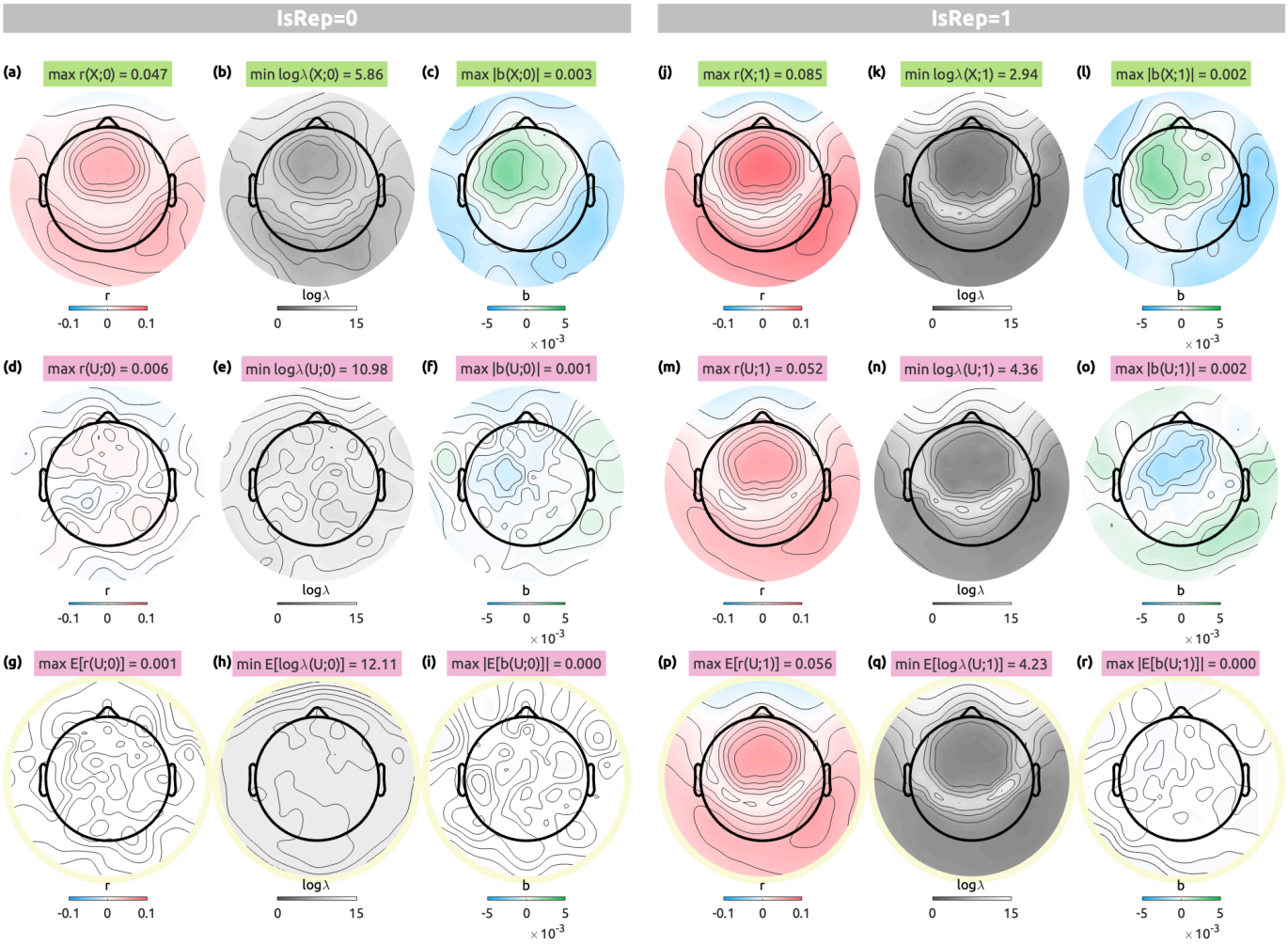


Figure 7. EEG linearized encoding analysis results with delays from 0 to 0.5 sec with an audio envelope (top row, **(a-c, j-l)**), a single case of a phase-randomized envelope (middle row, **(d-f, m-o)**), and an average of 100 phase-randomized envelopes (bottom row, circled in pale yellow, **(g-i, p-r)**). For each CV scheme ($\texttt{IsRep} = 0$, left panels, **(a-i)**; $\texttt{IsRep} = 1$, right panels, **(j-r)**), prediction accuracy (r, blue to red, **(a, d, g, j, m, p)**), logarithmic ridge hyperparameter ($\log_{10} \lambda$, gray to white, **(b, e, h, k, n, q)**), transfer function weights that are summed over delays (b, blue to green, **(c, f, i, l, o, r)**) are shown along the columns. Note that stimulus repetition not only slightly inflated true prediction accuracies but even the predicted 'brain activity pattern' from null features (**m,p**) which was literally indistinguishable from true predictions (**a, j**).

Figure 7—figure supplement 1. EEG linearized encoding analysis with normal noise and delays from 0 to 0.5 sec

Figure 7—figure supplement 2. EEG linearized encoding analysis with uniform noise and delays from 0 to 0.5 sec

Figure 7—figure supplement 3. EEG linearized encoding analysis with the phase-randomized envelope and delays from 0 to 0.3 sec

Figure 7—figure supplement 4. EEG linearized encoding analysis with normal noise and delays from 0 to 0.3 sec

Figure 7—figure supplement 5. EEG linearized encoding analysis with uniform noise and delays from 0 to 0.3 sec

Figure 7—figure supplement 6. EEG linearized encoding analysis with the phase-randomized envelope and delays from 0 to 1 sec

Figure 7—figure supplement 7. EEG linearized encoding analysis with normal noise and delays from 0 to 1 sec

Figure 7—figure supplement 8. EEG linearized encoding analysis with uniform noise and delays from 0 to 1 sec

Figure 7—figure supplement 9. EEG transfer function weights with all three noise models and delays from 0 to 0.5 sec

Figure 7—figure supplement 10. EEG transfer function weights with all three noise models and delays from 0 to 0.3 sec

Figure 7—figure supplement 11. EEG transfer function weights with all three noise models and delays from 0 to 1 sec

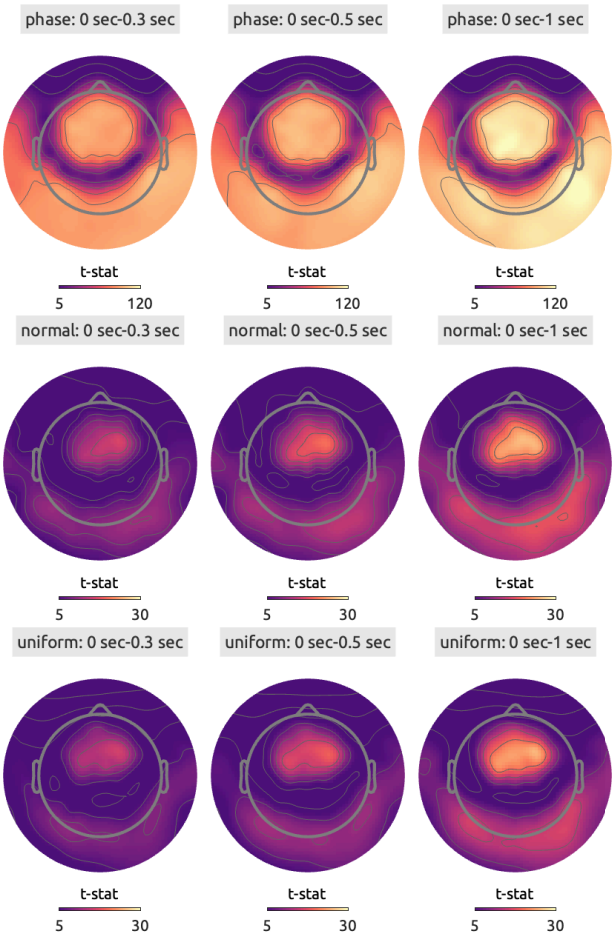


Figure 8. t-statistic maps comparing the null prediction accuracies between two CV schemes (e.g., ***Figure 7*****g** vs. ***Figure 7*****p**) to test the RDD effects in the EEG data for the phase-randomized envelope (top row), the normal noise (middle row), and the uniform noise (bottom row). All channels were found significant after FDR adjustment in all cases ($P_{FDR} < 0.01$).

Figure 9—figure Supplement 11) display strong "auditory components" even for normal and uniform noise in their eigenvectors while the corresponding eigenvariates were widely different from the weights estimated by the true feature.

Behavioral ratings

Continuous ratings of music-evoked emotions were sampled from the same 39 healthy participants who took part in the fMRI experiment above (*Sachs et al., 2020*). After the scanning session, participants listened to the same musical pieces again and rated their Emotionality (how happy/sad they felt) and Enjoyment (how much they enjoyed the piece) using a slider. A linearized encoding analysis was performed with the ratings as responses, the audio envelope as a true feature, the phase-randomized envelope as a null feature, and delays from 0 to 10 seconds (*Figure 11*). Once more, while the true envelope predicted Emotionality to some degrees and Enjoyment to a greater extent (*Figure 11***a**), the phase-randomized envelope also predicted both scales well above zero when the stimuli were repeated across CV partitions (*Figure 11***m**,**p**) unlike when the stimuli were not repeated (*Figure 11***g**).

The RDD effect was significant also in the behavioral ratings consistently across all noise models and delays (*Figure 12*; *Figure 11—figure Supplement 1*, *Figure 11—figure Supplement 2*, *Figure 11—figure Supplement 3*, *Figure 11—figure Supplement 4*, *Figure 11—figure Supplement 5*, *Figure 11—figure Supplement 6*, *Figure 11—figure Supplement 7*, *Figure 11—figure Supplement 8*). Similarly to other modalities, the transfer function weights reflected the inherent autocorrelation structure of the behavioral data (*Figure 11—figure Supplement 9*, *Figure 11—figure Supplement 10*, *Figure 11—figure Supplement 11*).

Discussion

The primary objective of cognitive neuroscience is to comprehend how the brain executes information-processing operations (*Kay, 2018*). Linearized encoding analysis serves as a robust method to evaluate a model (i.e., transfer function) that describes how the brain encodes sensory information from the environment and processes this information further (*Naselaris et al., 2011*). Prediction accuracy is crucial as it measures the model's ability to generalize to unseen stimulus-response pairs. This paper elucidates how information leakage in training examples (i.e., RDD) can artificially inflate prediction accuracy and demonstrates this through extensive simulations and real data analyses.

Firstly, it was mathematically shown that the expected prediction accuracy of the null feature could exceed zero when the null features are identically repeated across CV partitions (*Equation 20*). It is important to understand that RDD arises not from noise in the data but from the stimulus-driven similarity. In particular, the similarity between training and optimization sets disables regularization leading to an inflation of the null prediction accuracy.

Secondly, simulations showed that the RDD effect (i.e., the inflation of the prediction accuracy due to RDD; an interaction with `IsRep`) is more pronounced with a higher signal-to-noise ratio, greater flexibility (i.e., longer delay points or higher dimensional features), and more similar autocorrelation structures between the true and null features (*Table 4*; *Table 6*).

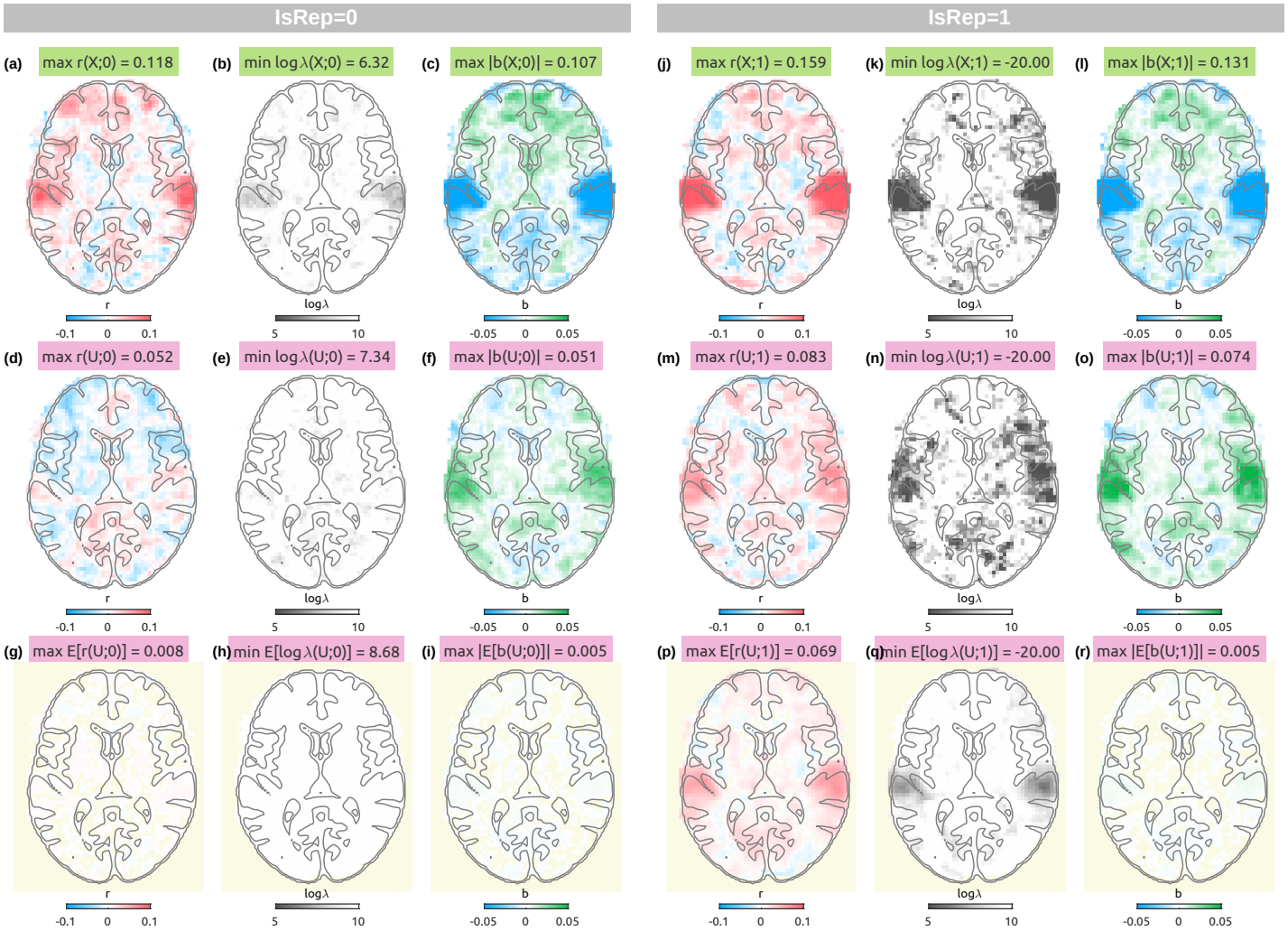


Figure 9. fMRI linearized encoding analysis results with delays from 3 to 9 sec with the audio envelope (top row, **(a-c, j-l)**), a single case of a phase-randomized envelope (middle row, **(d-f, m-o)**), and an average of 100 phase-randomized envelopes (bottom row, pale yellow background, **(g-i, p-r)**). For each CV scheme ($\texttt{IsRep} = 0$, left panels, **(a-i)**; $\texttt{IsRep} = 1$, right panels, **(j-r)**), prediction accuracy (r, blue to red, **(a, d, g, j, m, p)**), logarithmic ridge hyperparameter ($\log_{10} \lambda$, gray to white, **(b, e, h, k, n, q)**), transfer function weights that are summed over delays (b, blue to green, **(c, f, i, l, o, r)**) are shown along the columns. The analysis was done in the 3-D space, but transverse slices (Montreal Neurological Institute [MNI]-coordinate Z = 8 mm) are chosen to display anatomical structures implicated in a meta-analysis on music-evoked emotions (***Koelsch, 2020***) such as the Heschl's gyrus, planum tempolare, the inferior frontal cortex. The 3-D volumes can be viewed with the NeuroVault web viewer (https://identifiers.org/neurovault.collection:19626).

Figure 9—figure supplement 1. fMRI linearized encoding analysis with normal noise as null features and delays from 3 to 9 sec

Figure 9—figure supplement 2. fMRI linearized encoding analysis with uniform noise as null features and delays from 3 to 9 sec

Figure 9—figure supplement 3. fMRI linearized encoding analysis with the phase-randomized envelope and delays from 4 to 6 sec

Figure 9—figure supplement 4. fMRI linearized encoding analysis with normal noise as null features and delays from 4 to 6 sec

Figure 9—figure supplement 5. fMRI linearized encoding analysis with uniform noise as null features and delays from 4 to 6 sec

Figure 9—figure supplement 6. fMRI linearized encoding analysis with the phase-randomized envelope and delays from 0 to 12 sec

Figure 9—figure supplement 7. fMRI linearized encoding analysis with normal noise as null features and delays from 0 to 12 sec

Figure 9—figure supplement 8. fMRI linearized encoding analysis with uniform noise as null features and delays from 0 to 12 sec

Figure 9—figure supplement 9. fMRI transfer function weights with all three noise models and delays from 3 to 9 sec

Figure 9—figure supplement 10. fMRI transfer function weights with all three noise models and delays from 4 to 6 sec

Figure 9—figure supplement 11. fMRI transfer function weights with all three noise models and delays from 0 to 12 sec

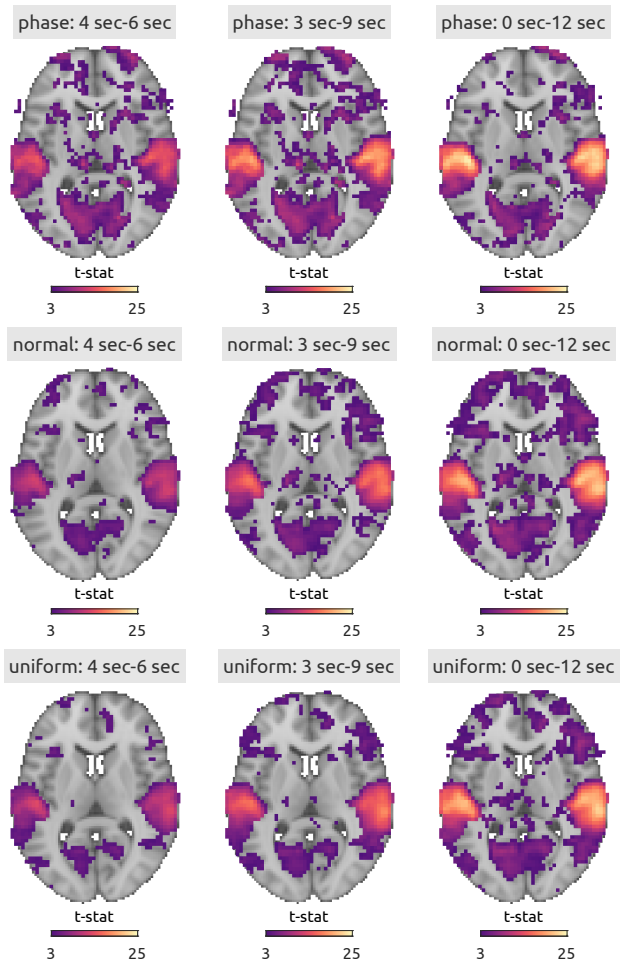


Figure 10. t-statistic maps on the transverse slices comparing the null prediction accuracies between two CV schemes (e.g., ***Figure 9*****g** vs. ***Figure 7*****p**) to test the RDD effects in the fMRI data with the phase-randomized envelope (top row), the normal noise (middle row), and the uniform noise (bottom row). Voxels were thresholded by statistical significance after FDR adjustment ($P_{FDR} < 0.01$). The background anatomical image is the MNI template included in FSL (`MNI152_T1_2mm_brain.nii.gz`). The 3-D volumes can be viewed with the NeuroVault web viewer (https://identifiers.org/neurovault.collection:19626).

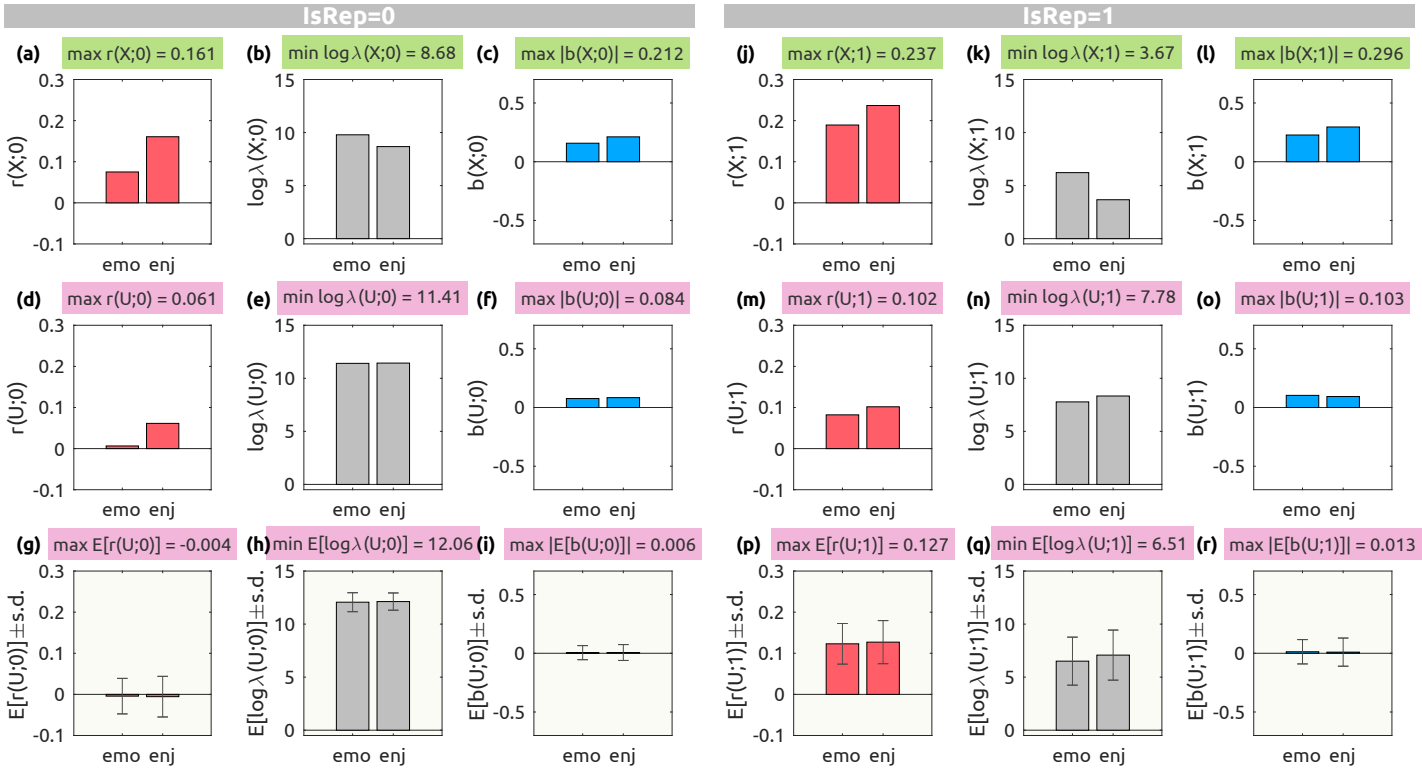


Figure 11. Behavioral linearized encoding analysis results with delays from 0 to 10 sec with an audio envelope (top row, **(a-c, j-l)**), a single case of the phase-randomized envelope (middle row, **(d-f, m-o)**), and an average of 100 phase-randomized envelopes (bottom row, pale yellow background, **(g-i, p-r)**). For each CV scheme (`IsRep` $= 0$, left panels, **(a-i)**; `IsRep` $= 1$, right panels, **(j-r)**), prediction accuracy (r, red bars, **(a, d, g, j, m, p)**), logarithmic ridge hyperparameter ($\log_{10} \lambda$, gray bars, **(b, e, h, k, n, q)**), transfer function weights that are summed over delays (b, blue bars, **(c, f, i, l, o, r)**) are shown along the columns. For the averaged metrics (bottom row), the standard deviations are shown as error bars. Emo.: emotionality, Enj.: enjoyment.

