THE SHUFFLE ESTIMATOR FOR EXPLAINABLE VARIANCE IN FMRI EXPERIMENTS[1]

BY YUVAL BENJAMINI AND BIN YU

Stanford University and University of California, Berkeley

In computational neuroscience, it is important to estimate well the proportion of signal variance in the total variance of neural activity measurements. This explainable variance measure helps neuroscientists assess the adequacy of predictive models that describe how images are encoded in the brain. Complicating the estimation problem are strong noise correlations, which may confound the neural responses corresponding to the stimuli. If not properly taken into account, the correlations could inflate the explainable variance estimates and suggest false possible prediction accuracies.

We propose a novel method to estimate the explainable variance in functional MRI (fMRI) brain activity measurements when there are strong correlations in the noise. Our shuffle estimator is nonparametric, unbiased, and built upon the random effect model reflecting the randomization in the fMRI data collection process. Leveraging symmetries in the measurements, our estimator is obtained by appropriately permuting the measurement vector in such a way that the noise covariance structure is intact but the explainable variance is changed after the permutation. This difference is then used to estimate the explainable variance. We validate the properties of the proposed method in simulation experiments. For the image-fMRI data, we show that the shuffle estimates can explain the variation in prediction accuracy for voxels within the primary visual cortex (V1) better than alternative parametric methods.

1. Introduction. Neuroscientists study how human perception of the outside world is physically encoded in the brain. Although the brain's processing unit, the neuron, performs simple manipulations of its inputs, hierarchies of interconnected neuron groups achieve complex perception tasks. By measuring neural activities at different locations in the hierarchy, scientists effectively sample different stages in the cognitive process.

Functional MRI (fMRI) is an indirect imaging technique, which allows researchers to sample a correlate of neural activities over a dense grid covering the brain. FMRI measures changes in the magnetic field caused by flow of oxygenated blood; these blood oxygen-level dependent (BOLD) signals are indicative

of neuronal activities. Because it is noninvasive, fMRI can record neural activity from a human subject's brain while the subject performs cognitive tasks that range from basic perception of images or sound to higher-level cognitive and motor actions. The vast data collected by these experiments allow neuroscientists to develop quantitative models, *encoding models* [Dayan and Abbott (2001)], that relate the cognitive tasks with the activity patterns these tasks evoke in the brain. Encoding models are usually fit separately to each point of the spatial activity grid, a *voxel*, recorded by fMRI. Each fitted encoding model extracts features of the perceptual input and summarizes them into a value reflecting the evoked activity at the voxel.

Encoding models are important because they can be quantitatively evaluated based on how well they can predict on new data. Prediction accuracy of different models is thus a yardstick to contrast competing models regarding the function of the neurons spanned by the voxel [Carandini et al. (2005)]. Furthermore, the relation between the spatial organization of neurons along the cortex and the function of these neurons can be recovered by feeding the model with artificial stimuli. Finally, predictions for multiple voxels taken together create a predicted fingerprint of the input; these fingerprints have been successfully used for extracting information from the brain [so-called "mind-reading", Nishimoto et al. (2011)] and building brain machine interfaces [Shoham et al. (2005)]. The search for simpler but more predictive encoding models is ongoing, as researchers try to encode more complex stimuli and predict higher levels of cognitive processing.

Because brain responses are not deterministic, encoding models cannot be perfect. A substantial portion of the fMRI measurements is noise that does not reflect the input. The noise may be caused by background brain activity, by noncognitive factors related to blood circulation or by the measurement apparatus. Regardless of the source, the noise cannot be predicted by encoding models that are deterministic functions of the inputs [Roddey, Girish and Miller (2000)]. To reduce the effect of noise, the same input can be displayed multiple times within the input sequence and all responses to the same input averaged, in an experimental design called event-related fMRI [Josephs, Turner and Friston (1997)]. See Pasupathy and Connor (1999), Haefner and Cumming (2008), for examples, and Huettel (2012) for a review. Typically, even after averaging, the noise level is high enough to be a considerable source of prediction error. Hence, it is standard practice to measure and report an indicator of the signal strength together with prediction success. We focus on *explainable variance*, the proportion of signal variance in the total variance of the measurements,[2] which we here formally define in equation (2.9). The comparison of explainable variance with prediction success [Roddey, Girish and Miller (2000), Sahani and Linden (2003)] informs how much room is left on the