Figure 11—figure supplement 1. Behavioral linearized encoding analysis with the normal noise and delays from 0 to 10 sec

Figure 11—figure supplement 2. Behavioral linearized encoding analysis with the uniform noise and delays from 0 to 10 sec

Figure 11—figure supplement 3. Behavioral linearized encoding analysis with the phase-randomized envelope and delays from 0 to 5 sec

Figure 11—figure supplement 4. Behavioral linearized encoding analysis with the normal noise and delays from 0 to 5 sec

Figure 11—figure supplement 5. Behavioral linearized encoding analysis with the uniform noise and delays from 0 to 5 sec

Figure 11—figure supplement 6. Behavioral linearized encoding analysis with the phase-randomized envelope and delays from 0 to 15 sec

Figure 11—figure supplement 7. Behavioral linearized encoding analysis with the normal noise and delays from 0 to 15 sec

Figure 11—figure supplement 8. Behavioral linearized encoding analysis with the uniform noise and delays from 0 to 15 sec

Figure 11—figure supplement 9. Behavioral transfer function weights with all three noise models and delays from 0 to 10 sec

Figure 11—figure supplement 10. Behavioral transfer function weights with all three noise models and delays from 0 to 5 sec

Figure 11—figure supplement 11. Behavioral transfer function weights with all three noise models and delays from 0 to 15 sec

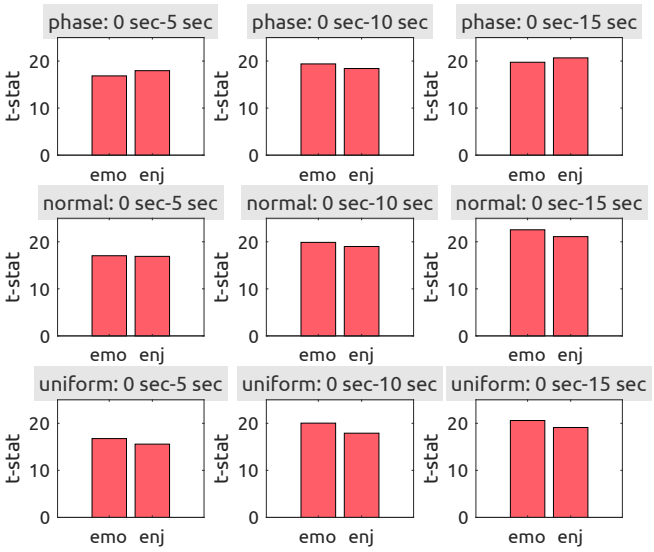


Figure 12. t-statistic bar plots comparing the null prediction accuracies between two CV schemes (e.g., ***Figure 11*****g** vs. ***Figure 7*****p**) to test the RDD effects in the behavioral data with the phase-randomzied envelope (top row), the normal noise (middle row), and the uniform noise (bottom row). All effects were statistical significant after FDR adjustment ($P_{FDR} < 0.01$). Emo.: emotionality, Enj.: enjoyment.

Lastly, the RDD effect was consistently observed across popular data modalities in cognitive neuroscience (***Figure 8***, ***Figure 10***, ***Figure 12***). In particular, the inflated prediction accuracy exhibited highly plausible spatial patterns even when predicted by uniform noise as a null feature. It is essential to emphasize that these patterns are driven by time-locked neural responses to repeated stimuli (widely known as inter-trial/-subject synchrony), not by random noise in the data. Therefore, when combined with *informal reverse inference* (i.e., falsely inferring a mental process from a brain activity pattern without accounting for base rates), which is also a common logical fallacy in cognitive neuroscience (***Poldrack, 2006***), RDD can lead to completely incorrect conclusions (e.g., *"The auditory cortex was encoding this uniform random noise that was never presented to the participant."*; ***Figure 9—figure Supplement 2***).

But my features are not just random noise!

As clearly demonstrated in the current paper, RDD can lead to spurious findings in cognitive neuroscience research when combined with questionable research practices such as adaptation of novel analysis frameworks without considering the original design of their data and informal reverse inference.

I used random noise to demonstrate that the RDD effect can be observed with a feature that is not expected to be encoded in the real data. Certainly, no one would sincerely expect the auditory cortex to encode a random feature that cannot be extracted from the stimulus. In practice, researchers hypothesize that certain information extracted from the stimulus is encoded in the response based on some theoretical and/or empirical grounds. However, RDD can inflate the prediction accuracy of their hypothesized feature well above the chance level even when it is not encoded. This, in turn, can misguide future research and contribute to contamination of the literature.

Also, note that the RDD effect was greater for the phase-randomized envelope than the normal or uniform noise (***Figure 8***, ***Figure 10***, ***Figure 12***). This is because the phase-randomized envelope preserves the autocorrelation structure of the stimulus, making it more similar to true features than random noise. Since the hypothesized feature is extracted from actual stimuli, it necessarily bears some resemblance to true features, regardless of the nonlinear operations involved. Therefore, the risk of RDD is greater when the hypothesized feature is derived from the stimulus rather than from random noise.

Often nested regression models are compared to determine the unique predictive contribution of a feature of interest. For example, in our previous study (***Leahy et al., 2021***), the prediction accuracy of a model with an audio envelope ("reduced model") was subtracted from the prediction accuracy of a model with the envelope and musical beats ("full model") to test the encoding of musical beats[6]. Even in such cases, RDD can still inflate the prediction accuracy of the hypothesized feature (or any additional random features) as shown in the simulation where additional random features moderately increased the RDD effect (***Table 8—source data 1***, the interaction between the number of features and stimulus repetition on null prediction: $\eta_p^2 = 0.101$). Once again, in a proper cross-validation without information leakage, irrelevant features would have been regularized. However, in the presence of RDD, the regularization is disabled, and the irrelevant features are not penalized, thus leading to an inflated prediction accuracy.

Is RDD really a different type of leakage than double-dipping?

All types of information leakage are essentially a circular fallacy (***Kaufman et al., 2012***). However, RDD is different from the more widely known type of circular fallacy—double dipping (***Kriegeskorte et al., 2009***)—in terms of the locus of the circularity. In double-dipping, it is the identical noise that is repeated in a selective analysis that uses the same dataset twice. The identical random noise appears once in setting up a selective analysis (e.g., channels or voxels of interest) and once again in testing the selective analysis, thus yielding spurious findings.

In reverse-double-dipping, it is the underlying signal that is repeated in a predictive analysis that uses apparently different datasets. Because the data points were acquired at different times with independent noise (but with the same stimuli), an identical underlying signal may seem less obvious. Because of the independence of noise, even researchers who are well-aware of the issue of double-dipping might not notice the circularity in their analyses.

Is RDD relevant to other analyses?

Given the deceptive nature of RDD, readers may wonder whether the displayed problem is specific to linearized encoding analysis or also relevant to other popular analyses in cognitive neuroscience. Here, I discuss the relevance of RDD to other analyses.

[6]While the comparison of nested models is a standard practice in Ordinary Least Squares (OLS) regression, it can be complicated with regularized regression such as ridge. The major problem is over-regularization of relevant features due to irrelevant features in the full model when a single penalty hyperparameter is used for all features (feature spaces). A multi-penalty model can address this issue better (e.g., ***La Tour et al., 2022***; ***Kim et al., 2024***).

Beta image encoding

Unlike the FIR model that estimates transfer function weights for each time point, the beta image encoding model is on top of the classical GLM that estimates 'beta' weights for each stimulus. This approach is popular in visual fMRI experiments where the gazing and visual attention of participants is difficult (or unnecessary) to resolve in time (***Kay et al., 2008***) or auditory fMRI experiments with short (1-2 seconds) stimuli (***Moerel et al., 2018***). Thus, instead of directly handling the autocorrelated BOLD time series, an average activation amplitude (i.e., beta weight) during a short trial is first estimated using a GLM, either using a theoretical transfer function called the 'canonical hemodynamic response function (cHRF)' (***Henson et al., 1999***) or by fitting a regularized FIR model on a split data (***Prince et al., 2022***).

For more general noise covariance structures, the beta estimation can be described as a Generalized Least Squares (GLS) problem (***Lage-Castellanos et al., 2019***):

$$\boldsymbol{\xi} \equiv \hat{\boldsymbol{\beta}} = \left(\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Omega}^{-1}\mathbf{y}, \tag{23}$$

where $\boldsymbol{\xi} \in \mathbb{R}^{M\times 1}$ is the estimated stimulus-response vector for M stimuli, $\boldsymbol{\Phi} \in \mathbb{R}^{T\times M}$ is the cHRF-convolved design matrix for T time points, $\Omega \in \mathbb{R}^{T\times T}$ is the autocovariance matrix of the noise in the BOLD time series, and $\mathbf{y} \in \mathbb{R}^{T\times 1}$ is the BOLD time series. Then, the estimated beta weights are subjected to an encoding model:

$$\boldsymbol{\xi} = \mathbf{F}\mathbf{g} + \mathbf{e}, \tag{24}$$

where $\mathbf{F} \in \mathbb{R}^{M\times F}$ is a matrix that describes F features for M presented stimuli, $\mathbf{g} \in \mathbb{R}^{F\times 1}$ is a feature-response vector of the voxel, and $\mathbf{e} \in \mathbb{R}^{M\times 1}$ is unknown noise.

Let us consider a case where two sets of beta images with M identical stimuli were partitioned into CV folds (e.g., even runs vs. odd runs where all M stimuli were presented in each run in randomized orders). Then, the expected null prediction accuracy by the Red Team would be also positive:

$$\mathbb{E}\left[\mathrm{corr}\left(\hat{\boldsymbol{\xi}}_{2,\mathbf{H}}, \boldsymbol{\xi}_2\right)\right] \overset{(18)}{\approx} \mathbf{s}_1^{\mathsf{T}}\mathbf{H}_1\left(\mathbf{H}_1^{\mathsf{T}}\mathbf{H}_1\right)^{-1}\mathbf{H}_1^{\mathsf{T}}\mathbf{s}_1 \geq 0, \tag{25}$$

where $\mathbf{s}_i = \mathbf{F}_i\mathbf{g} \in \mathbb{R}^{M\times 1}$ is the underlying signal pattern (a "response profile") for M stimuli in the i-th CV partition ($i = 1$, training; $i = 2$, testing), $\mathbf{H}_i \in \mathbb{R}^{M\times F}$ is the null feature matrix.

Although it is standard practice to randomize the presentation order of stimuli across runs and participants, the beta image estimation process can reorganize the response profiles, aligning them in a consistent order across all runs (e.g., sequentially from the first to the M-th stimuli). Therefore, the risk of RDD in the beta image encoding analysis remains a concern.

Stimulus reconstruction

Reconstruction of unseen stimuli (e.g., images or sounds) based on neural data demonstrates the remarkable potential of neuroimaging techniques for "mind-reading" (***Kay et al., 2008***; ***Santoro et al., 2017***; ***Han et al., 2019***). In practice, a set of linear models decodes features from neural data (e.g., beta images), followed by a reconstruction step where a simple classifier or a deep neural network (such as variational autoencoder) synthesizes the stimulus from the decoded features.

For instance, a multivariate decoding model based on beta images can be described as:

$$\mathbf{F} = \mathbf{W}\boldsymbol{\Xi} + \mathbf{e} \tag{26}$$

where $\mathbf{F} \in \mathbb{R}^{M \times F}$ is an F-dimensional feature matrix for M stimuli, $\mathbf{W} \in \mathbb{R}^{V \times F}$ is a spatial filter to decode the features, $\mathbf{\Xi} \in \mathbb{R}^{M \times V}$ is the matrix of vectorized beta images for M stimuli and V voxels. Since the decoding model is also a linear model like the encoding model, in principle, RDD can also occur. However, the goal of the analysis makes the requirement of "unseen stimuli" more explicitly. For example, if one fit a linear model with $N-1$ subjects' responses to Stimulus A, and then tries to recover, once again, Stimulus A from a response of the i-th subject, the problem of the circularity in this design would be more visible. Having said that, it is still possible that latent similarity of stimuli could be overlooked. Especially if there are too many stimuli to manually inspect, it is possible that certain stimuli are indeed non-identical but still highly similar (e.g., utterances by the same speaker; repetitions of musical phrases). Checking inter-stimulus correlation for all features prior to partitioning them into training and test sets will prevent such cases of hidden 'twinning'.

Multivariate classification

Multivariate classification analysis has been widely used in the cognitive neuroscience community, commonly known as multivoxel (or multivariate) pattern analysis (MVPCA; ***Kriegeskorte et al., 2006***) or single-trial classification in electrophysiological data such as EEG, MEG, and ECoG (***Müller-Gerking et al., 1999***; ***Pistohl et al., 2012***; ***Quandt et al., 2012***), either on the whole set of response units (e.g., whole-brain classification; ***Ryali et al., 2010***) or a striding set of local neighbors (e.g., a "searchlight"; ***Kriegeskorte et al., 2006***). While a highly accurate classifier is necessary to build a brain-computer interface system, classification analysis can be useful even with low but significant accuracy—often the case in cognitive neuroscience research—as a multivariate model that tests the existence of certain information in the brain.

In a simple case of two classes, a linear classifier can be constructed (***Hastie et al., 2009***) as:

$$C = \mathrm{sign}\left(\mathbf{a}^{\mathsf{T}}\mathbf{w} + w_0\right), / \tag{27}$$

where $\mathrm{sign} : \mathbb{R} \rightarrow \{-1, +1\}$ is a sign function, $\mathbf{a} \in \mathbb{R}^{V \times 1}$ is an activation pattern (e.g., an M/EEG topography or beta values in a region of interest) for an unknown class instance, $\mathbf{w} \in \mathbb{R}^{V \times 1}$ is a classification weight vector, w_0 is a scalar bias term, and $C \in \{-1, +1\}$ is a class label.

Classification can be seen as model-free as compared to model-based encoding or decoding analysis because the classification does not require a definition of a 'model' that describes which feature of the stimulus contributes to which response to what extent. Note that the trained weight vector only represents a separation of given training examples in a functional space, regardless of the features of the stimuli. For example, one may try to classify 'happy music' vs. 'sad music' based on the EEG responses. However, if the chosen exemplars of the classes are imbalanced (e.g., all exemplars of 'happy music' naturally happened to be faster and louder than 'sad music'), classification could be strongly driven by different acoustics, not necessarily due to perceived or evoked emotions from the music. That is, without carefully matching the training examples for all relevant features, the interpretation of classification analysis may remain unclear.

Moreover, in principle, the repetition of stimuli (or at least classes) across data points is not just unavoidable but in fact necessary for the classification. In order to train a classifier, not only linear but also general, balanced training and testing examples of all classes are required (***Hastie et al.,***

2009). In other words, it is impossible to train a classifier for an 'unseen class' while it can be trained for an 'unseen instance' of a known class. Therefore, it can be noted that RDD does not concern classification analysis although the forward RDD (i.e., double-dipping) greatly concerns the feature selection process of classification.

Representation similarity analysis

Finally, let us consider a method that evaluates the second-order isomorphism across representational systems (***Kriegeskorte et al., 2008***), as widely known as representational similarity analysis (RSA). This flexible method can define a 'model' distance matrix between stimuli either based on class labels (as in classification analysis) or feature descriptors (as in encoding analysis). The original formulation of RSA does not involve cross-validation as the RSA is neither a predictive model nor classification (***Kriegeskorte et al., 2008***). A later extension introduced a cross-validated, squared Mahalanobis distance estimator—known as "crossnobis"—to enhance both reliability and interpretability (***Diedrichsen and Kriegeskorte, 2017***). When estimating a crossnobis distance between two conditions in a leave-one-out cross-validation (LOOCV) scheme, it is assumed that the number of responses for both conditions remains balanced across all CV partitions. Thus, if identical stimuli are repeated across partitions as in case of RDD, then the crossnobis distance for the brain RDM would be biased by the stimulus-specific (not condition-specific) activity. However, even so, null descriptors would define a model RDM that is irrelevant to the brain RDM. Therefore, a false conclusion that "a brain region represents null information" cannot be wrongly supported by the association between the model and brain RDMs.

RSA primarily focuses on second-order association across RDMs, typically assessed using linear (non-negative) regression. However, this association has been seldomly tested for its generalizability. For instance, two model RDMs—one based on object identity (e.g., a human face vs. a house) and another based on pixel intensity—and a brain RDM from trials where a subject views these images. In a training set, the association weights can be estimated as

$$\mathbf{d}_{\text{brain}} = \beta_0 + \mathbf{d}_{\text{identity}}\beta_1 + \mathbf{d}_{\text{intensity}}\beta_2 \tag{28}$$

where $\mathbf{d}$ is a vector of flattened upper triangular elements excluding the diagonal of an RDM. In a test set, the relationship between RDMs for 'unseen' pictures can be predicted by the weights obtained from the training set, similarly to any predictive regression models (i.e., the third-order similarity). In such a case (i.e., a replication of RSA findings), the repetition of stimuli across training and test sets (i.e., 'unseen' pictures were actually 'seen' pictures) could inflate the prediction accuracy. However, introducing null features would disrupt the model RDMs and would greatly decrease the weights (i.e., $\hat{\beta}_i$) already in the training set, making it unlikely to observe a high prediction accuracy. In conclusion, the risk of RDD in RSA seems minimal.

How do we detect and prevent RDD?

The linearized encoding model has been utilized in various neuroimaging studies for many years (***Kay et al., 2008***), so it may be surprising that this pitfall has gone underrecognized for so long. In fact, many leading groups in this field have employed a hold-out validation design rather than cross-validation. In this approach, the testing dataset is deliberately acquired separately from the training

set by the experiment design. That is, training stimuli are repeated and averaged, while different stimuli are presented only once for training (e.g., ***Nishimoto et al., 2011***; ***Huth et al., 2016***; ***Han et al., 2019***). As a result, model predictions are evaluated only once per subject. This design prevents accidental stimulus repetition across CV partitions and thus avoids the risk of RDD. However, when the data is collected for a different purpose (e.g., intersubject synchrony) and researchers later apply encoding analysis, the likelihood of encountering RDD increases. This risk is particularly high for novice researchers who may misinterpret the valid recommendation to construct a 'subject-independent model' (***Crosse et al., 2021***), which involves cross-validation across subjects. Without careful implementation, they may inadvertently create a "stimulus-specific" model, as illustrated in ***Figure 6***, leading to RDD pitfall. Thus, in this section, I provide practical recommendations for detecting and preventing RDD.

Inter-trial correlation diagnosis

As repeatedly shown throughout the paper, RDD occurs when the identical stimulus is presented more than once across CV partitions. Thus, a simple way to test a CV design for the risk of RDD is to compute the correlation across data epochs before partitioning the data into training, optimization, and test sets. If the correlation is high, RDD is likely to occur. This was already illustrated in the toy example. When no stimulus was repeated across CV partitions the inter-trial correlation (ITC) was on average zero across 200 random samplings (***Figure 2***q). However, when the stimulus was repeated across CV partitions, the ITC was on average about 0.8 (***Figure 3***q). Because this correlation can be cheaply computed prior to costly optimization and modeling fitting, this can be a useful diagnostic tool to assess risk for RDD.

Checking ITC is also useful to detect latent similarity across stimuli as well. For example, two audio files are named differently but contain similar music, the researcher would not know about the leakage in the training examples. The ITC can be a useful tool to detect such latent similarity.

For users' convenience, an automatic validation test for a given CV design based on feature-ITC and response-ITC is implemented as a default option in the MATLAB package for Linearized Encoding Analysis (LEA; https://github.com/seunggookim/lea).

Hold-out validation

The most straightforward but also most expensive way to prevent RDD is to use a hold-out validation design. In this design, the testing dataset is deliberately acquired separately from the training set by the experiment design. For example, the training stimuli are repeated and averaged, while different stimuli are presented only once for training. This design prevents accidental stimulus repetition across training and test partitions and thus avoids the risk of RDD.

However, the hold-out validation requires more data points and higher SNR to work as compared to cross-validation. In cross-validation, the out-sample prediction performance is repeatedly assessed and averaged across different CV partition schemes to achieve an accurate estimation in the presence of sampling variance. In hold-out validation, the out-sample prediction performance is assessed only once. Thus, the sampling variance needs to be already suppressed to achieve a similarly accurate estimation of the true prediction performance.

Single-use stimulus

Another straightforward design to prevent RDD is not to re-use a stimulus during the whole study. That is, a stimulus is presented to only one participant, only once, and never used again (i.e., a single-use stimulus). This design prevents any accidental stimulus repetition across CV partitions and thus avoids the risk of RDD. That is, any CV partitioning (across stimulus or participants or both) cannot occur RDD.

However, this design may be not practical in most cases as it requires a large number of stimuli to be presented to each participant in order to ensure the sampling variance due to stimuli is sufficiently reduced. With a small number of stimuli, the attribution of the observed difference between participants would be ambiguous—to what extent should it be attributed to inter-subject variability and inter-stimulus variability?

Group-level modeling

In practice, acquiring neuroimaging data of high quality is still highly time-consuming and expensive. In particular, acquiring extensive data from vulnerable populations (e.g., patients, young children, elderly people) poses not only financial and logistical but also ethical concerns. Thus, many researchers may be motivated to combine the limited data across multiple subjects to further reduce the sampling variance.

As discussed, if an identical set of stimuli were used (as a standard procedure) across subjects, RDD could arise. In such a case, one option is to simply average all data across subjects within each group to test the group difference by cross-encoding (see below). If only a single long stimulus was used, multiple pseudo-trial segments should be created with appropriate gaps to avoid carry-over effect (e.g., for fMRI data, a gap of at least six seconds is recommended to account for the hemodynamic delay; for M/EEG data, a gap of one second or more), and design a CV across pseudo-trial segments. To detect potential repetition across segments, checking ITC of features is also recommended. While a one-size-fits-all hard threshold cannot be recommended because the temporal autocorrelation alters the expected variance of the null correlation, a strong correlation across segments (e.g., $r > 0.5$) may indicate a risk of RDD. In that case, design a different CV scheme or a different segmentation scheme.

A group-level inference is typically performed by testing the null hypothesis that the mean subject-wise prediction accuracy is indifferent between groups (e.g., clinical cases vs. healthy controls), using either a parametric test (e.g., a t-test based on the t-distribution) or a non-parametric test (e.g., a permutation test based on a null distribution generated by swapping group labels). When analyzing a group difference after averaging data within each group, a group difference can be tested by cross-encoding (i.e., predicting a mean response to the 'unseen' segment B in the 'unseen' patient group by the weights fitted to a mean response to the 'seen' segment A in the control group) with bootstrapping across subjects (i.e., resampling with replacement when creating the mean response). If the prediction is at or below chance level, the group difference can be concluded.

Limitations

While the current paper provides a comprehensive analysis of RDD, some limitations should be acknowledged.

First, only a single-penalty ridge regression was considered here. However, other types of regularization (e.g., multi-penalty ridge, LASSO, elastic net) and different modeling techniques (e.g., support vector regression, non-linear kernel ridge regression) were not explored. While the key mechanism of RDD (i.e., disabling regularization due to the similarity of the underlying signal between the training and optimization sets) is likely to be present in these other methods, the extent to which the RDD effect generalizes to these other methods remains an open question.

Also, no neural spike data or intracranial recordings—which is gradually becoming more accessible for human patients—were analyzed in this paper. The RDD effect was only shown in the context of continuous-valued data acquired from non-invasive methods (e.g., EEG, fMRI, and behavioral ratings). Sparse spike data or discrete firing rate data may behave differently from continuous-valued data. However, given that the RDD effect is driven by the similarity of the underlying signal, RDD is likely to occur when the data is transformed in such a way that stimulus-evoked, time-locked response is often demonstrated (e.g., high gamma power envelope of a local population of neurons; ***Ray et al., 2008***; ***Jacobs and Kahana, 2009***).

Conclusion

The current paper shows that RDD is a critical but underrecognized risk in encoding analysis and stimulus reconstruction. This circular fallacy—which I named *reverse double-dipping*—may lead to spurious findings that irrelevant information is encoded in neural signal. When carefully designed, the model-based approach will remain a powerful tool for information-based neuroimaging.