---

[2]Depending on context, this proportion is also known as the intraclass correlation, effect-size and R^2.

collected data for improving prediction through better models. Explainable variance is also an important quality control metric before fitting encoding models, and can help choose regularization parameters for model training.

In this paper we develop a new method to estimate the explainable variance in fMRI responses and use it to reanalyze data from an experiment conducted by the Gallant lab at UC Berkeley [Kay et al. (2008), Naselaris et al. (2009)]. Their work examines the representation of visual inputs in the human brain using fMRI by ambitiously modeling a rich class of images from natural scenes rather than artificial stimulus. An encoding model was fit to each of more than 10,000 voxels within the visual cortex. The prediction accuracy of their fitted models on a separate validation image set was surprisingly high given the richness of the input class, inspiring many studies of rich stimuli class encoding [Pasley et al. (2012), Pereira, Detre and Botvinick (2011)]. Still, accuracy for the voxels varied widely (see Figure 2), and more than a third of the voxels had prediction accuracy not significantly better than random guessing. Researchers would like to know whether accuracy rates reflect (a) overlooked features which might have improved the modeling, or instead reflect (b) the noise that cannot be predicted regardless of the model used. As we show in this paper, valid measures of explainable variance can shed light on this question.

Measuring explainable variance on correlated noise. We face the statistical problem of estimating the explainable variance, assuming the measurement vector is composed of a random mean-effects signal evoked by the images with additive auto-correlated noise [Scheffé (1959)]. In fMRI data, many of the sources of noise would likely affect more than one measurement. Furthermore, low frequency correlation in the noise has been shown to be persistent in fMRI data [Fox and Raichle (2007)]. Ignoring the correlation would greatly bias the signal variance estimation (see Figure 7 below) and would cause us to overestimate the explainable variance. This overestimation of signal variance may be a contributing factor to replicability concerns raised in neuroscience [Vul et al. (2009)].

Classical analysis-of-variance methods account for correlated noise by (a) estimating the full noise covariance, and (b) deriving the variances of the signal and the averaged noise based on that covariance. The two steps can be performed separately by methods of moments [Scheffé (1959)] or simultaneously using restricted maximum likelihood [Laird, Lange and Stram (1987)]. In both cases, some parametric model for the correlation is needed for the methods to be feasible, for example, a fast decay is assumed in Woolrich et al. (2001). These approaches, however, are sensitive to misspecifications of the correlation parameters. In fMRI signals, the correlation of the noise might vary with the specifics of the preprocessing method in a way, that is, not easy to parametrize. As we show in Section 6, if the autocorrelation model is too simplistic, it might not capture the correlation well and then overestimate the signal; on the other hand, if it is too flexible, the noise

might be overestimated and, furthermore, the numeric optimizations involved in estimating the correlation might sometimes fail to converge.

An alternative way [Sahani and Linden (2003), Hsu, Borst and Theunissen (2004)] to handle the noise correlation when estimating variances is to restrict the analysis to measurements that, based on the data collection, should be independent. Many neuroscience experiments are divided into several sessions, or *blocks*, to better reflect the inherent variability and to allow the subject rest. Fewer have a *block design*, where the same stimulus sequence is repeated for multiple blocks. Under block design the signal level can be estimated by comparing repeated measures across different blocks: regardless of the within-block-correlation, the noise should decay as $1/b$ when averaged over b blocks with the same stimulus sequence. Block designs, however, are quite limiting for fMRI experiments, because the long reaction time of fMRI limits the number of stimuli that can be displayed within an experimental block [Huettel (2012)]. The methods above also do not use repeats within a block to improve their estimates. These problems call for a method that can make use of patterns in the data collection to estimate the signal and noise variances under less restrictive designs.