Materials and Methods

Simulation methods

Predictors

True features were sampled from an F-dimensional multivariate Gaussian distribution with a correlation between adjacent parameters as: $\mathbf{x} \sim \mathcal{N}_F(\mathbf{0}, \mathbf{\Sigma}_X)$ where the covariance is defined as:

$$\mathbf{\Sigma}_X = \begin{bmatrix} 1 & \rho_X & \rho_X^2 & \cdots & \rho_X^{F-1} \\ \rho_X & 1 & \rho_X & \cdots & \rho_X^{F-2} \\ \rho_X^2 & \rho_X & 1 & \cdots & \rho_X^{F-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_X^{F-1} & \rho_X^{F-2} & \rho_X^{F-3} & \cdots & 1 \end{bmatrix}. \tag{29}$$

The f-th feature $\mathbf{x}^{(f)} = [x_1^{(f)}, x_2^{(f)}, \ldots, x_T^{(f)}]^\mathsf{T}$ is given with an AR(1) temporal autocorrelation ϕ_X as:

$$x_t^{(f)} = x_t^{(f)} + \phi_X x_{t-1}^{(f)}. \tag{30}$$

Null features $\mathbf{u}$ were created in the same way as $\mathbf{x}$, but independently. Only causal (i.e., nonnegative) delays were considered from $\{0\}$ to $\{0, \ldots, D-1\}$ in creating a design matrix by horizontally

concatenating Toeplitz matrices as:

$$\mathbf{X} = \begin{bmatrix} x_1^{(1)} & \cdots & x_{1-(D-1)}^{(1)} & x_1^{(2)} & \cdots & x_{1-(D-1)}^{(2)} & \cdots & x_1^{(F)} & \cdots & x_{1-(D-1)}^{(F)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ x_T^{(1)} & \cdots & x_{T-(D-1)}^{(1)} & x_T^{(2)} & \cdots & x_{T-(D-1)}^{(2)} & \cdots & x_T^{(F)} & \cdots & x_{T-(D-1)}^{(F)} \end{bmatrix}. \tag{31}$$

For convenience, all predictors were standardized to zero-mean and unit variance.

Responses

Responses $\mathbf{Y}_i \in \mathbb{R}^{T \times V}$ were generated by plugging in a design matrix $\mathbf{X}_i$, true weights $\mathbf{B} \in \mathbb{R}^{FD \times V}$, and noise $\mathbf{E}_i \in \mathbb{R}^{T \times V}$ to the FIR model ***Equation 1*** for the i-th partition as $\mathbf{Y}_i = \mathbf{X}_i\mathbf{B} + \mathbf{E}_i$ where $\mathbf{Y}_i = \left[\mathbf{y}_i^{(1)}, \mathbf{y}_i^{(2)}, \cdots, \mathbf{y}_i^{(V)}\right]$, $\mathbf{B} = \left[\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \cdots, \mathbf{b}^{(V)}\right]^\mathsf{T}$, and $\mathbf{E}_i = \left[\mathbf{e}_i^{(1)}, \mathbf{e}_i^{(2)}, \cdots, \mathbf{e}_i^{(V)}\right]$.

True weights of the v-th variate $\mathbf{b}^{(v)}$ (a column of a response variable) were created from the multivariate Gaussian distribution with an AR(1) temporal correlation ϕ_B along the delays and a covariance σ_b across predictors corresponding to the rows of the design matrix as:

$$\mathbf{b}^{(v)} = \left[b_{1,1}^{(v)}, \cdots, b_{1,D}^{(v)}, b_{2,1}^{(v)}, \cdots, b_{2,D}^{(v)}, \cdots, b_{F,1}^{(v)}, \cdots, b_{F,D}^{(v)}\right]^\mathsf{T} \tag{32}$$

where $b_{f,d}^{(v)}$ is a scalar coefficient for the v-th variate, the f-th feature, and the d-th delay. Noise $\mathbf{e}^{(v)}$ at the v-th variate was also created as Gaussian noise with an AR(1) temporal correlation ϕ_E as $e_t^{(v)} = e_t^{(v)} + \phi_E e_{t-1}^{(v)}$. In case of multivariate models ($V \geq 2$), weights and noise have spatial autocorrelation: $\mathbf{b}^{(v)} = \mathbf{b}^{(v)} + \theta_B\mathbf{b}^{(v-1)} + \theta_B\mathbf{b}^{(v+1)}$ and $\mathbf{e}^{(v)} = \mathbf{e}^{(v)} + \theta_E\mathbf{e}^{(v-1)} + \theta_E\mathbf{e}^{(v+1)}$ where a variate v is a neighbor of $v-1$ and $v+1$ except for boundaries ($v = 1$ and $v = V$).

Before summing the signal $\mathbf{XB}$ and error $\mathbf{E}$, the mean variance of error across variates was scaled so that it achieved the intended signal-to-noise ratio as $S = 10\log_{10}\sigma_S^2/\sigma_E^2$. Then, all response variates were standardized to zero-mean and unit-variance.

Optimization and evaluation

Ridge parameters were optimized via a grid search (λ-grid = $10^{[-10,-9,\ldots,10]}$) for each variate independently. To implement a nested 4-by-3-fold cross-validation (CV), in total 4 "trials" were generated for each random sampling. For the 4-fold-outer-loop, 3 trials (outer-training) vs. 1 trial (outer-test) were partitioned. Then, within each 3-fold-inner-loop with the 3 trials, 2 trials (inner-training) vs. 1 trial (inner-optimization) were partitioned. In total 12 different partitions (each was called a "CV-fold") were used for training (50%), optimization (25%), and test (25%) models. Prediction accuracy of Pearson correlation was averaged across the CV-folds. Random sampling was repeated for 1000 times for each combination of model parameters.

Computational considerations

To minimize the number of inversion operations, a general linear model (GLM) formulation was used. That is, for each possible λ, a regularized covariance matrix was inverted only once for all variates (i.e., voxels or channels). The sum of squared errors was then temporally stored, allowing the optimal λ to be determined for each variate. This approach is more efficient than a naïve method of inverting the covariance matrix separately for each variate, which would redundantly repeat the same calculation for all variates. Also, when computing Predictions, variates with the same optimal λ were grouped together into GLM models to reduce the number of inversions.

Computation was carried out using an in-house high-performance computing (HPC) server, where a user is allowed to utilize up to 192 CPUs of Intel Xeon Gold 6130 [2.10 GHz] in parallel. The actual utilization varied between 32–192 CPUs, depending on the demands of other users. An individual job of 1,000 random samplings took 100–400 seconds of CPU time. A total of 45,899 jobs, which amounted to approximately 7 months of CPU time, was completed in about one week on the HPC server.

Real data methods

For clarity, the original studies (***Kaneshiro et al., 2020***; ***Sachs et al., 2020***) focused on inter-subject synchrony in neural responses (EEG and fMRI) without specific hypotheses about the encoded features. Thus, these studies did not suffer from RDD.

EEG data

Data source The original study investigated the electroencephalographic correlates of temporal structure and beat in Western-style Indian pop music with Hindi lyrics (i.e., Bollywood music) (***Kaneshiro et al., 2020***). The raw dataset of 48 healthy participants was downloaded from the Stanford Digital Repository (https://purl.stanford.edu/sd922db3535).

Data acquisition Scalp electrical potential data were acquired using a 128-channel Geodesic EEG System 300 (Electrical Geodesics, Inc., Oregon, USA) at a sampling rate of 1 kHz with vertex reference and electrode impedances less than 60 kΩ.

Preprocessing The authors' preprocessed data (`CleanEEG_*`) from the repository were used with the channel location file (`GSN-HydroCel-125.sfp`). As explained in the original publication (***Kaneshiro et al., 2020***), the authors' preprocessing steps involved bandpass filtering (0.3–50 Hz), downsampling (125 Hz), ocular artifacts removal using independent component analysis, bad channel interpolation, average re-referencing, and epoching. In addition, based on previous EEG encoding analyses where the low-frequency bands of the EEG signal mostly carried the audio envelope encoding (***Di Liberto et al., 2015***, ***2020***), we further bandpass filtered the EEG data to δ and θ bands (1–8 Hz) using the FIR filter in the MATLAB Signal Processing Toolbox (R2022b, RRID:SCR_001622).

Data dimensions The analyzed EEG data were comprised of 125 channels, 32,878–33,982 time points (263.02–271.86 seconds at 125 Hz), 96 runs (48 subjects × 2 run) per stimulus with 12 stimuli (3 versions of 4 songs). Two runs with the same stimuli were averaged for each participant to increase the signal-to-noise ratio of the evoked response. In the original study (***Kaneshiro et al., 2020***), a participant was randomly assigned to one of 4 versions of th 4 songs (e.g., Intact-Song-A, Measure-shuffled-Song-B, Reserved-Song-C, Phase-scrambled-Song-D). The current study only analyzed the 3 versions except for phase-scrambled version, which did not evoke strong responses.

fMRI data

Data source The original study investigated the intersubject correlation in fMRI time series while participants were listening to sad music (***Sachs et al., 2020***). The published data include 2 sad musical pieces and 1 happy musical piece. The raw dataset of 39 healthy participants was

downloaded from OpenNeuro (RRID:SCR_005031, https://openneuro.org/datasets/ds003085/versions/1.0.0).

Data acquisition Blood-oxygen-level-dependent (BOLD) signals were acquired using a 3-T Prisma magnetic resonance imaging system with a 32-channel head coil (Siemens Healthineers, Erlangen, Germany). Eight-fold accelerated multiband, T2*-weighted, gradient-echo, echo-planar imaging sequence was used to acquire 40 transverse slices at the sampling rate of 1 Hz and at the isotropic spatial resolution of 3 mm. Anatomical scans were also collected using a T1-weighted contrast sequence at the isotropic spatial resolution of 1 mm.

Preprocessing After correcting for the susceptibility artifacts using the reversed phase-encoding images with `topup` in FSL (v6.0.2; https://fsl.fmrib.ox.ac.uk/fsl/), SPM12 (v6225; https://www.fil.ion.ucl.ac.uk/spm/software/spm12/) was used for slice timing correction and realignment of the functional 4-D time series data. Advanced Normalization Tools (v2.3.5; https://github.com/ANTsX/ANTs) were used to perform symmetric diffeomorphic registration between individuals' anatomical 3-D images and the standard template 3-D image (`MNI152_T1_2mm_brain.nii.gz`) from FSL as well as the coregistration (rigid-body affine transform) between the anatomical 3-D image and the temporally averaged functional 3-D image within each subject. The rigid-body and diffeomorphic transformations were combined and applied to each volume of the realigned 4-D functional image, which was resampled only once at the isotropic resolution of 3 mm. Thereafter, the functional images were spatially smoothed with an isotropic Gaussian kernel (with a full width at half maximum of 6 mm). ICA-AROMA (v0.4.4.beta) was used to automatically reject 'noise' components to attenuate head motion artifacts and non-BOLD image intensity perturbations (https://github.com/maartenmennes/ICA-AROMA).

Data dimensions The analyzed fMRI data consisted of 62,062 "brain" voxels (based on a brain mask created using `bet` in FSL), 178–525 time points (or seconds), 3 stimuli, and 39 subjects. Due to limited field of views in some subjects, only 61,572 voxels had valid values in the linearized encoding analysis results. The authors of the original study excluded 3 subjects from their analysis for either motion artifacts and emotional ratings but included in the shared data repository. In the current analysis, all 39 subjects were analyzed for no apparent motion artifacts after ICA-AROMA.

Behavioral data

Data source As part of the fMRI study (***Sachs et al., 2020***), behavioral data of 39 healthy participants were downloaded from OpenNeuro (https://openneuro.org/datasets/ds003085/versions/1.0.0).

Data acquisition After the fMRI session, participants listened to the same stimuli while rating their evoked instantaneous emotions using a physical slider ("fader") in a silent room. Emotional scales of *Emotionality* (the intensity of the evoked feelings of sadness or happiness, depending on the intended emotion of each piece) and *Enjoyment* (the momentary feelings of enjoyment) were rated, one at a time. In total, a participant listened to an identical stimulus for three times including the fMRI session. The slider position values (an integer from 0 to 127) were sampled at about 30.3 Hz.

Preprocessing The imported time series were linearly detrended and downsampled at 5 Hz after an anti-aliasing low-pass filtering using the `resample` function in the MATLAB Signal Processing Toolbox (R2022b).

Data dimensions The analyzed behavioral data comprised 2 emotional scales, 841–2,576 time points depending on the stimulus (178–525 seconds at 5 Hz), 3 stimuli, and 39 subjects.

Linearized encoding analysis

Features True features $\mathbf{X}$ were the audio envelope extracted from the music samples for the well-established auditory response in various types of human brain data including fMRI (***Giraud et al., 2000***; ***Harms et al., 2005***; ***Overath et al., 2012***), EEG (***Aiken and Picton, 2008***), and intracranial electrocorticography (***Kubanek et al., 2013***). The envelope was created by summing the output of the "cochlear model" with a 128-channel filterbank, of which characteristic frequencies ranged loglinearly from 180 to 7,040 Hz, as in the NSL Auditory-cortical MATLAB Toolbox (v2001; http://nsl.isr.umd.edu/downloads.html), and then further downsampled to the sampling rate of the human data (EEG: 125 Hz; fMRI: 1 Hz, behavioral ratings: 5 Hz). Null features $\mathbf{U}$ were either (a) uniform noise, (b) normal noise, or (c) the phase-randomized envelope. Phase-randomization preserves the amplitude spectrum of the envelope, thus preserves the temporal autocorrelation. This is necessary to correctly estimate the null correlation distribution of the autocorrelated noise. In previous studies (***Leahy et al., 2021***; ***Kaneshiro et al., 2020***), phase-randomization was used to generate an empirical null distribution for non-parameteric statistical inference. The phase randomization was done using fast Fourier transform (FFT), random rotation of phases while preserving complex conjugation for positive and negative frequencies, and followed by an inverse FFT. While this method ('unwindowed Fourier transform' in ***Theiler et al., 1992***) may introduce spurious high frequencies for non-stationary signals, the envelope of stimuli that include zeros (i.e., silent periods) at the beginning and end of the musical stimuli can be assumed as stationary (i.e., the values of the last samples are similar to the first samples). To estimate the central tendency, 100 noise realizations were created for all cases. All features (either true or null) were standardized prior to fitting.

FIR modeling As in the simulations, an FIR model was regularized by ridge hyperparameters that are specific for response units (i.e., EEG channels, fMRI voxels, rating scales). Accounting for the inherent time scales of the measures, different delays of the features were used. To demonstrate the effect of the model's flexibility, 3 cases of delays (common choice, shorter, and longer) were used. For EEG, delays of [0 to 0.3 sec; 39 samples], [0 to 0.5 sec; 64 samples], and [0 to 1 sec; 126 samples] were used. For fMRI, [4 to 6 sec; 3 samples], and [3 to 9 sec; 7 samples], [0 to 12 sec; 13 samples]. For behaviors, [0 to 5 sec; 26 samples], [0 to 10 sec; 51 samples], and [0 to 15 sec; 76 samples]. To avoid transient onset/offset effects at the boundaries of the musical stimuli, the first and last 15 seconds were excluded from the analysis. Both stimuli and responses were standardized before the FIR modeling.

Cross-validation In all cases, the CV partition was 3-by-2 nested k-fold (i.e., 33%, 33%, 33% for training, optimization, test sets) for the given structures of the real data (i.e., 3 stimuli per

participant). To compare CV schemes, prediction accuracies were averaged across CV folds and models (either subject-specific or stimulus-specific).

Statistical inference The RDD effect is defined as the difference between null prediction accuracies: $\mathrm{RDD} = \bar{r}_{\mathrm{stim}}(\mathbf{U};\, \mathtt{IsRep} = 1) - \bar{r}_{\mathrm{subj}}(\mathbf{U};\, \mathtt{IsRep} = 0)$. The null hypothesis is that the expected RDD effect is zero $\mathcal{H}_0 : \mathbb{E}(\mathrm{RDD}) = 0$ and the alterative hypothesis is that RDD effect is positive $\mathcal{H}_A : \mathbb{E}(\mathrm{RDD}) > 0$. Non-parameteric P-values were computed by permutation test ($K =$ 10,000). That is, for 200 (100 randomizations $\times$ 2 CV-designs) vectors of prediction accuracies, the binary variable `IsRep` was randomly permuted ($\max = C(200, 100) > 10^{57}$), and the two-sample t-statistics between two CV-designs were calculated for 10,000 times to form a null distribution. Resulting one-sided P-values were further corrected for the multiple response units using false discovery rate (FDR) adjustment (***Yekutieli and Benjamini, 1999***) to control the family-wise error rate as $P_{\mathrm{FDR}} < 0.01$.

Software implementation All encoding analyses of the multimodal datasets (EEG, fMRI, behavior) were done using an MATLAB package named Linearized Encoding Analysis (LEA; https://github.com/seunggookim/lea), developed by the author.

Acknowledgments

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The author confirms being the sole contributor of this work and approved it for publication. Max Planck Society supported this work. The author thanks Dr. Daniela Sammler, Dr. Vincent Shi-chen Chien, Dr. Vincent K. M. Cheung, and Mr. Seung-Cheol Baek for their constructive feedback and expert insights on an earlier version of the manuscript. The author acknowledges the free image sources from www.irasutoya.com (Takashi Mifune), which reserves the copyrights of the source illustrations.

Data and code availability

All analysis code, simulation data, and real data analysis results are available at Zenodo: https://doi.org/10.5281/zenodo.15100830. An executable demo of simulation is available at CodeOcean: https://codeocean.com/capsule/4591394/. The EEG raw dataset is available at Standford Digital Repository: https://purl.stanford.edu/sd922db3535. The fMRI and behavioral raw datasets are available at OpenNeuro: https://openneuro.org/datasets/ds003085. A MATLAB package for Linearized Encoding Analysis (LEA) on multimodal data is available at Zenodo release: https://doi.org/10.5281/zenodo.15107756 and GitHub: https://github.com/seunggookim/lea. An executable demo of LEA on real data is available at MATLAB Online: https://s.gwdg.de/7cOQmw. fMRI analysis results in 3-D NIfTI format can be viewed and downloaded at NeuroVault: https://identifiers.org/neurovault.collection:19626.

Declaration of AI-assisted technologies in the writing process

During the preparation of this work, the author, as a non-native English speaker, used ChatGPT-4o (OpenAI) via GitHub Copilot to correct typographical and grammatical errors. The author reviewed and edited the content as needed and takes full responsibility for the final manuscript.

References

Aiken SJ, Picton TW. Human cortical responses to the speech envelope. Ear and hearing. 2008; 29(2):139–157. https://doi.org/10.1097/AUD.0b013e31816453dc.

Allen EJ, St-Yves G, Wu Y, Breedlove JL, Prince JS, Dowdle LT, Nau M, Caron B, Pestilli F, Charest I, et al. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. Nature neuroscience. 2022; 25(1):116–126. https://doi.org/10.1038/s41593-021-00962-x.

Baldassano C, Hasson U, Norman KA. Representation of real-world event schemas during narrative perception. Journal of Neuroscience. 2018; 38(45):9689–9699. https://doi.org/10.1523/jneurosci.0251-18.2018.

Brunswik E. Organismic achievement and environmental probability. Psychological review. 1943; 50(3):255. https://doi.org/10.1037/h0060889.

Caucheteux C, Gramfort A, King JR. Evidence of a predictive coding hierarchy in the human brain listening to speech. Nature human behaviour. 2023; 7(3):430–441. https://doi.org/10.1038/s41562-022-01516-2.

Chi T, Ru P, Shamma SA. Multiresolution spectrotemporal analysis of complex sounds. The Journal of the Acoustical Society of America. 2005; 118(2):887–906. https://doi.org/10.1121/1.1945807.

Crosse MJ, Zuk NJ, Di Liberto GM, Nidiffer AR, Molholm S, Lalor EC. Linear modeling of neurophysiological responses to speech and other continuous stimuli: methodological considerations for applied research. Frontiers in neuroscience. 2021; 15:705621.

Di Liberto GM, O'sullivan JA, Lalor EC. Low-frequency cortical entrainment to speech reflects phoneme-level processing. Current Biology. 2015; 25(19):2457–2465. https://doi.org/10.1016/j.cub.2015.08.030.

Di Liberto GM, Pelofi C, Bianco R, Patel P, Mehta AD, Herrero JL, de Cheveigné A, Shamma S, Mesgarani N. Cortical encoding of melodic expectations in human temporal cortex. eLife. 2020 mar; 9:e51784. https://doi.org/10.7554/elife.51784.

Diedrichsen J, Kriegeskorte N. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. PLoS computational biology. 2017; 13(4):e1005508.

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. https://doi.org/10.1002/hbm.460020402.

Gibson JJ. The Ecological Approach to the Visual Perception of Pictures. Leonardo. 1978; 11(3):227–235. https://doi.org/10.2307/1574154.

Giraud AL, Lorenzi C, Ashburner J, Wable J, Johnsrude I, Frackowiak R, Kleinschmidt A. Representation of the temporal envelope of sounds in the human brain. Journal of neurophysiology. 2000; 84(3):1588–1598. https://doi.org/10.1152/jn.2000.84.3.1588.

Hamilton LS, Huth AG. The revolution will not be controlled: natural stimuli in speech neuroscience. Language, Cognition and Neuroscience. 2020; 35(5):573–582. https://doi.org/10.1080/23273798.2018.1499946.

Han K, Wen H, Shi J, Lu KH, Zhang Y, Fu D, Liu Z. Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex. NeuroImage. 2019; 198:125–136.

Harms MP, Guinan Jr JJ, Sigalovsky IS, Melcher JR. Short-term sound temporal envelope characteristics determine multisecond time patterns of activity in human auditory cortex as shown by fMRI. Journal of neurophysiology. 2005; 93(1):210–222. https://doi.org/10.1152/jn.00712.2004.

977 **Hasson U**, Nir Y, Levy I, Fuhrmann G, Malach R. Intersubject Synchronization of Cortical Activity During Natural
978 Vision. Science. 2004; 303(5664):1634–1640. https://doi.org/10.1126/science.1089506.

979 **Hastie T**, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference,
980 and prediction, vol. 2. Springer; 2009. https://doi.org/10.1007/978-0-387-84858-7.

981 **Haxby JV**, Guntupalli JS, Nastase SA, Feilong M. Hyperalignment: Modeling shared information encoded in
982 idiosyncratic cortical topographies. eLife. 2020 jun; 9:e56601. https://doi.org/10.7554/elife.56601.

983 **Henson RNA**, Rugg MD, Shallice T, Josephs O, Dolan RJ. Recollection and Familiarity in Recognition Memory: An
984 Event-Related Functional Magnetic Resonance Imaging Study. Journal of Neuroscience. 1999; 19(10):3962–
985 3972.

986 **Hoerl AE**, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics.
987 1970; 12(1):55–67. https://doi.org/10.1080/00401706.1970.10488634.

988 **Huth AG**, De Heer WA, Griffiths TL, Theunissen FE, Gallant JL. Natural speech reveals the semantic maps that
989 tile human cerebral cortex. Nature. 2016; 532(7600):453–458.

990 **Jacobs J**, Kahana MJ. Neural representations of individual stimuli in humans revealed by gamma-band electro-
991 corticographic activity. Journal of neuroscience. 2009; 29(33):10203–10214.

992 **Kaneshiro B**, Nguyen DT, Norcia AM, Dmochowski JP, Berger J. Natural music evokes correlated EEG responses
993 reflecting temporal structure and beat. NeuroImage. 2020; 214:116559. https://doi.org/10.1016/j.neuroimage.
994 2020.116559.

995 **Kaufman S**, Rosset S, Perlich C, Stitelman O. Leakage in data mining: Formulation, detection, and avoidance.
996 ACM Trans Knowl Discov Data. 2012 dec; 6(4). https://doi.org/10.1145/2382577.2382579.

997 **Kay KN**. Principles for models of neural information processing. NeuroImage. 2018; 180:101–109.

998 **Kay KN**, Naselaris T, Prenger RJ, Gallant JL. Identifying natural images from human brain activity. Nature. 2008;
999 452(7185):352–355.

1000 **Kim SG**. On the encoding of natural music in computational models and human brains. Frontiers in Neuro-
1001 science. 2022; 16. https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2022.928841, doi:
1002 10.3389/fnins.2022.928841.

1003 **Kim SG**, De Martino F, Overath T. Linguistic modulation of the neural encoding of phonemes. Cerebral Cortex.
1004 2024; 34(4):bhae155.

1005 **Koelsch S**. A coordinate-based meta-analysis of music-evoked emotions. NeuroImage. 2020; 223:117350.

1006 **Kriegeskorte N**, Goebel R, Bandettini P. Information-based functional brain mapping. Proceedings of the
1007 National Academy of Sciences. 2006; 103(10):3863–3868.

1008 **Kriegeskorte N**, Mur M, Bandettini PA. Representational similarity analysis-connecting the branches of sys-
1009 tems neuroscience. Frontiers in systems neuroscience. 2008; 2:249.

1010 **Kriegeskorte N**, Simmons WK, Bellgowan PSF, Baker CI. Circular analysis in systems neuroscience: the dangers
1011 of double dipping. Nature Neuroscience. 2009; 12(5):535–540. https://doi.org/10.1038/nn.2303.

1012 **Kubanek J**, Brunner P, Gunduz A, Poeppel D, Schalk G. The tracking of speech envelope in the human cortex.
1013 PloS one. 2013; 8(1):e53398. https://doi.org/10.1371/journal.pone.0053398.

1014 **La Tour TD**, Eickenberg M, Nunez-Elizalde AO, Gallant JL. Feature-space selection with banded ridge regression.
1015 NeuroImage. 2022; 264:119728.

Lage-Castellanos A, Valente G, Formisano E, De Martino F. Methods for computing the maximum performance of computational models of fMRI responses. PLOS Computational Biology. 2019 03; 15(3):1–25.

Leahy J, Kim SG, Wan J, Overath T. An analytical framework of tonal and rhythmic hierarchy in natural music using the multivariate temporal response function. Frontiers in Neuroscience. 2021; 15:665767.

Moerel M, De Martino F, Kemper VG, Schmitter S, Vu AT, Uğurbil K, Formisano E, Yacoub E. Sensitivity and specificity considerations for fMRI encoding, decoding, and mapping of auditory cortex at ultra-high field. NeuroImage. 2018; 164:18–31.

Müller-Gerking J, Pfurtscheller G, Flyvbjerg H. Designing optimal spatial filters for single-trial EEG classification in a movement task. Clinical Neurophysiology. 1999; 110(5):787–798.

Naselaris T, Kay KN, Nishimoto S, Gallant JL. Encoding and decoding in fMRI. Neuroimage. 2011; 56(2):400–410.

Nastase SA, Goldstein A, Hasson U. Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. NeuroImage. 2020; 222:117254. https://www.sciencedirect.com/science/article/pii/S1053811920307400, doi: https://doi.org/10.1016/j.neuroimage.2020.117254.

Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL. Reconstructing visual experiences from brain activity evoked by natural movies. Current biology. 2011; 21(19):1641–1646.

Nuzzo R. How scientists fool themselves –and how they can stop. Nature. 2015; 526(7572):182–185. https://doi.org/10.1038/526182a, doi: 10.1038/526182a.

Overath T, Zhang Y, Sanes DH, Poeppel D. Sensitivity to temporal modulation rate and spectral bandwidth in the human auditory system: fMRI evidence. Journal of Neurophysiology. 2012; 107(8):2042–2056. https://doi.org/10.1152/jn.00308.2011, doi: 10.1152/jn.00308.2011, pMID: 22298830.

Pistohl T, Schulze-Bonhage A, Aertsen A, Mehring C, Ball T. Decoding natural grasp types from human ECoG. NeuroImage. 2012; 59(1):248–260.

Poldrack RA. Can cognitive processes be inferred from neuroimaging data? Trends in cognitive sciences. 2006; 10(2):59–63.

Prince JS, Charest I, Kurzawski JW, Pyles JA, Tarr MJ, Kay KN. Improving the accuracy of single-trial fMRI response estimates using GLMsingle. Elife. 2022; 11:e77599.

Quandt F, Reichert C, Hinrichs H, Heinze HJ, Knight RT, Rieger JW. Single trial discrimination of individual finger movements on one hand: A combined MEG and EEG study. NeuroImage. 2012; 59(4):3316–3324.

Ray S, Crone NE, Niebur E, Franaszczuk PJ, Hsiao SS. Neural correlates of high-gamma oscillations (60–200 Hz) in macaque local field potentials and their potential implications in electrocorticography. Journal of Neuroscience. 2008; 28(45):11526–11536.

Ryali S, Supekar K, Abrams DA, Menon V. Sparse logistic regression for whole-brain classification of fMRI data. NeuroImage. 2010; 51(2):752–764.

Sachs ME, Habibi A, Damasio A, Kaplan JT. Dynamic intersubject neural synchronization reflects affective responses to sad music. NeuroImage. 2020; 218:116512. https://doi.org/10.1016/j.neuroimage.2019.116512.

Santoro R, Moerel M, Martino FD, Valente G, Ugurbil K, Yacoub E, Formisano E. Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. Proceedings of the National Academy of Sciences. 2017; 114(18):4799–4804.

Sonkusare S, Breakspear M, Guo C. Naturalistic Stimuli in Neuroscience: Critically Acclaimed. Trends in Cognitive Sciences. 2019; 23(8):699–714. https://www.sciencedirect.com/science/article/pii/S1364661319301275, doi: https://doi.org/10.1016/j.tics.2019.05.004.

Theiler J, Eubank S, Longtin A, Galdrikian B, Doyne Farmer J. Testing for nonlinearity in time series: the method of surrogate data. Physica D: Nonlinear Phenomena. 1992; 58(1):77–94. https://www.sciencedirect.com/science/article/pii/016727899290102S, doi: https://doi.org/10.1016/0167-2789(92)90102-S.

Varoquaux G. Cross-validation failure: Small sample sizes lead to large error bars. NeuroImage. 2018; 180:68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061.

Wu MCK, David SV, Gallant JL. Complete functional characterization of sensory neurons by system identification. Annu Rev Neurosci. 2006; 29(1):477–505. https://doi.org/10.1146/annurev.neuro.29.051605.113024.

Yekutieli D, Benjamini Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. Journal of Statistical Planning and Inference. 1999; 82(1):171–196. https://doi.org/10.1016/S0378-3758(99)00041-5.

Table 1. Definition of variables, parameters, and notations.

Variables	Description
$\mathbf{X} \in \mathbb{R}^{T \times FD}$	True finite impulse response design matrix
$\mathbf{U} \in \mathbb{R}^{T \times FD}$	Null finite impulse response design matrix
$\mathbf{y} \in \mathbb{R}^{T \times 1}$	Response vector
$\mathbf{b} \in \mathbb{R}^{FD \times 1}$	Regression coefficient vector
$\mathbf{s} \in \mathbb{R}^{T \times 1}$	Signal time series
$\mathbf{e} \in \mathbb{R}^{T \times 1}$	Noise time strategies
$\mathbf{L} \in \mathbb{R}^{FD \times FD}$	Tikhonov regularization matrix
Parameters	**Description**
T	Number of time points
F	Number of features
D	Number of delays
P	Number of predictors (= FD)
ϕ	AR(1) temporal correlation parameter
θ	AR(1) spatial correlation parameter
ρ	Multicollinearity parameter
λ	regularization parameter
Notations	**Description**
$\|\cdot\|$	l_2-norm of a vector
$\|\cdot\|_{\mathrm{F}}$	Frobenius norm of a matrix
$(\cdot)^{\mathsf{T}}$	Transposition
$(\cdot)_{(i)}$	i-th column vector of a matrix
$\mathrm{diag}(\cdot)$	Diagonal matrix with the diagonal elements from the given vector
$\mathcal{L}(\cdot;\cdot)$	Optimization function
$\cdot^*$	Optimal solution
$\hat{\cdot}$	Estimate of a variable
$(\cdot)_i$	i-th cross-validation set
$\bar{\cdot}$	Mean
$\mathbb{E}[\cdot]$	Expectation of a random variable
$\propto$	Proportional to
$\approx$	Approximately equal to
$\equiv$	Equivalent to

Table 2. Parameters of the simulations

Category	Parameter	Notation
Complexity of model	Number of delays	D
	Number of features	F
Dimensionality of data	Number of variates	V
Strength of signal	Signal-to-noise ratio (SNR)	S
AR(1) temporal autocorrelation (if $T \geq 2$)	of true predictors	ϕ_X
	of null predictors	ϕ_U
	of true weight	ϕ_B
	of noise	ϕ_E
AR(1) spatial autocorrelation (if $V \geq 2$)	of true weight	θ_B
	of noise	θ_E
Multicollinearity (if $F \geq 2$)	of true predictors	ρ_X
	of null predictors	ρ_U
Presence of information leakage	Binary flag for the stimulus repetition	`IsRep`

Number of time points $T = 100$, λ-grid = $10^{[-10,-9,\ldots,15]}$, Number of samplings $K = 1000$

Table 3. Strong effects ($\eta_p^2 \geq 0.160$) on the prediction accuracies in univariate models with univariate responses using true predictors $\mathbf{X}$

Contrast	SS	F-stat	η_p^2
S	8.377×10^{1}	1.867×10^{5}	0.985
$\phi_X : \phi_E$	2.645×10^{-1}	5.895×10^{2}	0.175
$S : \phi_E$	2.037×10^{-1}	4.539×10^{2}	0.140
$S : \phi_X : \phi_E$	3.187×10^{-1}	7.101×10^{2}	0.203

$df_1 = 1$, $df_2 = 2788$, $P_{\text{Bonferroni}} \leq 0.0001$, $\eta_p^2 \geq 0.14$, SS: sum of squares

Table 3—source data 1. Full anova table: https://zenodo.org/records/15101528/files/uni-uni_anova-x.xlsx

Table 4. Strong effects ($\eta_p^2 \geq 0.160$) on the prediction accuracies in univariate models with univariate responses using null predictors $\mathbf{U}$

Contrast	SS	F-stat	η_p^2
`IsRep`	1.569×10^{1}	3.864×10^{4}	0.933
S	2.199	5.413×10^{3}	0.660
D	1.504	3.704×10^{3}	0.571
ϕ_X	4.312×10^{-1}	1.062×10^{3}	0.276
ϕ_U	3.607×10^{-1}	8.881×10^{2}	0.242
S : `IsRep`	2.209	5.438×10^{3}	0.661
`IsRep` : D	1.458	3.589×10^{3}	0.563
$\phi_X : \phi_U$	6.987×10^{-1}	1.720×10^{3}	0.382
ϕ_X : `IsRep`	4.435×10^{-1}	1.092×10^{3}	0.281
ϕ_U : `IsRep`	3.357×10^{-1}	8.266×10^{2}	0.229
$S : D$	2.164×10^{-1}	5.328×10^{2}	0.160
$\phi_X : \phi_U$: `IsRep`	7.277×10^{-1}	1.792×10^{3}	0.391
S : `IsRep` : D	2.408×10^{-1}	5.929×10^{2}	0.175

$df_1 = 1$, $df_2 = 2788$, $P_{\text{Bonferroni}} \leq 0.0001$, $\eta_p^2 \geq 0.14$, SS: sum of squares

Table 4—source data 1. Full anova table: https://zenodo.org/records/15101528/files/uni-uni_anova-u.xlsx

Table 5. Strong effects ($\eta_p^2 \geq 0.160$) on the prediction accuracies of multivariate models with univariate responses using true predictors $\mathbf{X}$

Contrast	SS	F-stat	η_p^2
S	5.729×10^{2}	5.984×10^{5}	0.972
ρ_X	2.374×10^{1}	2.479×10^{4}	0.593
D	1.258×10^{1}	1.314×10^{4}	0.436
`IsRep`	6.991	7.302×10^{3}	0.301
F	6.196	6.471×10^{3}	0.276
ϕ_X	4.010	4.188×10^{3}	0.198
ϕ_E	3.557	3.715×10^{3}	0.179
$\phi_X : \phi_E$	5.048	5.273×10^{3}	0.237
$\rho_X : D$	3.213	3.356×10^{3}	0.165
ρ_X : `IsRep`	2.841	2.967×10^{3}	0.149

$df_1 = 1$, $df_2 = 16\,984$, $P_{\text{Bonferroni}} \leq 0.0001$, $\eta_p^2 \geq 0.14$, SS: sum of squares

Table 5—source data 1. Full anova table: https://zenodo.org/records/15101528/files/mult-uni_anova-x.xlsx

Table 6. Strong effects ($\eta_p^2 \geq 0.160$) on the prediction accuracies of multivariate models with univariate responses using null predictors $\mathbf{U}$

Contrast	SS	F-stat	η_p^2
`IsRep`	4.161×10^2	1.057×10^5	0.862
S	6.946×10^1	1.765×10^4	0.510
ρ_U	5.282×10^1	1.342×10^4	0.441
D	3.429×10^1	8.715×10^3	0.339
S : `IsRep`	6.933×10^1	1.762×10^4	0.509
ρ_U : `IsRep`	5.296×10^1	1.346×10^4	0.442
`IsRep` : D	3.469×10^1	8.816×10^3	0.342

$df_1 = 1$, $df_2 = 16\,984$, $P_{\text{Bonferroni}} \leq 0.0001$, $\eta_p^2 \geq 0.14$, SS: sum of squares

Table 6—source data 1. Full anova table: https://zenodo.org/records/15101528/files/mult-uni_anova-u.xlsx

Table 7. Strong effects ($\eta_p^2 \geq 0.160$) on the prediction accuracies of multivariate models with multivariate responses using true predictors $\mathbf{X}$

Contrast	SS	F-stat	η_p^2
S	1.161×10^4	1.417×10^7	0.976
ρ_X	3.606×10^2	4.404×10^5	0.557
D	2.177×10^2	2.659×10^5	0.432
`IsRep`	8.921×10^1	1.090×10^5	0.237
F	8.851×10^1	1.081×10^5	0.236
ϕ_X	7.909×10^1	9.660×10^4	0.216
ϕ_E	7.119×10^1	8.696×10^4	0.199
$\phi_X : \phi_E$	1.057×10^2	1.292×10^5	0.269
$\rho_X : D$	5.227×10^1	6.384×10^4	0.154
$S : \phi_X : \phi_E$	4.999×10^1	6.106×10^4	0.148

$df_1 = 1$, $df_2 = 350\,191$, $P_{\text{Bonferroni}} \leq 0.0001$, $\eta_p^2 \geq 0.14$, SS: sum of squares

Table 7—source data 1. Full anova table: https://zenodo.org/records/15101528/files/mult-mult_anova-x.xlsx

Table 8. Strong effects ($\eta_p^2 \geq 0.160$) on the prediction accuracies of multivariate models on multivariate responses with null predictors $\mathbf{U}$

Contrast	SS	F-stat	η_p^2
`IsRep`	7.636×10^3	2.446×10^6	0.875
S	1.269×10^3	4.066×10^5	0.537
ρ_U	8.979×10^2	2.876×10^5	0.451
D	7.222×10^2	2.313×10^5	0.398
S : `IsRep`	1.270×10^3	4.068×10^5	0.538
ρ_U : `IsRep`	8.967×10^2	2.872×10^5	0.451
`IsRep` : D	7.226×10^2	2.315×10^5	0.398

$df_1 = 1$, $df_2 = 349\,964$, $P_{\text{Bonferroni}} \leq 0.0001$, $\eta_p^2 \geq 0.14$, SS: sum of squares

Table 8—source data 1. Full anova table: https://zenodo.org/records/15101528/files/mult-mult_anova-u.xlsx

Appendix 1

Optimization processes with identical signals

In a scenario where the strong underlying signals are identical in the training and optimization sets ($\mathbf{s}_1 = \mathbf{s}_3$ and $\|\mathbf{s}_1\| \gg 0$), the optimization prediction accuracy for a given $\lambda \geq 0$ is:

$$\begin{aligned}\mathbb{E}\left[\text{corr}\left(\hat{\mathbf{y}}_{3,\mathbf{X}}[\lambda],\, \mathbf{y}_3\right)\right] = \mathbb{E}\left[\frac{\left(\hat{\mathbf{y}}_{3,\mathbf{X}}[\lambda]\right)^{\mathsf{T}}\mathbf{y}_3}{\|\hat{\mathbf{y}}_{3,\mathbf{X}}\|\|\mathbf{y}_3\|}\right] &\propto \mathbb{E}\left[\left(\hat{\mathbf{y}}_{3,\mathbf{X}}[\lambda]\right)^{\mathsf{T}}\mathbf{y}_3\right] = \mathbb{E}\left[\mathbf{y}_1^{\mathsf{T}}\mathbf{P}_{\mathbf{X}}^{\mathsf{T}}[\lambda]\mathbf{y}_3\right] \\ &= \mathbb{E}\left[(\mathbf{s}_1 + \mathbf{e}_1)^{\mathsf{T}}\mathbf{P}_{\mathbf{X}}^{\mathsf{T}}[\lambda](\mathbf{s}_1 + \mathbf{e}_3)\right] \\ &= \mathbf{s}_1^{\mathsf{T}}\left\{\mathbf{X}_1\left(\mathbf{X}_1^{\mathsf{T}}\mathbf{X}_1 + \lambda\mathbf{I}\right)^{-1}\mathbf{X}_1^{\mathsf{T}}\right\}^{\mathsf{T}}\mathbf{s}_1\end{aligned} \tag{A1.1}$$

If $\mathbf{X} \in \mathbb{R}^{T\times P}$ where $P = FD$ is orthonormal and full-rank, ***Equation A1.1*** can be simplified as a quadratic form of a scaled Ordinary Least Squares projection matrix and the signal as (***Hastie et al., 2009***, pp. 64):

$$\mathbf{s}_1^{\mathsf{T}}\left\{\mathbf{X}_1\left(\mathbf{X}_1^{\mathsf{T}}\mathbf{X}_1 + \lambda\mathbf{I}\right)^{-1}\mathbf{X}_1^{\mathsf{T}}\right\}^{\mathsf{T}}\mathbf{s}_1 = \frac{1}{1+\lambda}\mathbf{s}_1^{\mathsf{T}}\mathbf{X}_1\left(\mathbf{X}_1^{\mathsf{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathsf{T}}\mathbf{s}_1 = \frac{\|\mathbf{s}_1\|}{1+\lambda}, \tag{A1.2}$$

for which the optimal λ to maximize the prediction accuracy is zero ($\because \|\mathbf{s}_1\| > 0$):

$$\lambda^* = \arg\max_{\lambda}\frac{\|\mathbf{s}_1\|}{1+\lambda} = 0. \tag{A1.3}$$

More generally, the singular value decomposition of $\mathbf{X}$ can be used; $\mathbf{X} = \mathbf{UDV}^{\mathsf{T}}$ where $\mathbf{U} \in \mathbb{R}^{T\times P}$ and $\mathbf{V} \in \mathbb{R}^{P\times P}$ are orthonormal, and the diagonal matrix $\mathbf{D} \in \mathbb{R}^{P\times P}$ contains the singular values: $d_1 \geq d_2 \geq \cdots \geq d_P \geq 0$. Then, ***Equation A1.1*** can be written as (***Hastie et al., 2009***, Eq. 3.47):

$$\mathbf{s}_1^{\mathsf{T}}\left\{\mathbf{X}_1\left(\mathbf{X}_1^{\mathsf{T}}\mathbf{X}_1 + \lambda\mathbf{I}\right)^{-1}\mathbf{X}_1^{\mathsf{T}}\right\}^{\mathsf{T}}\mathbf{s}_1 = \mathbf{s}_1^{\mathsf{T}}\left(\sum_{j=1}^{P}\mathbf{u}_{(j)}\frac{d_j^2}{d_j^2+\lambda}\mathbf{u}_{(j)}^{\mathsf{T}}\right)\mathbf{s}_1. \tag{A1.4}$$

For any $d_j = 0$ (i.e., $\mathbf{X}$ is rank-deficit), λ needs to be positive for the prediction accuracy value to be defined. Nonetheless, the prediction accuracy is maximized when $\lambda^* = \epsilon \approx 0$ for ϵ is the smallest positive value.

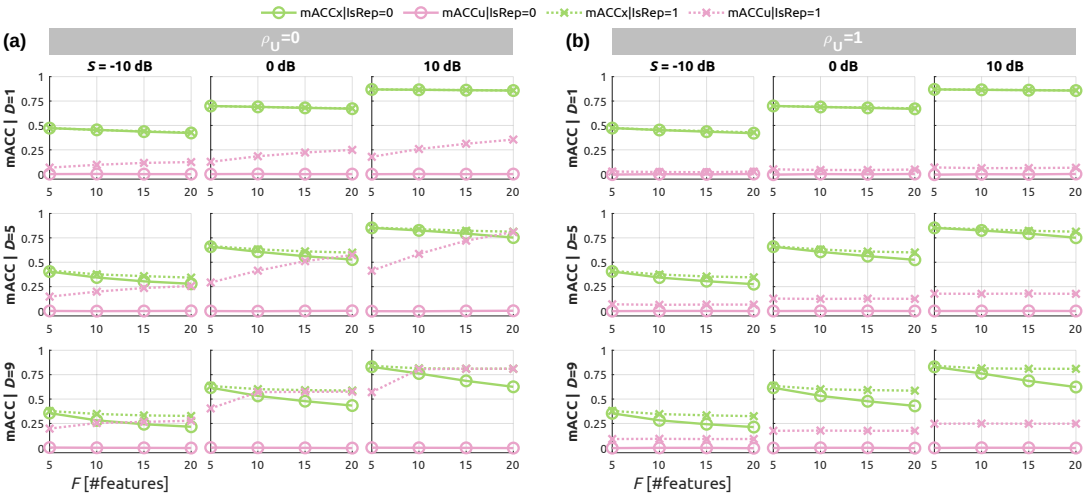


Figure 5—figure supplement 1. [TODO: FIXME] Simulations of multivariate-feature, univariate-response models when the multicolinearity of true feature is zero. Mean prediction accuracies are plotted over the number of features F when multicolinearity **(a)** $\rho_U = 0$ and **(b)** $\rho_U = 1$. Marker styles and colors are identical to those in ***Figure 4***. Each column represents a specified SNR level. The number of delays is $D = 1$ in the top rows and $D = 9$ in the bottom rows. Other parameters were as follows: $K = 1000$, $\phi_U = 0$, $\phi_B = 0$, $\phi_E = 0$, $\rho_X = 0$.

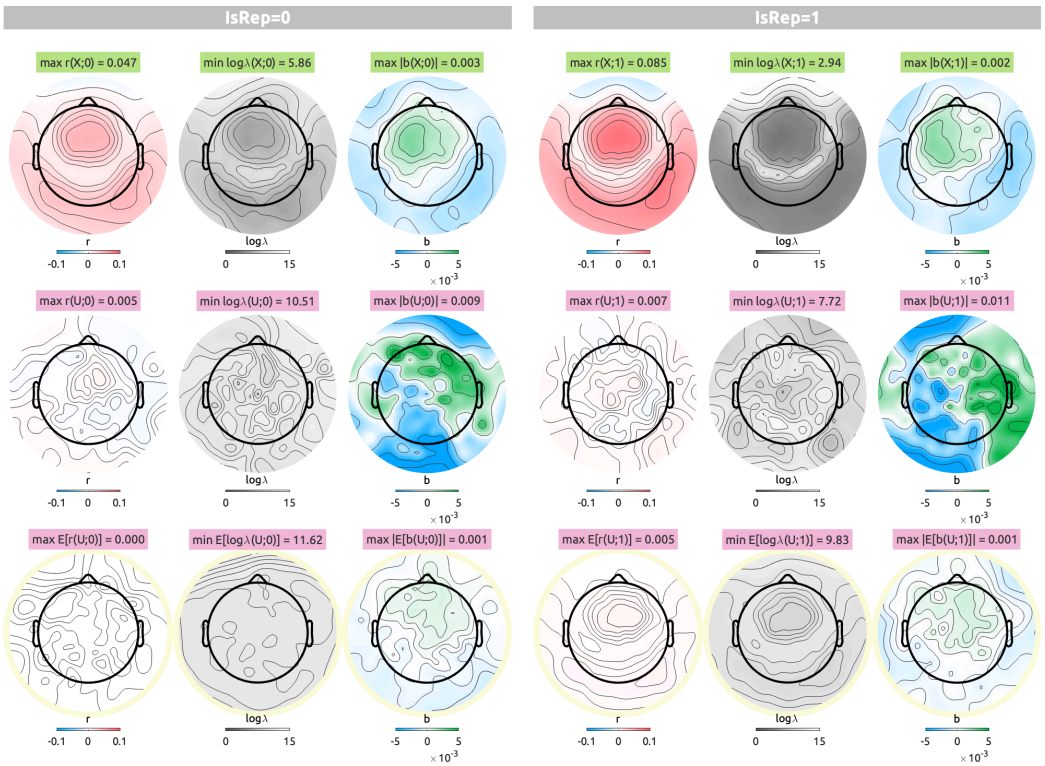


Figure 7—figure supplement 1. EEG linearized encoding analysis results with delays from 0 to 0.5 sec with an audio envelope (top row), a single case of the normal noise (middle row), and an average of 100 normal noises (bottom row; circled in pale yellow). For each CV scheme ($\texttt{IsRep} = 0$, left panels; $\texttt{IsRep} = 1$, right panels), prediction accuracy (r; blue to red), logarithmic ridge hyperparameter ($\log_{10} \lambda$; gray to white), summed weights (b; blue to green) are shown along the columns.

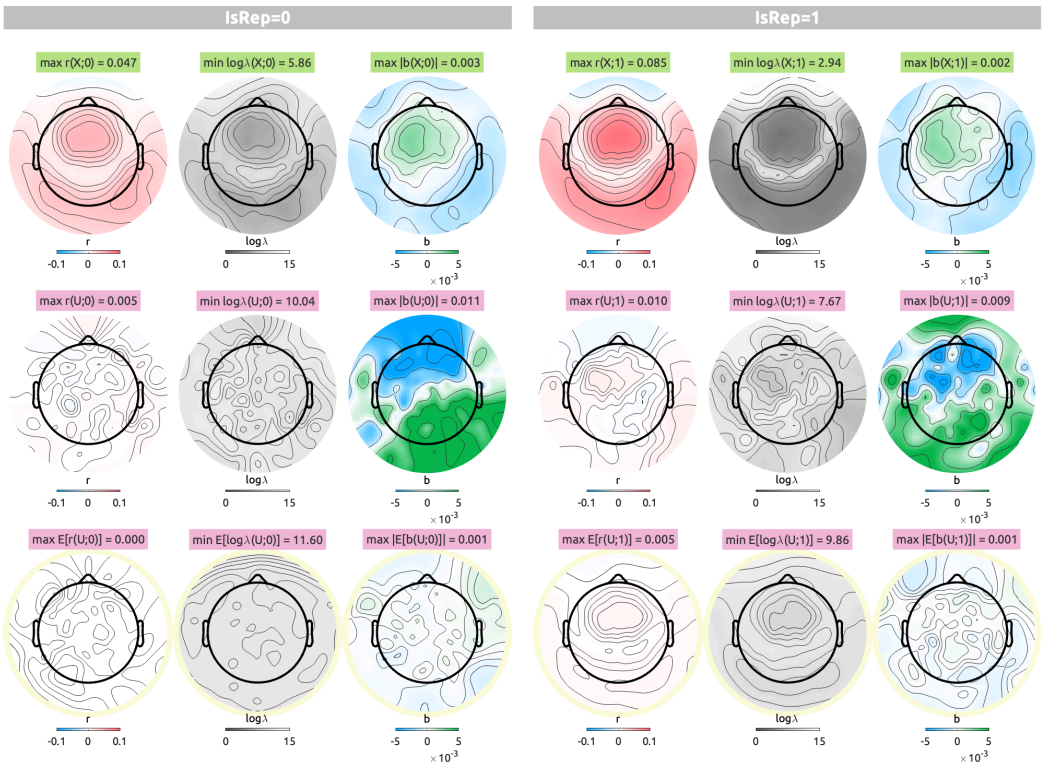


Figure 7—figure supplement 2. EEG topographies of linearized encoding analysis results with delays from 0 to 0.5 sec with the uniform noise as the null feature. The visualization scheme is identical to ***Figure 7—figure Supplement 1***.

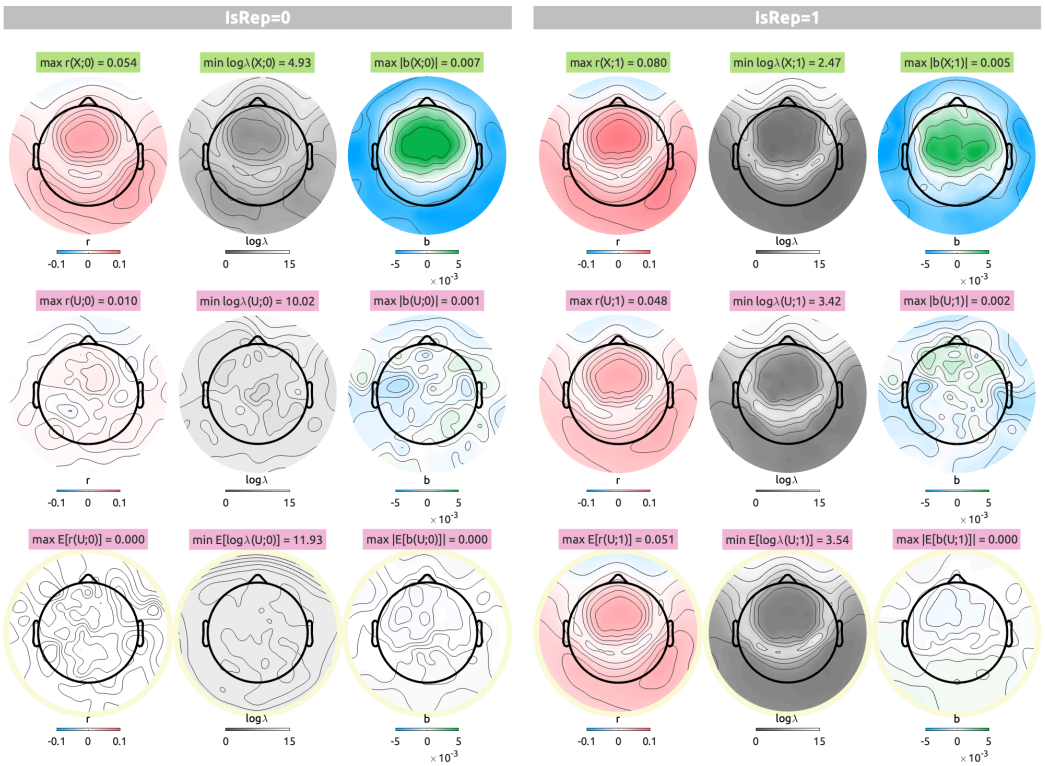


Figure 7—figure supplement 3. EEG topographies of linearized encoding analysis results with a delay from 0 to 0.3 sec with the phase-randomized envelope as the null feature. The visualization scheme is identical to ***Figure 7—figure Supplement 1***.

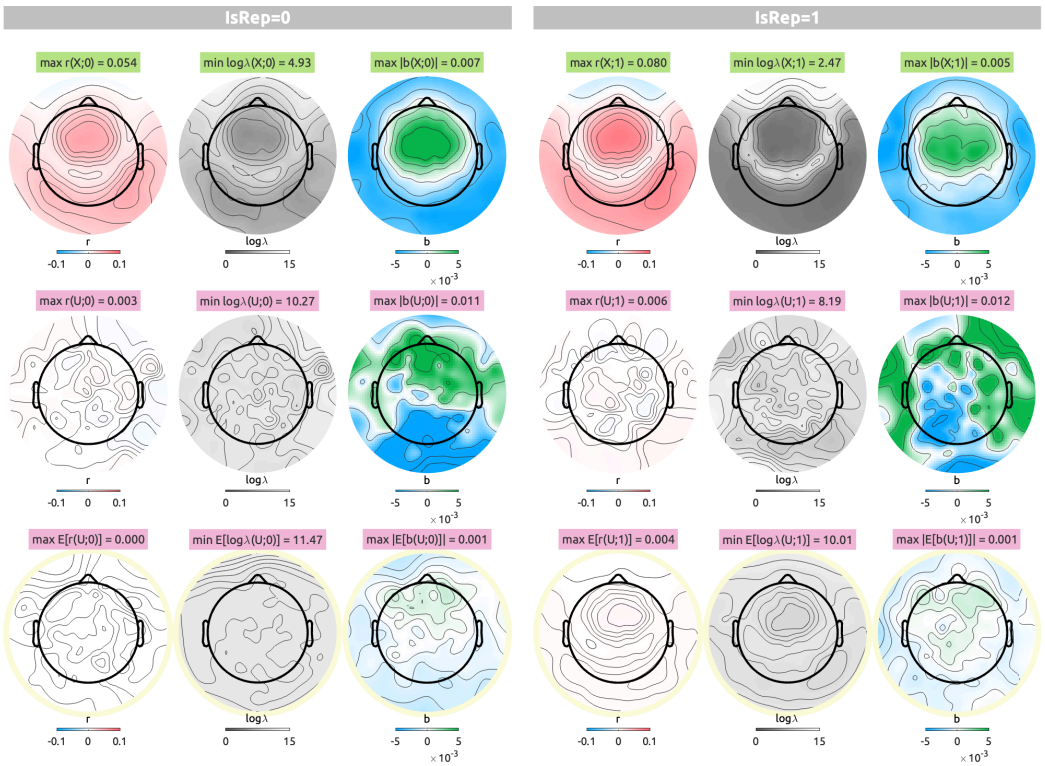


Figure 7—figure supplement 4. EEG topographies of linearized encoding analysis results with a delay from 0 sec to 0.3 sec with the normal noise as the null feature. The visualization scheme is identical to ***Figure 7—figure Supplement 1***.

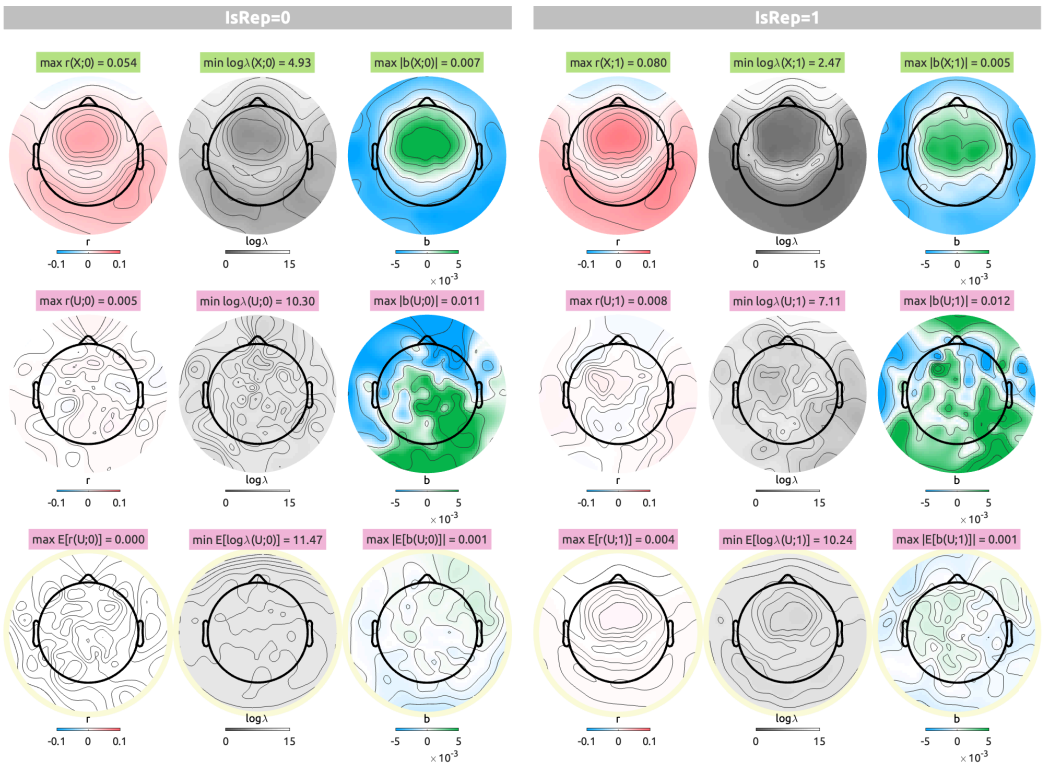


Figure 7—figure supplement 5. EEG topographies of linearized encoding analysis results with a delay from 0 to 0.3 sec with the uniform envelope as the null feature. The visualization scheme is identical to ***Figure 7—figure Supplement 1***.

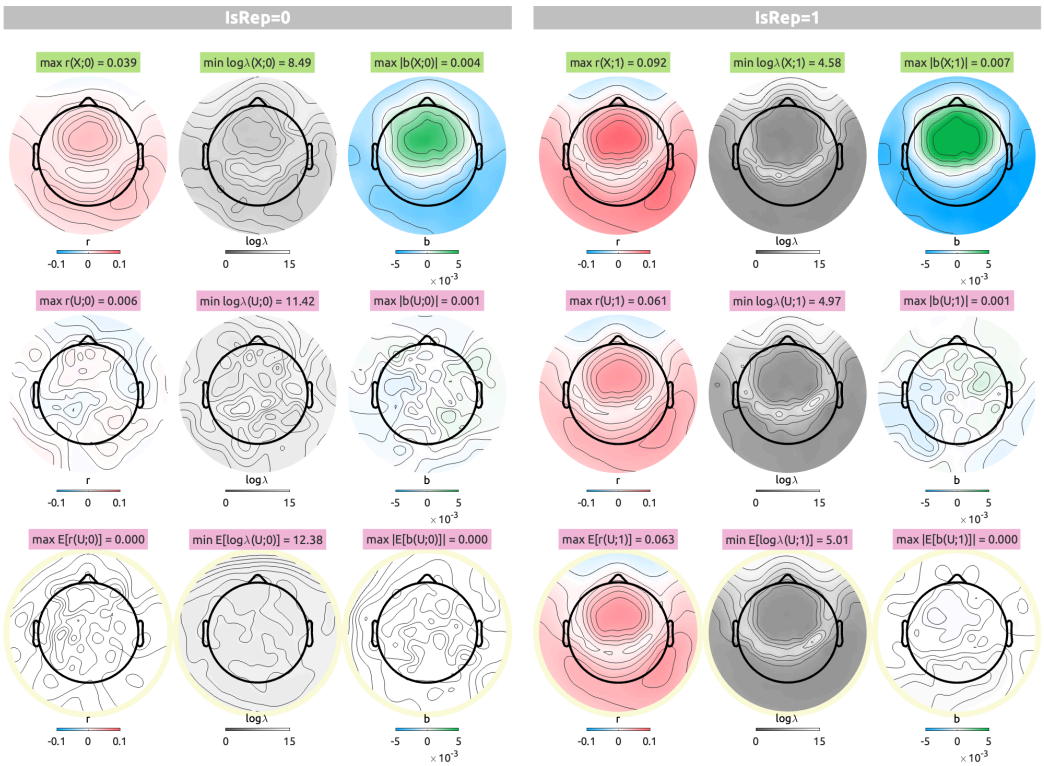


Figure 7—figure supplement 6. EEG topographies of linearized encoding analysis results with delays from 0 to 1 sec with the phase-randomized envelope as the null feature. The visualization scheme is identical to ***Figure 7—figure Supplement 1***.

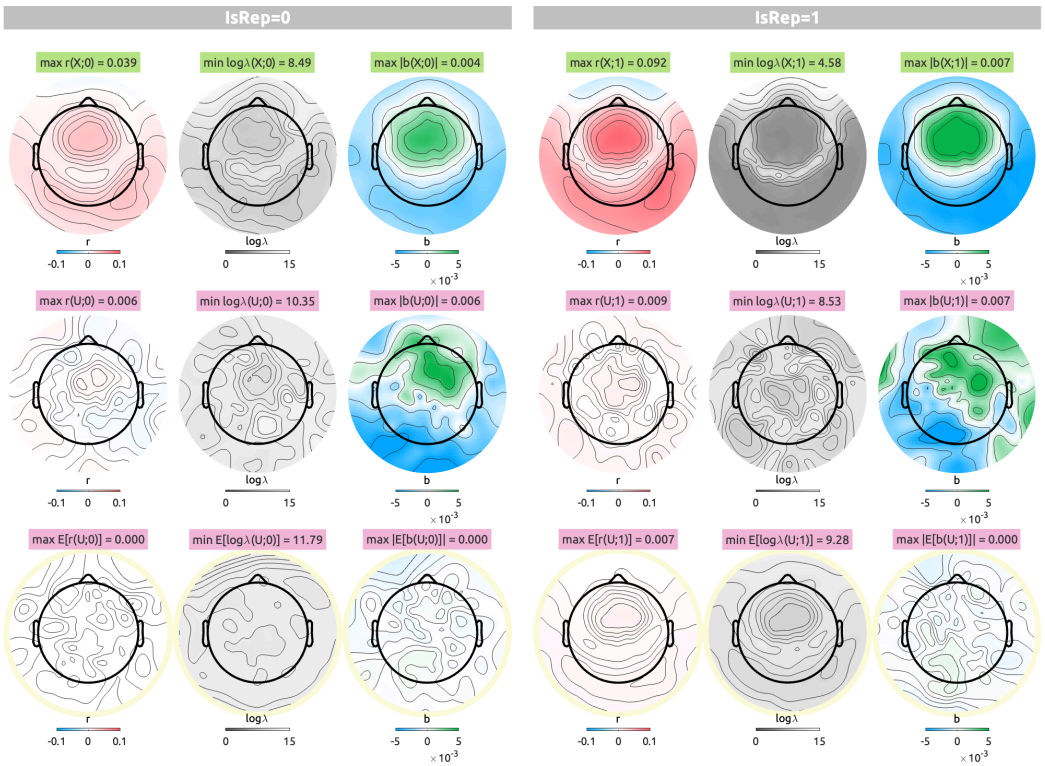


Figure 7—figure supplement 7. EEG topographies of linearized encoding analysis results with delays from 0 to 1 sec with the normal noise as the null feature. The visualization scheme is identical to ***Figure 7—figure Supplement 1***.

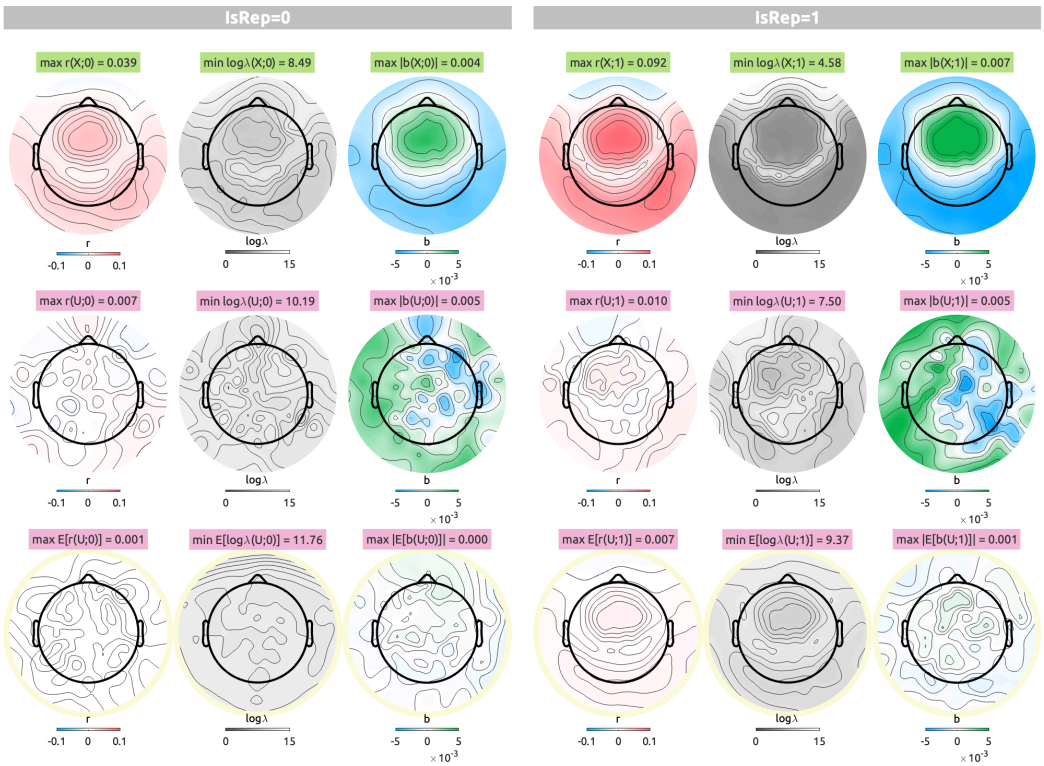


Figure 7—figure supplement 8. EEG topographies of linearized encoding analysis results with delays from 0 to 1 sec with the uniform noise as the null feature. The visualization scheme is identical to ***Figure 7—figure Supplement 1***.

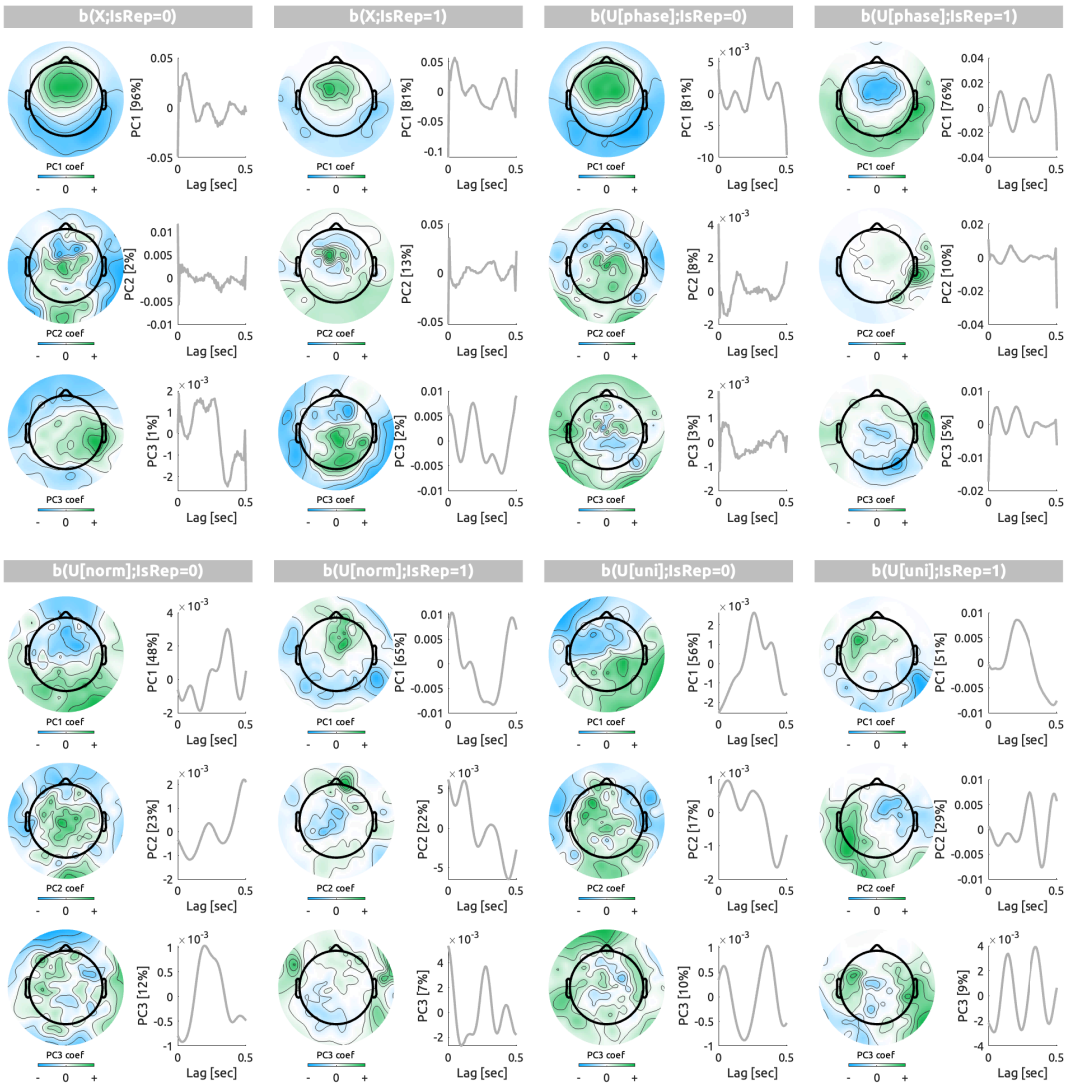


Figure 7—figure supplement 9. First three principal components (PCs) of the transfer function weights (b) of the models with delays from 0 to 0.5 sec are displayed by the topography of eigenvectors and the time series of eigenvariates with the explained variance noted. The weights are grouped by features (X, true audio envelope; $U[phase]$, phase-randomized envelope; $U[norm]$, normal noise; $U[uni]$, uniform noise) and the CV schemes ($\texttt{IsRep} = 0$, $\texttt{IsRep} = 1$).

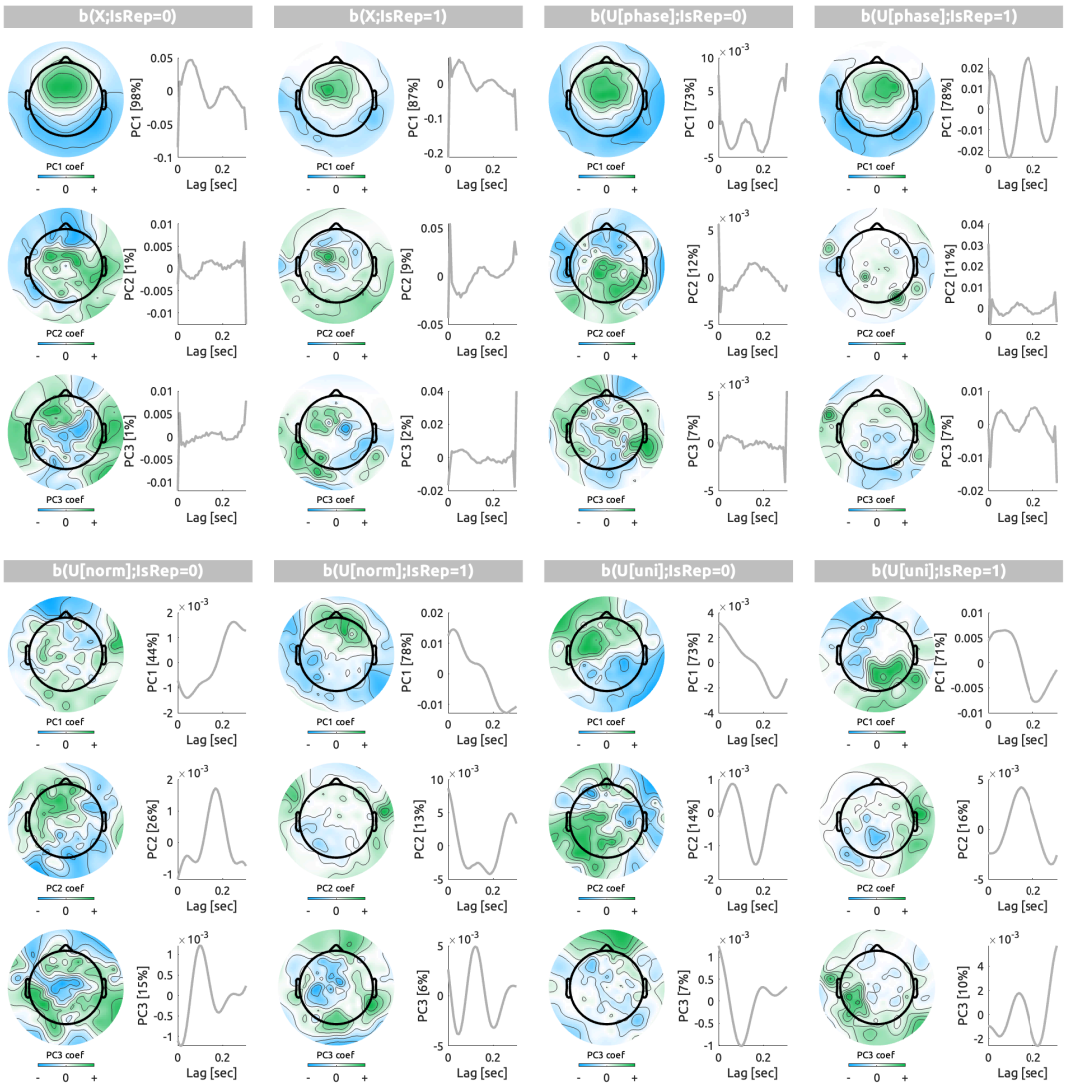


Figure 7—figure supplement 10. First three principal components (PCs) of the transfer function weights (b) of the models with delays from 0 to 0.3 sec are displayed. The visualization scheme is identical to ***Figure 7—figure Supplement 9***.

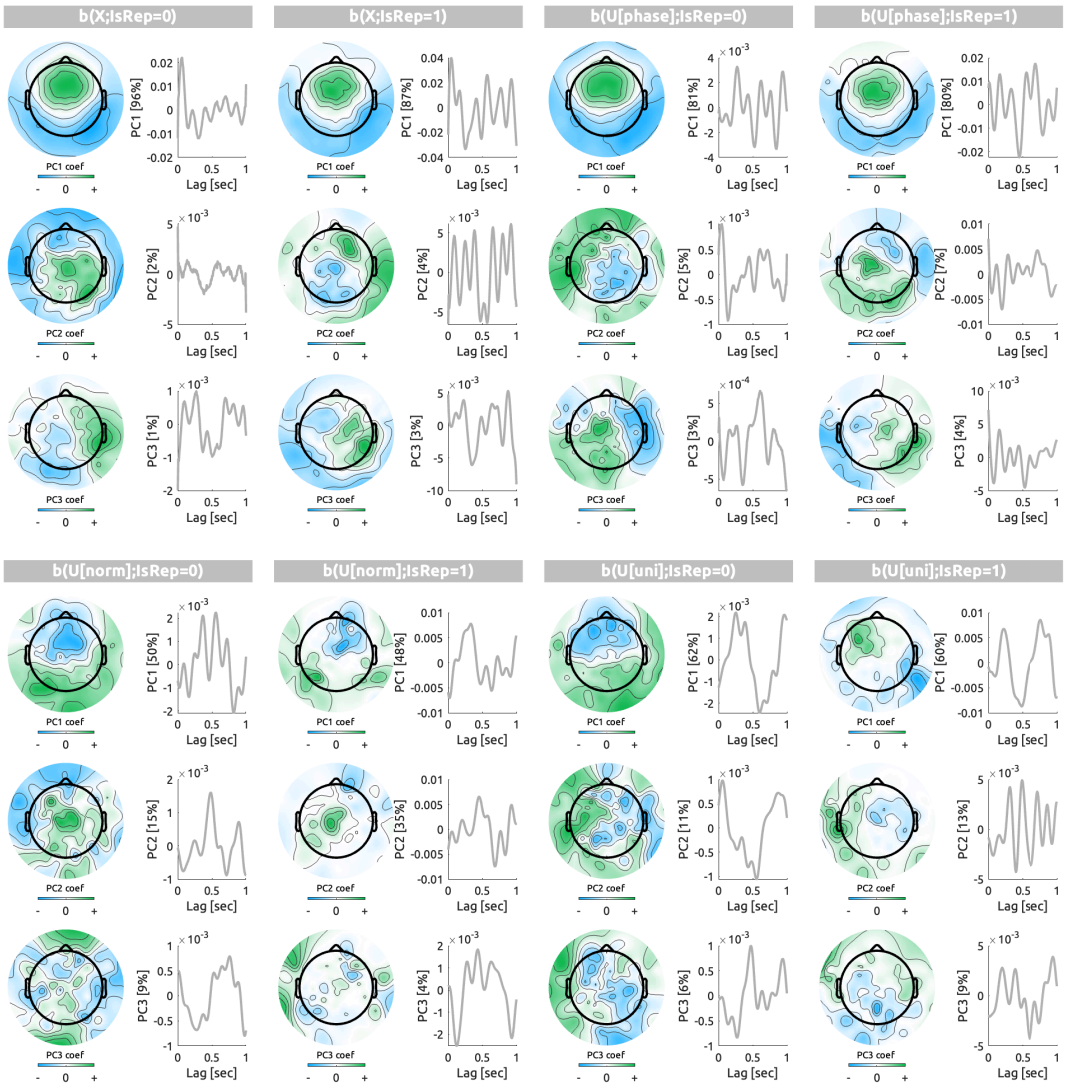


Figure 7—figure supplement 11. First three principal components (PCs) of the transfer function weights (*b*) of the models with delays from 0 to 1 sec are displayed. The visualization scheme is identical to ***Figure 7—figure Supplement 9***.

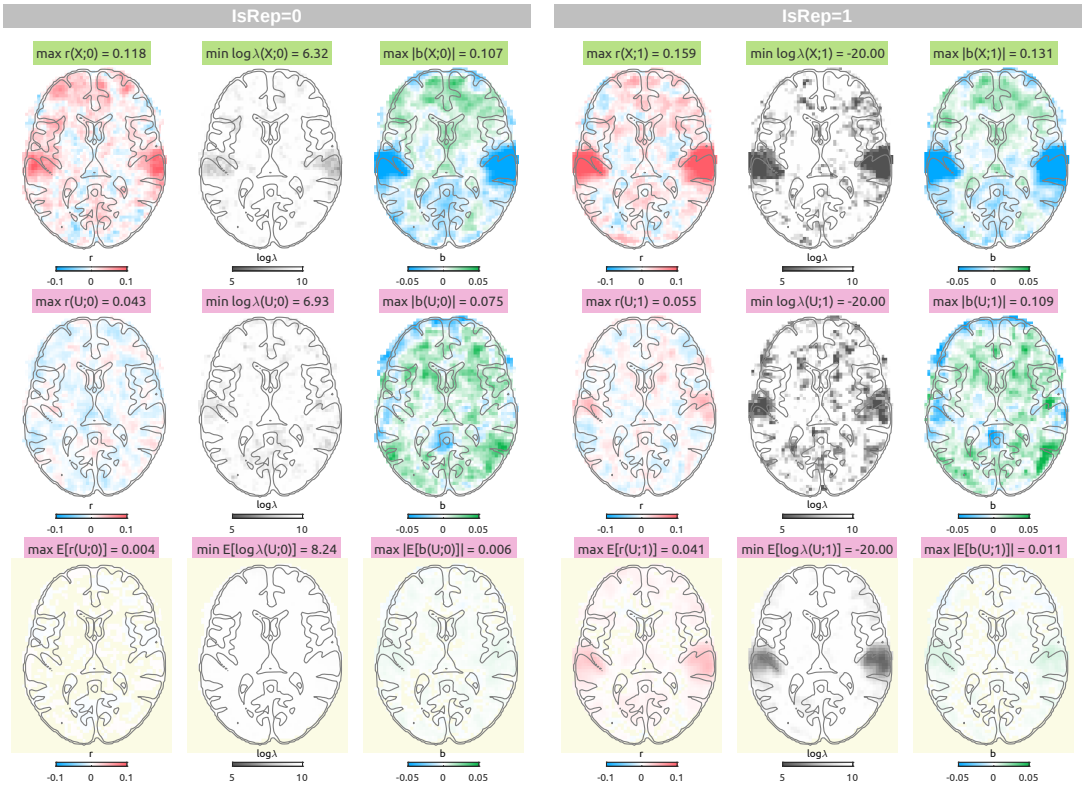


Figure 9—figure supplement 1. fMRI linearized encoding analysis results with delays from 3 to 9 sec with an audio envelope (top row), a single case of the normal noise (middle row), and an average of 100 phase-randomized envelopes (bottom row, pale yellow background). For each CV scheme ($\texttt{IsRep} = 0$, left panels; $\texttt{IsRep} = 1$, right panels), prediction accuracy (r, blue to red), logarithmic ridge hyperparameter ($\log_{10} \lambda$, gray to white), transfer function weights that are summed over delays (b, blue to green) are shown along the columns. The 3-D volumes can be viewed with the NeuroVault web viewer (https://identifiers.org/neurovault.collection:19626).

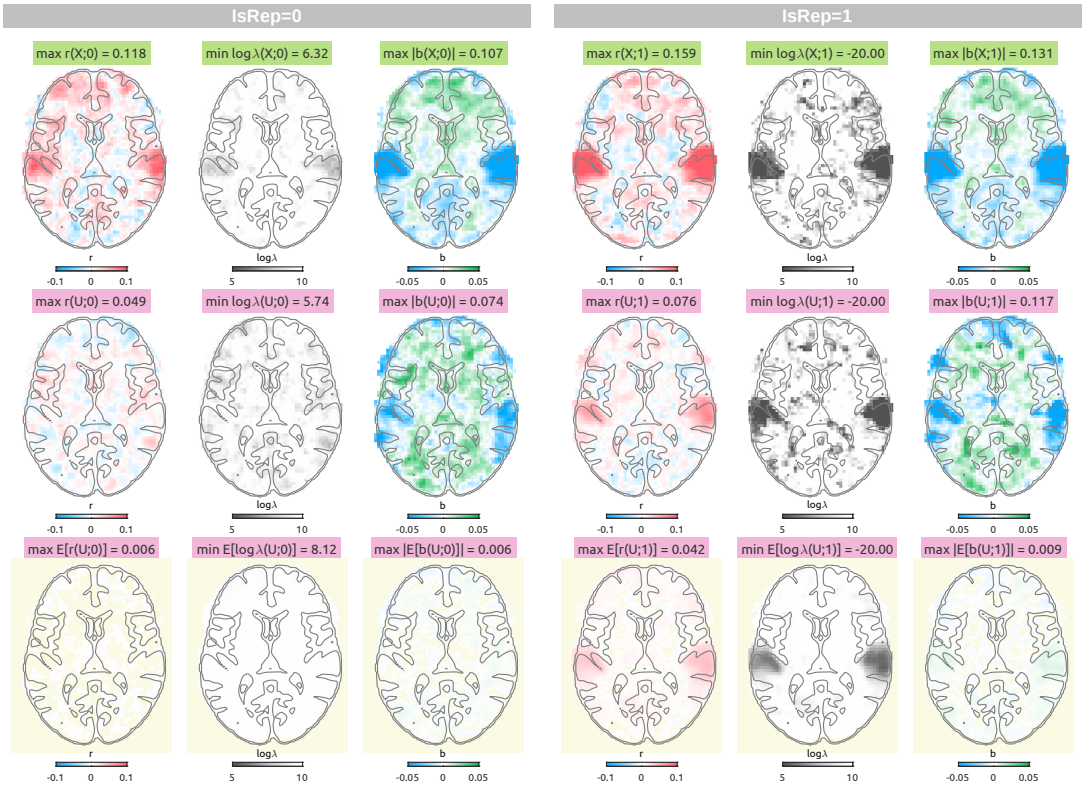


Figure 9—figure supplement 2. fMRI linearized encoding analysis results with delays from 3 to 9 sec with the uniform noise as the null feature. The visualization scheme is identical to ***Figure 9—figure Supplement 1***.

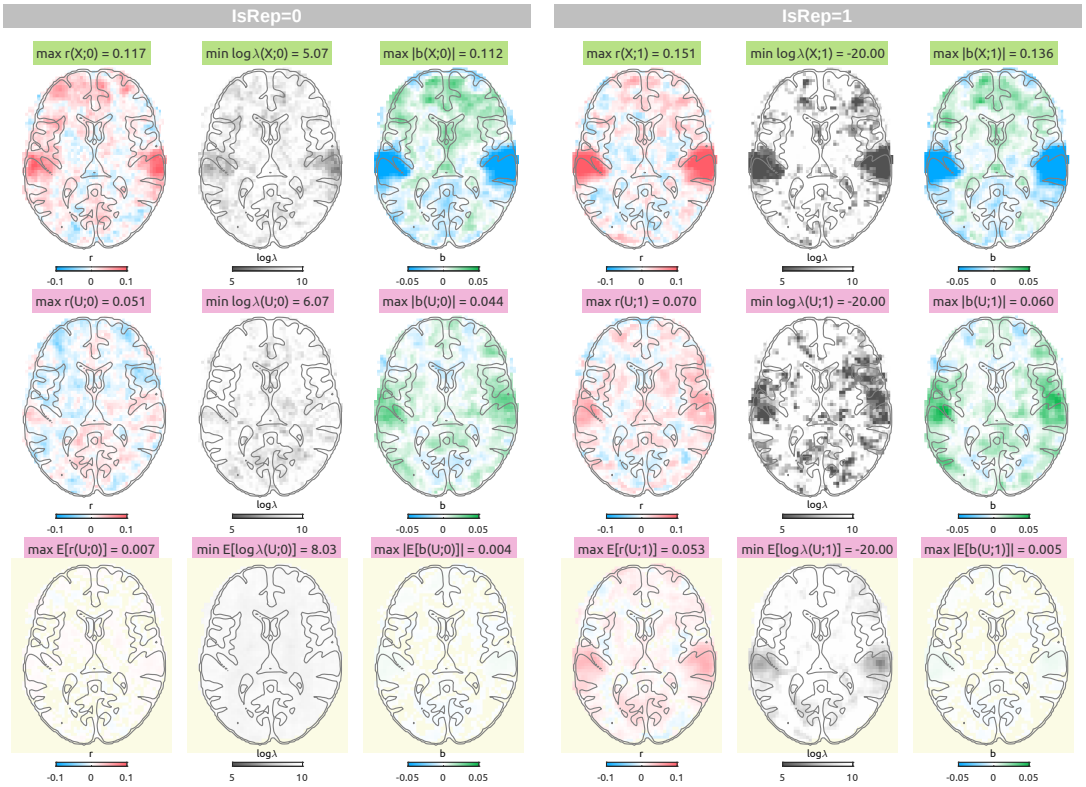


Figure 9—figure supplement 3. fMRI linearized encoding analysis results with delays from 4 to 6 sec with the phase-randomized envelope as the null feature. The visualization scheme is identical to ***Figure 9—figure Supplement 1***.

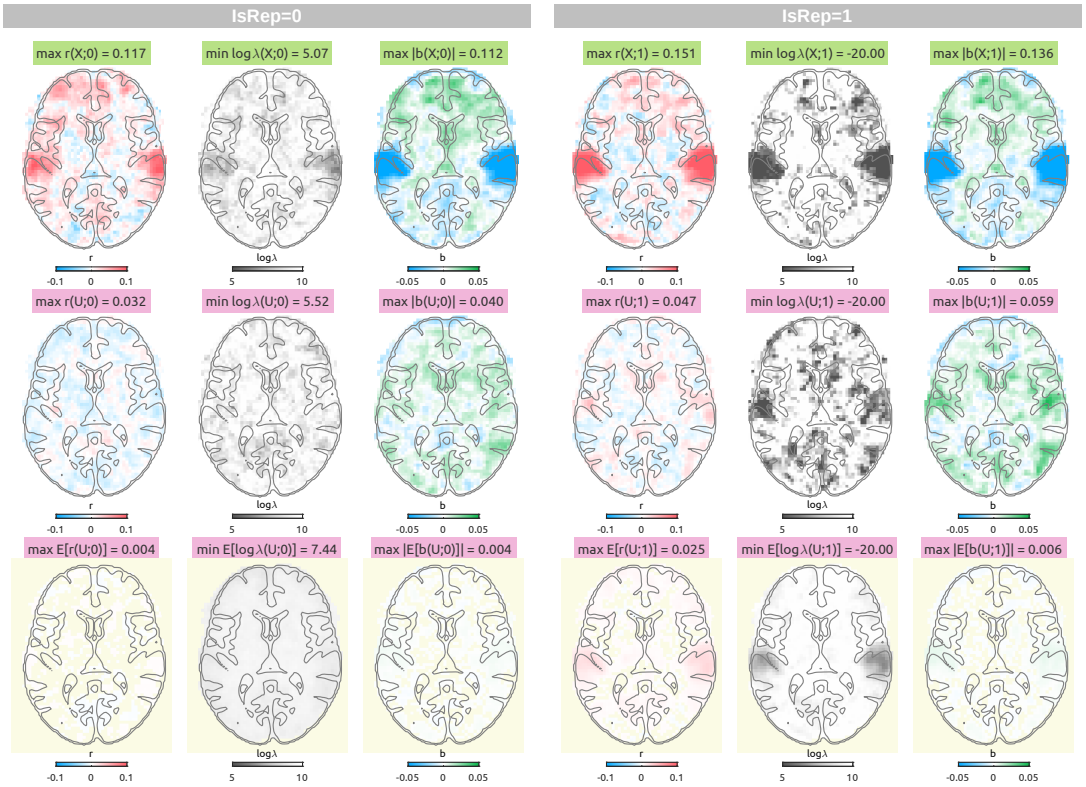
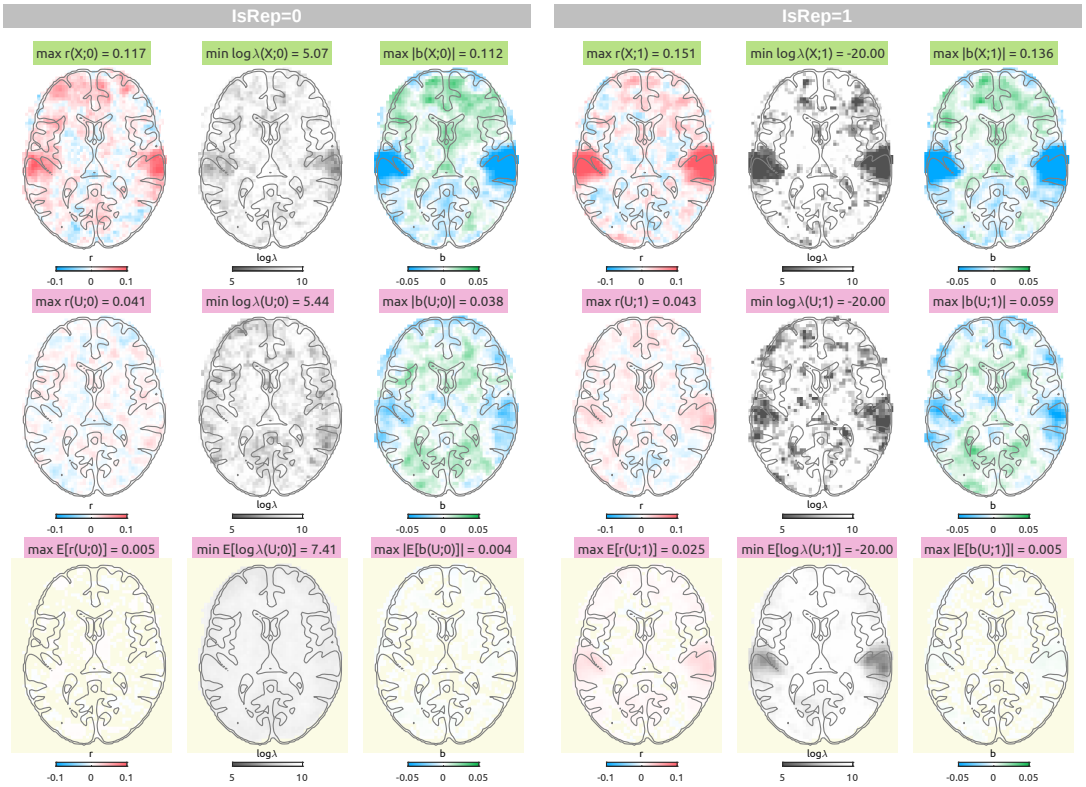


Figure 9—figure supplement 4. fMRI linearized encoding analysis results with delays from 4 to 6 sec with the normal noise as the null feature. The visualization scheme is identical to ***Figure 9—figure Supplement 1***.



1110

Figure 9—figure supplement 5. fMRI linearized encoding analysis results with delays from 4 to 6 sec with the uniform noise as the null feature. The visualization scheme is identical to ***Figure 9—figure Supplement 1***.

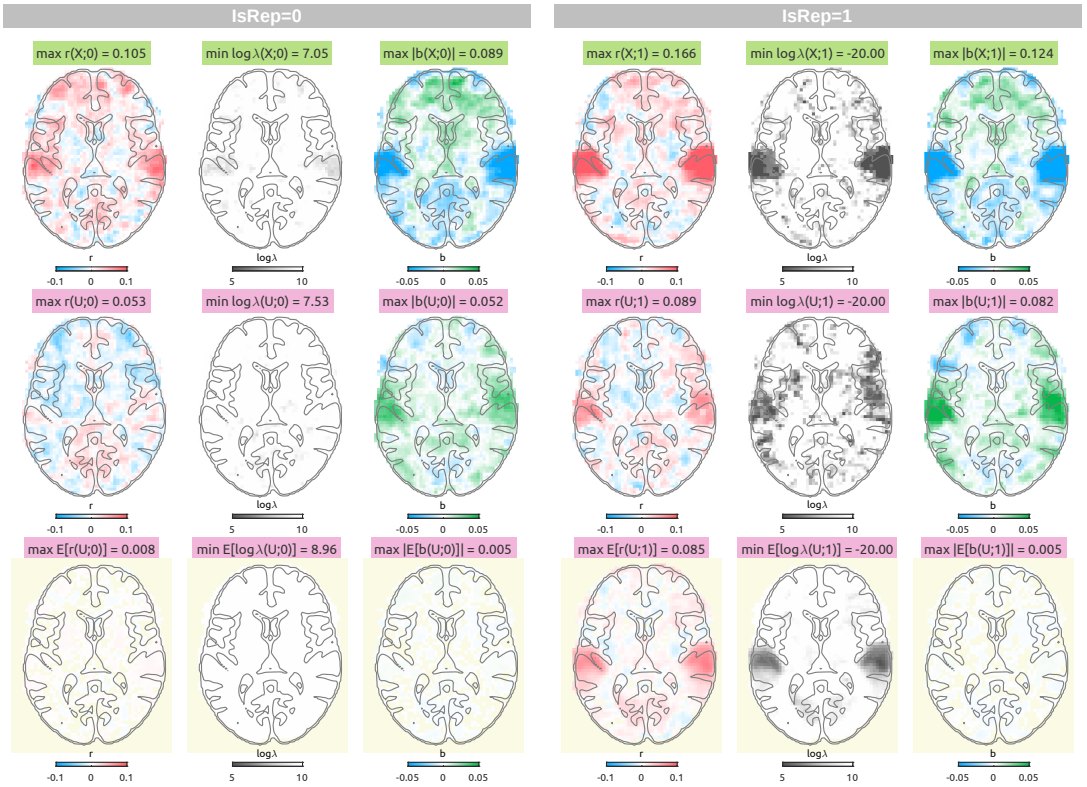


Figure 9—figure supplement 6. fMRI linearized encoding analysis results with delays from 0 to 12 sec with the phase-randomized envelope as the null feature. The visualization scheme is identical to ***Figure 9—figure Supplement 1***.

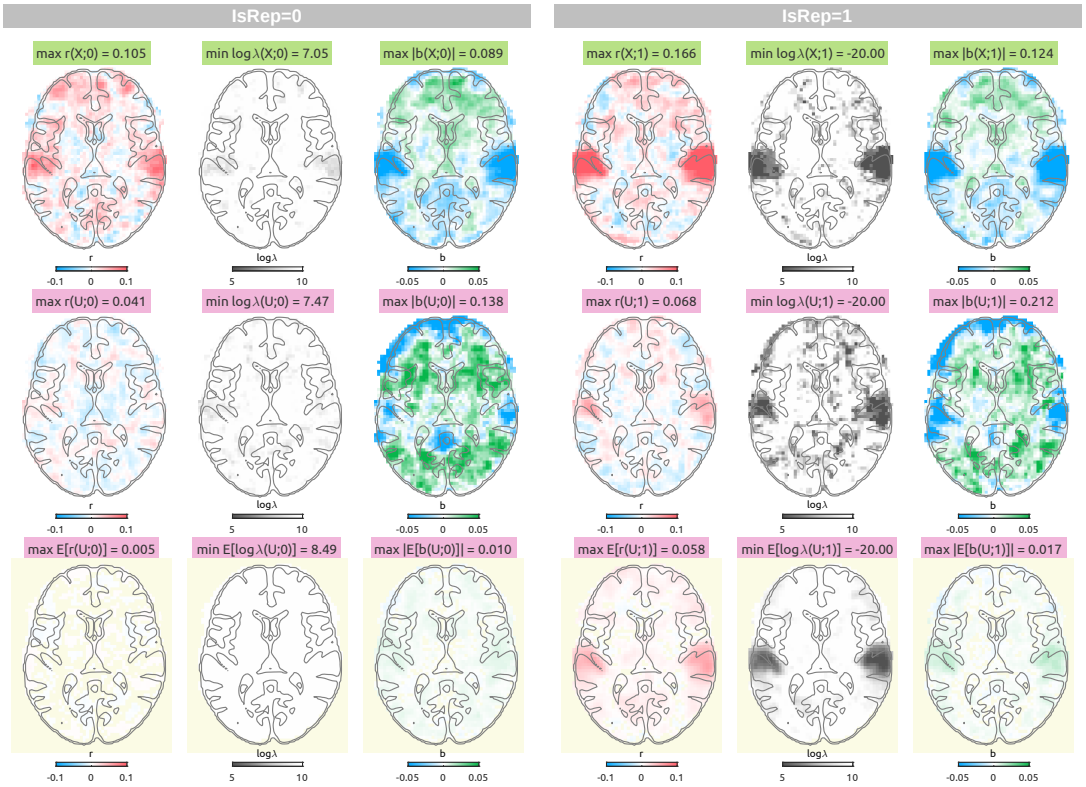


Figure 9—figure supplement 7. fMRI linearized encoding analysis results with delays from 4 to 6 sec with the normal noise as the null feature. The visualization scheme is identical to ***Figure 9—figure Supplement 1***.

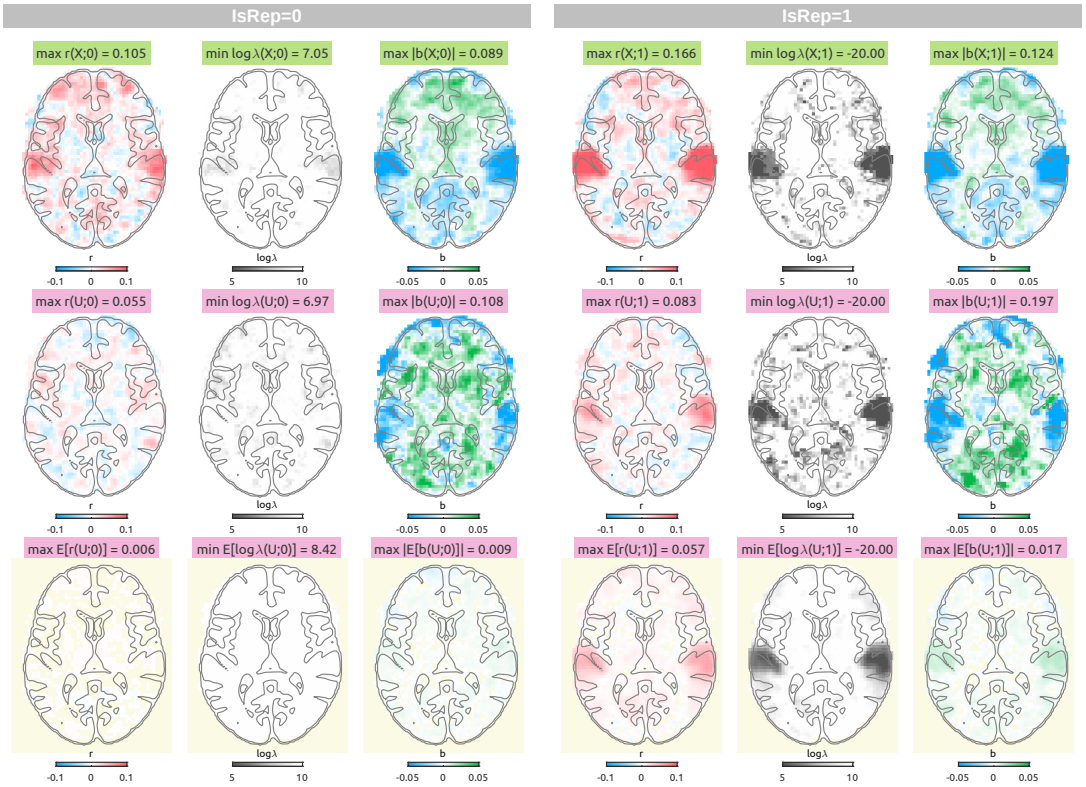


Figure 9—figure supplement 8. fMRI linearized encoding analysis results with delays from 0 to 12 sec with the uniform noise as the null feature. The visualization scheme is identical to ***Figure 9—figure Supplement 1***.

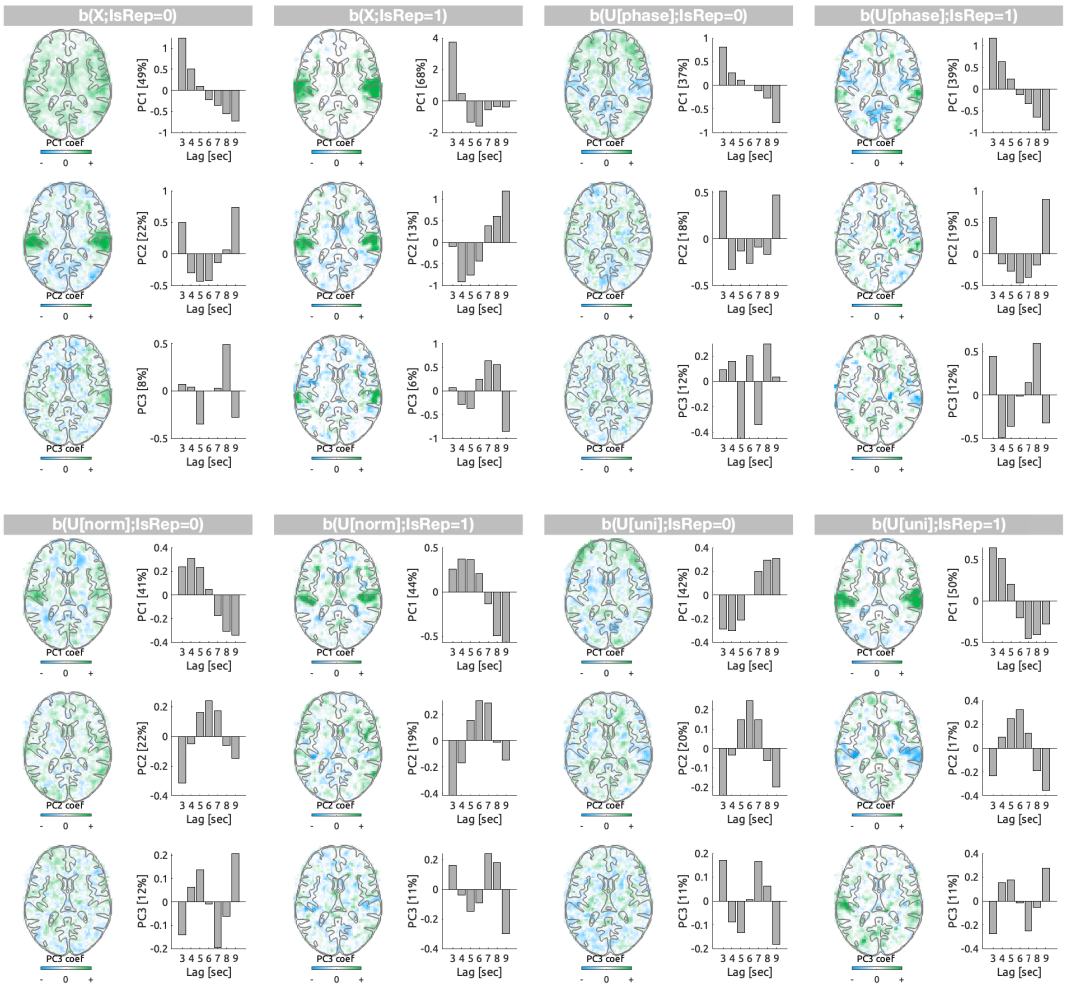


Figure 9—figure supplement 9. First three principal components (PCs) of the transfer function weights (b) of the models with delays from 3 to 9 sec are displayed in the transverse slice of eigenvectors and the time series of eigenvariates with the explained variance noted. The weights are grouped by features (X, true audio envelope; $U[phase]$, phase-randomized envelope; $U[norm]$, normal noise; $U[uni]$, uniform noise) and the CV schemes ($\texttt{IsRep} = 0$, $\texttt{IsRep} = 1$).

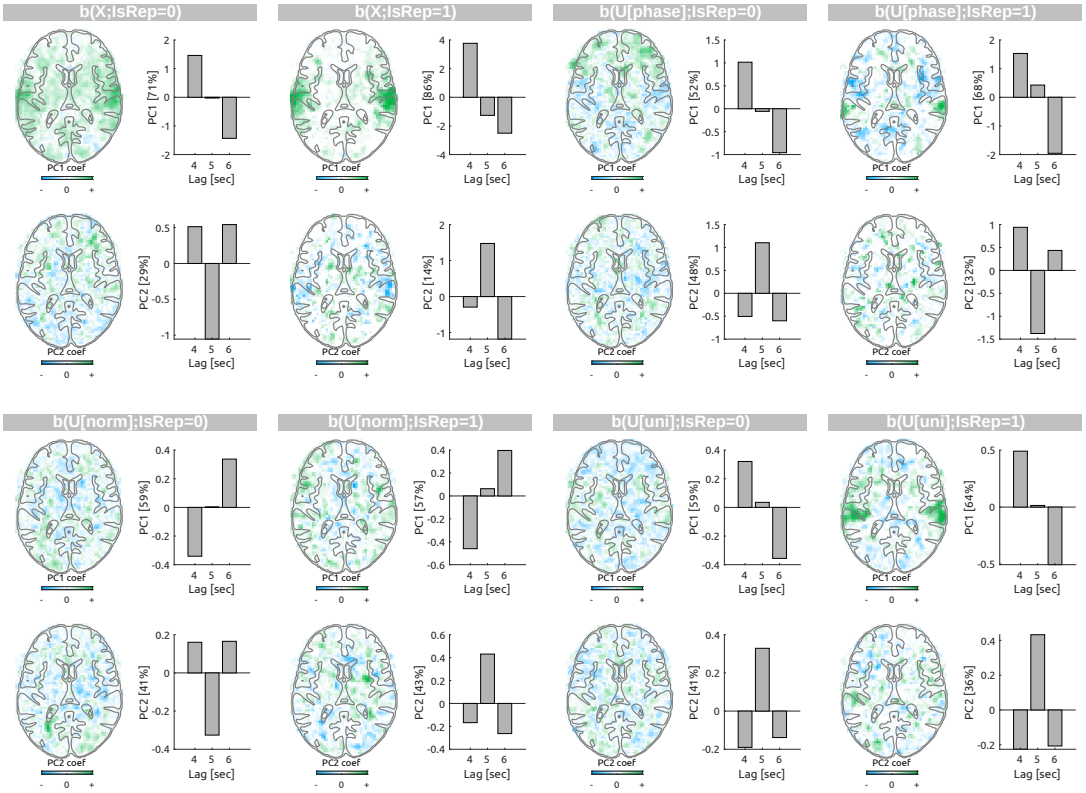


Figure 9—figure supplement 10. First two principal components (PCs) of the transfer function weights (*b*) of the models with delays from 4 to 6 sec are displayed. The visualization scheme is identical to ***Figure 9—figure Supplement 9***.

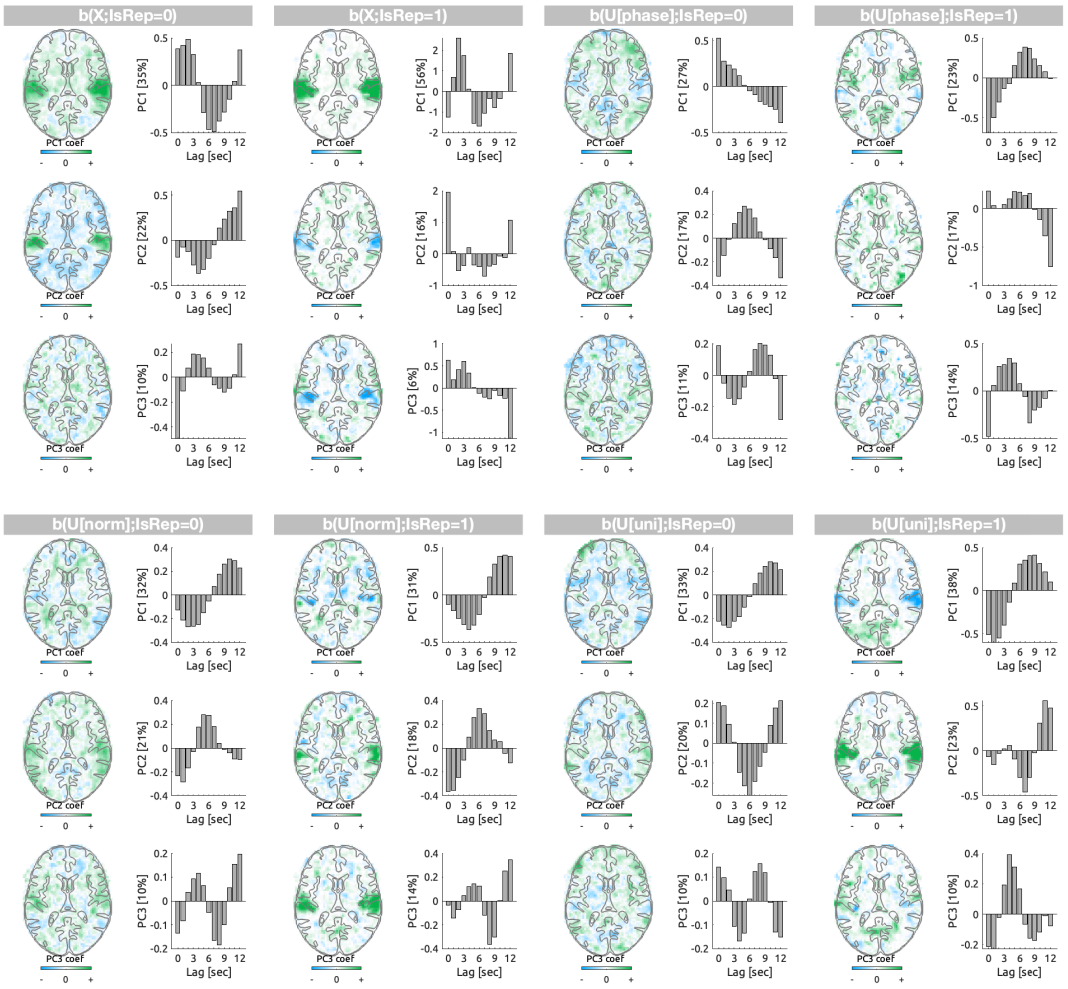


Figure 9—figure supplement 11. First three principal components (PCs) of the transfer function weights (*b*) of the models with delays from 0 to 12 sec are displayed. The visualization scheme is identical to ***Figure 9—figure Supplement 9***.

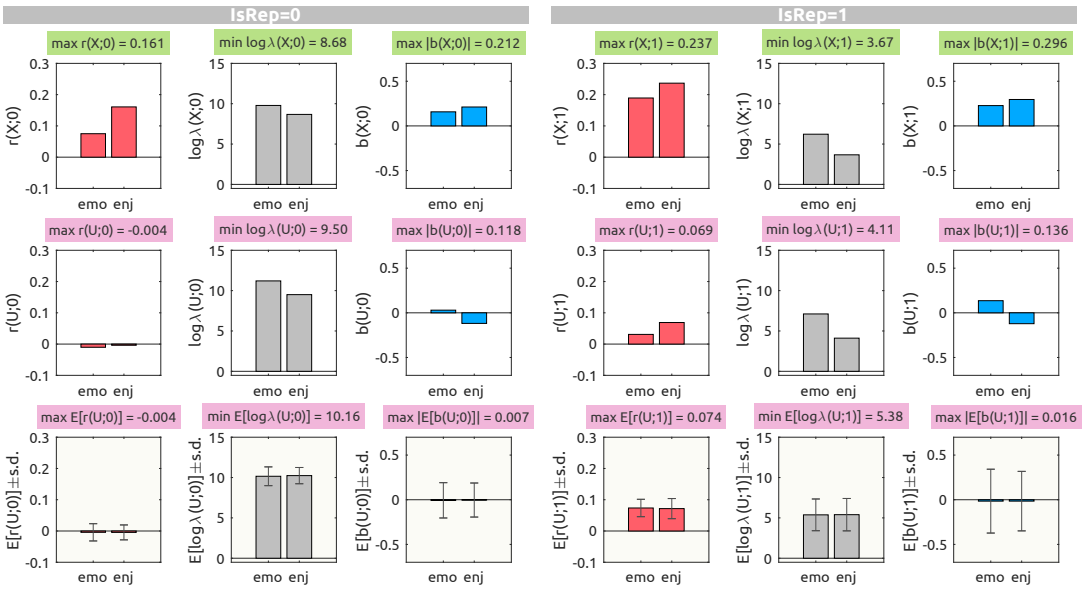


Figure 11—figure supplement 1. Behavioral linearized encoding analysis results with delays from 0 to 10 sec with an audio envelope (top row), a single case of a phase-randomized envelope (middle row), and an average of 100 phase-randomized envelopes (bottom row, pale yellow background). For each CV scheme ($\texttt{IsRep} = 0$, left panels; $\texttt{IsRep} = 1$, right panels), prediction accuracy (r, red bars), logarithmic ridge hyperparameter ($\log_{10} \lambda$, gray bars), transfer function weights that are summed over delays (b, blue bars) are shown along the columns.

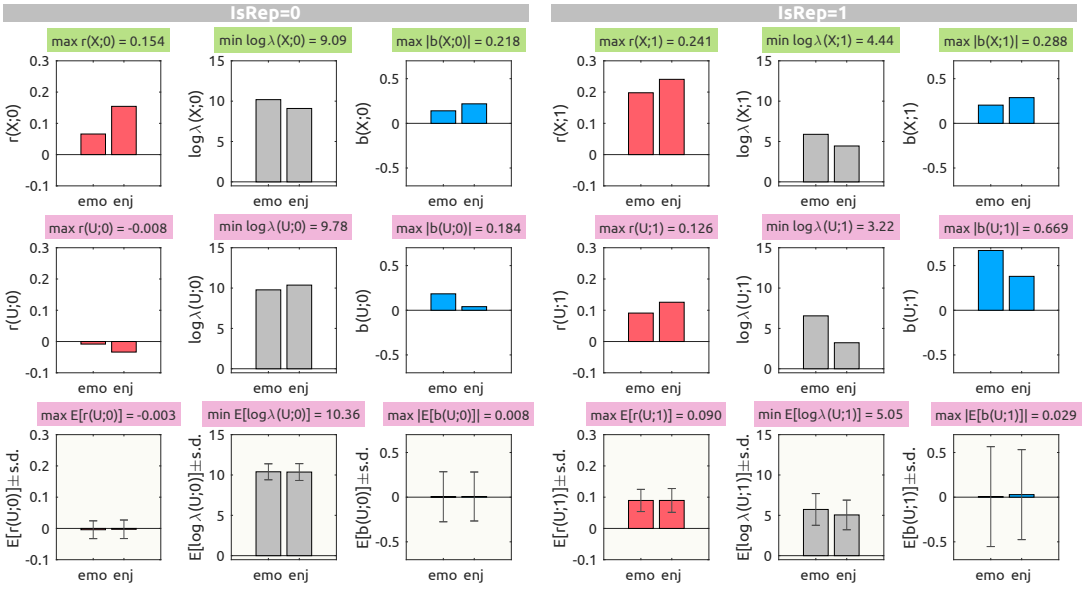
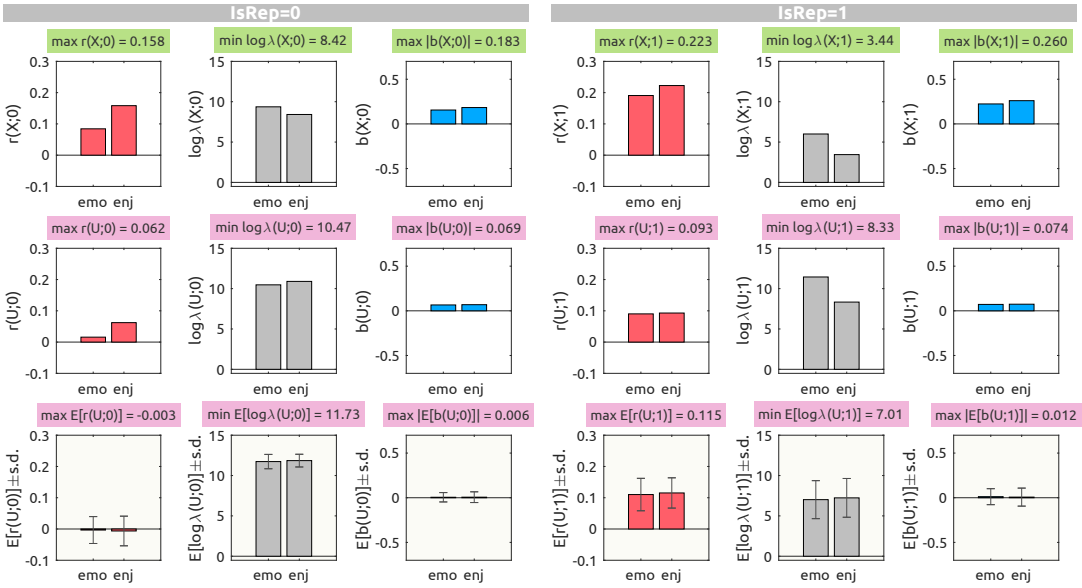
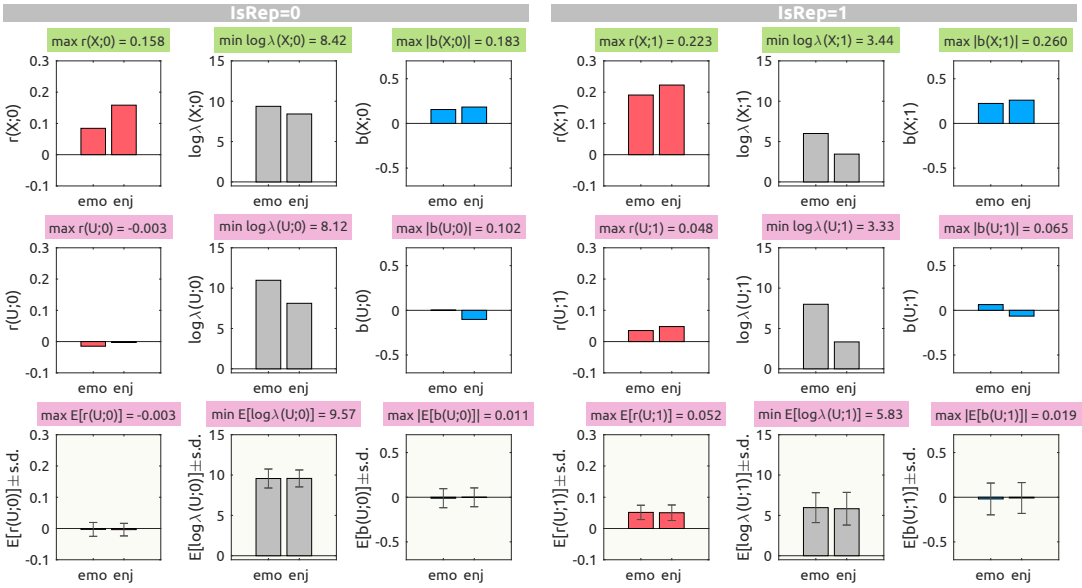


Figure 11—figure supplement 2. Behavioral linearized encoding analysis results with delays from 0 to 10 sec with the uniform noise as the null feature. The visualization scheme is identical to ***Figure 11—figure Supplement 1***.



1119

Figure 11—figure supplement 3. Behavioral linearized encoding analysis results with delays from 0 to 5 sec with the phase-randomized envelope as the null feature. The visualization scheme is identical to ***Figure 11—figure Supplement 1***.



1120

Figure 11—figure supplement 4. Behavioral linearized encoding analysis results with delays from 0 to 5 sec with the phase-randomized envelope as the null feature. The visualization scheme is identical to ***Figure 11—figure Supplement 1***.

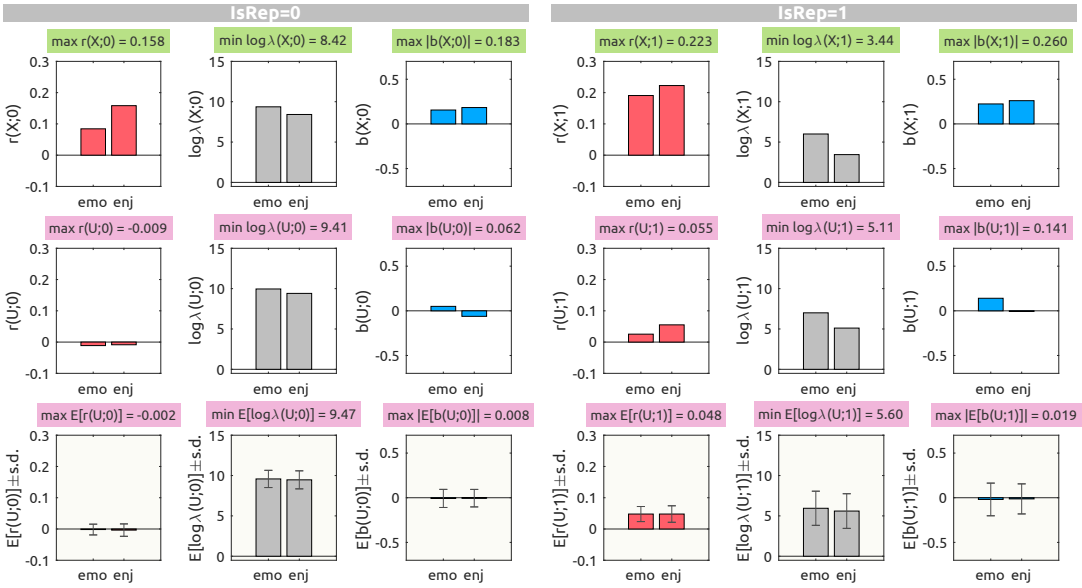


Figure 11—figure supplement 5. Behavioral linearized encoding analysis results with delays from 0 to 5 sec with the uniform noise as the null feature. The visualization scheme is identical to ***Figure 11—figure Supplement 1***.

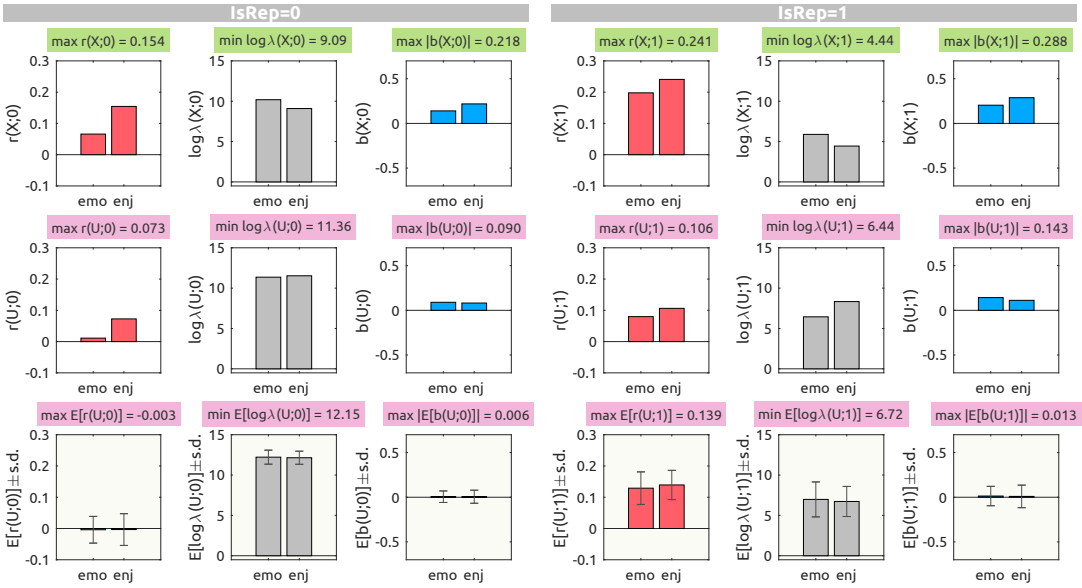


Figure 11—figure supplement 6. Behavioral linearized encoding analysis results with delays from 0 to 15 sec with the phase-randomized envelope as the null feature. The visualization scheme is identical to ***Figure 11—figure Supplement 1***.

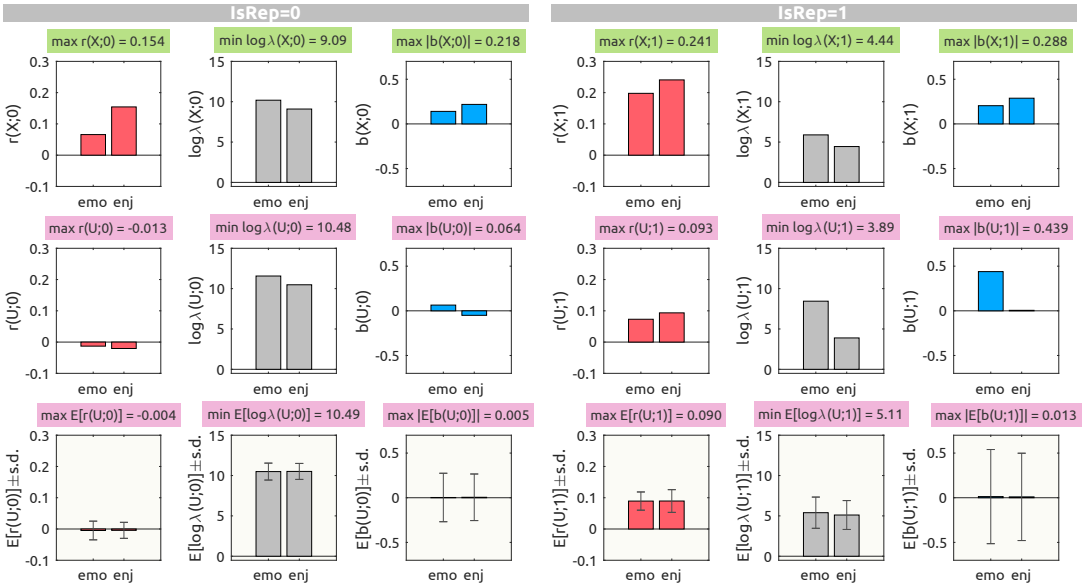


Figure 11—figure supplement 7. Behavioral linearized encoding analysis results with delays from 0 to 15 sec with the phase-randomized envelope as the null feature. The visualization scheme is identical to ***Figure 11—figure Supplement 1***.

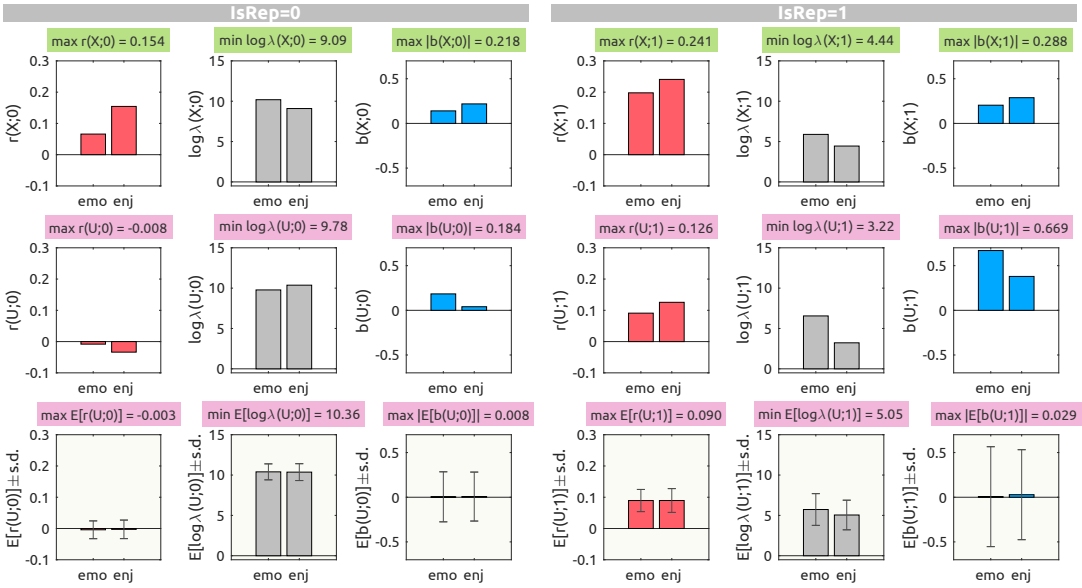


Figure 11—figure supplement 8. Behavioral linearized encoding analysis results with delays from 0 to 15 sec with the uniform noise as the null feature. The visualization scheme is identical to ***Figure 11—figure Supplement 1***.

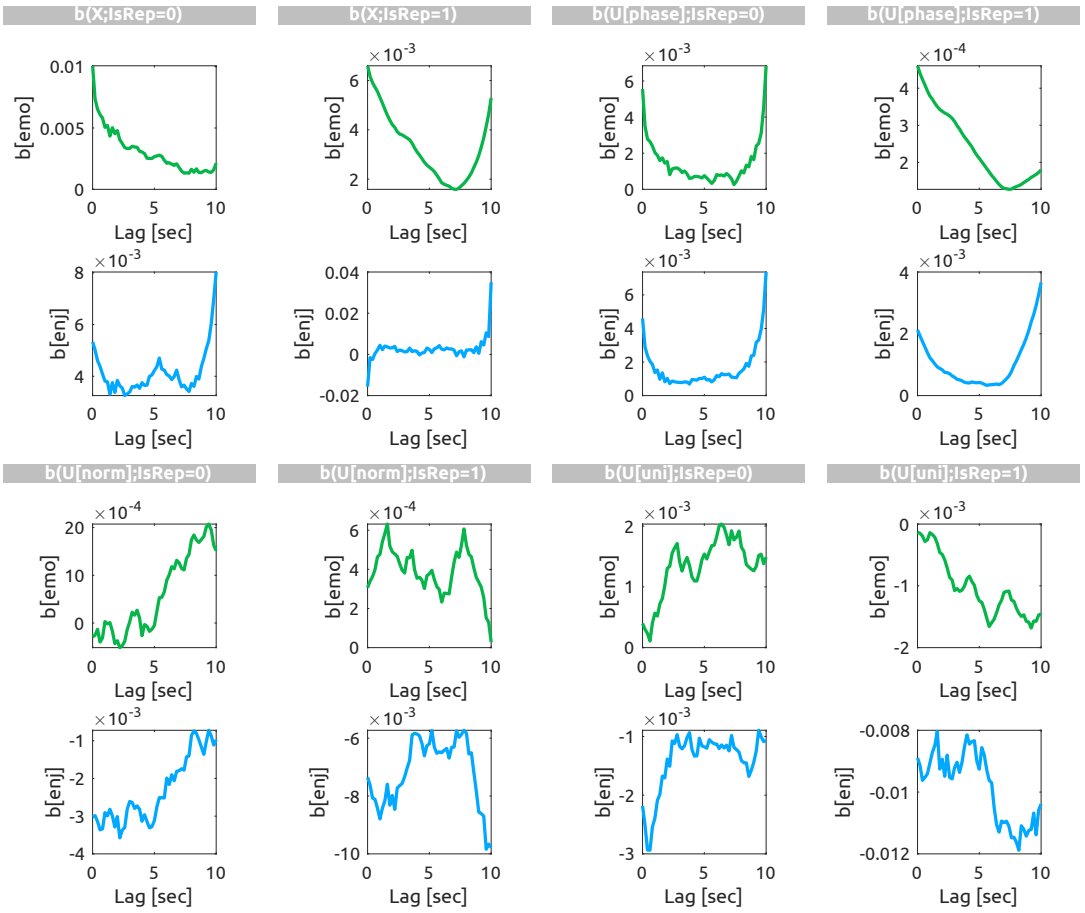


Figure 11—figure supplement 9. Transfer function weights (b) of Emotionality (green) and Enjoyment (blue) for the models with delays from 0 to 15 sec. The weights are grouped by features (X, true audio envelop; $U[phase]$, phase-randomized envelop; $U[norm]$, normal noise; $U[uni]$, uniform noise) and the CV schemes ($\texttt{IsRep} = 0$, $\texttt{IsRep} = 1$). Note that each weight time series is individually scaled.

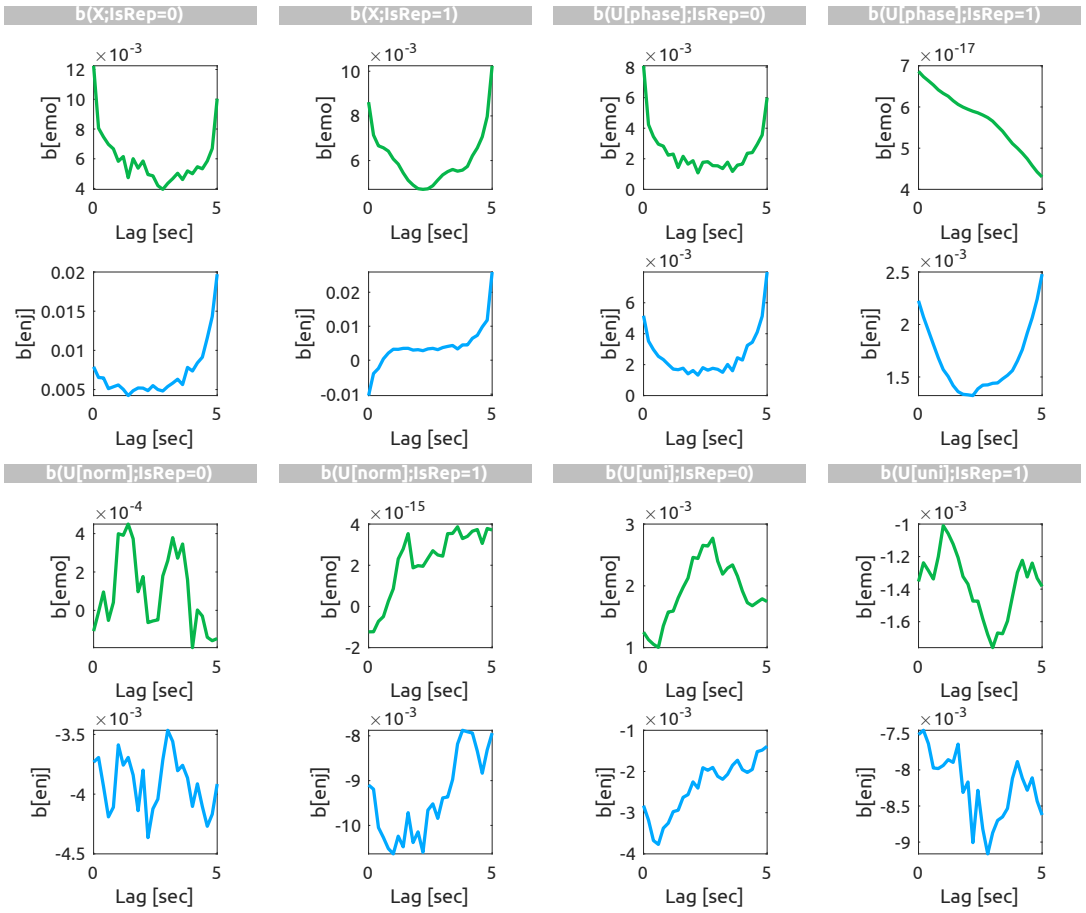


Figure 11—figure supplement 10. Transfer function weights (b) of the models with delays from 0 to 5 sec are displayed. The visualization scheme is identical to ***Figure 11—figure Supplement 9***.

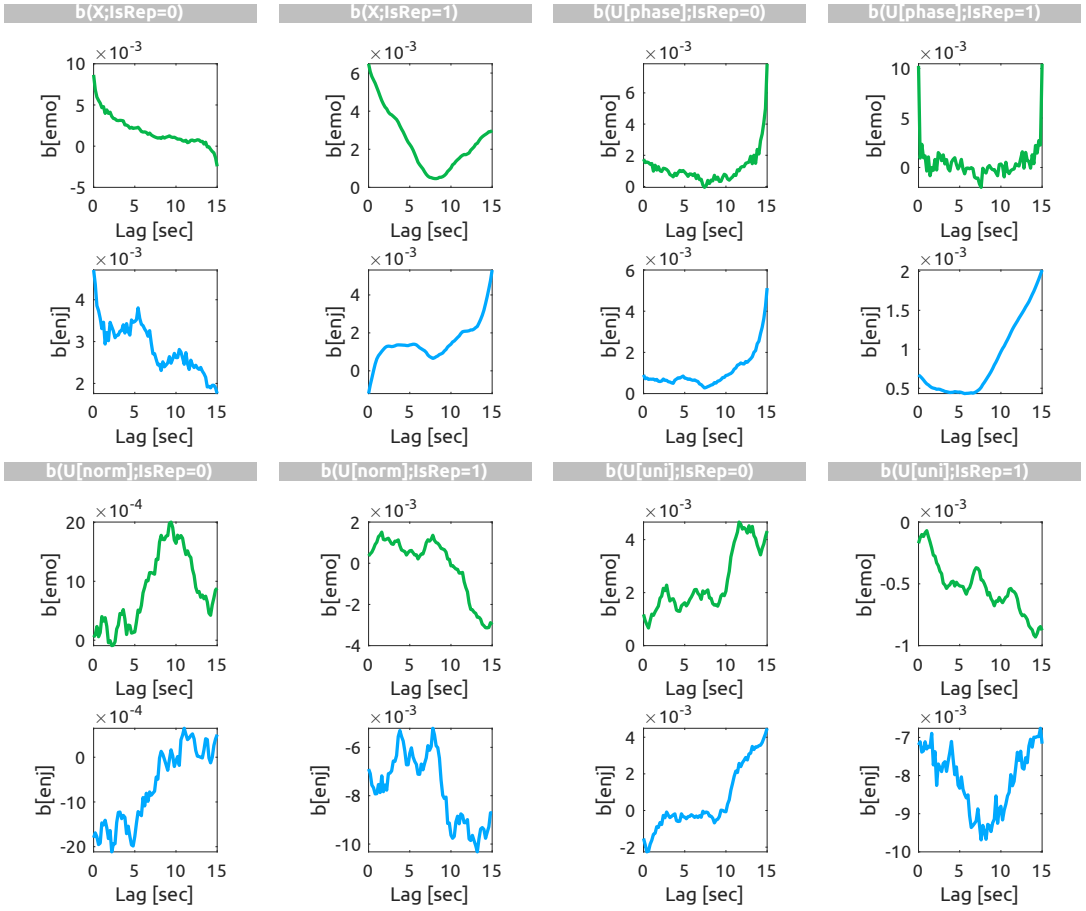


Figure 11—figure supplement 11. Transfer function weights (b) of the models with delays from 0 to 15 sec are displayed. The visualization scheme is identical to ***Figure 11—figure Supplement 9***.