We introduce novel variance estimators for the signal and noise levels, which we call *shuffle estimators*. Shuffle estimators resemble bias correction methods: we think of the noise component as a bias and try to remove it by resampling. The key idea is to artificially create a second data vector that will have similar noise patterns as our original data. We do this by permuting, or *shuffling*, the original data along symmetries that are based on the data collection, such as temporal stationarity or independence across blocks. As we prove in Section 3, the variance due to signal will be reduced in the shuffled data when some repeated measures for the same image are shuffled into different categories. An unbiased estimator of the signal level can be derived based on this reduction in variance. The method does not require parametrization of the noise correlation, and is flexible to incorporate different structures in the data collection.

We validate our method on both simulated and fMRI data. For the fMRI experiment, we estimate upper bounds for prediction accuracy based on the explainable variance of each voxel in the primary visual cortex (V1). The upper bounds we estimate (in Section 6) are highly correlated ($r > 0.9$) to the accuracy of the prediction models used by the neuroscientists. We therefore postulate that explainable variance, as estimated by the shuffle estimators, can "predict" optimal accuracy even for areas that do not have a good encoding model. Alternative estimates for explainable variance showed substantially less agreement with the prediction results of the voxels.

In Section 2 we describe the fMRI experiment in greater detail and motivate the random effects model underlying our analysis. In Section 3 we introduce the shuffle estimators method for estimating the signal and noise levels and prove the estimators are unbiased. In Section 4 we focus on the relation between explainable

variance and prediction for random effects model with correlated noise. The simulations in Section 5 verify unbiasedness of the signal estimates for various noise regimes and show that the estimates are comparable to parametric methods with the correct noise model. In Section 6 we estimate the explainable variance for multiple voxels from the fMRI experiment and show the shuffle estimates outperform alternative estimates in explaining variation in prediction accuracies of the voxels. Section 7 concludes this paper with a discussion of our method. The proofs and the conditions of consistency for the estimator are available in the supplementary materials [Benjamini and Yu (2013)].

2. Preliminaries.

2.1. *An fMRI experiment.* In this experiment, carried out by the Gallant lab at UC Berkeley [Kay et al. (2008)], a human subject viewed natural images while scanned by fMRI.[3] The two primary goals of the experiment were (a) to find encoding models that have high predictive accuracy across many voxels in the early visual areas; and (b) to use such models to identify the input image, from a set of candidate images, based on the evoked brain patterns. The experiment created the first noninvasive machinery to successfully identify natural images based on brain patterns, and its success spurred many more attempts to encode and decode neural activities evoked by various cognitive tasks [Pasley et al. (2012), Pereira, Detre and Botvinick (2011)]. We focus only on the prediction task, but note that gains in prediction would improve the accuracy of identification as well. A complete description of the experiment can be found in the supplementary materials of the original paper [Kay et al. (2008)]. This is background for our work, which begins in Section 2.2.

The data of this experiment are composed of the set of natural images, and the fMRI scans recorded corresponding to the images. The images were sampled from a library of gray-scale photos depicting natural scenes, objects, etc. Two *nonoverlapping* random samples were taken: the training sample (1750 images) was used for fitting the models; and the validation sample ($m = 120$ images) was used for measuring the prediction accuracy. Images were sequentially displayed in a randomized order, each image appearing multiple times ($n = 13$). BOLD contrasts—correlates of neural activity—were continuously recorded by the fMRI along a three-dimensional grid covering the visual cortex, as the subject watched the images. For each voxel in the grid, the responses were temporally discretized so that a single value (per voxel) was associated with a single displayed image.

Data from the training sample was used to fit a quantitative receptive field model for each voxel, describing the fMRI response as a function of the input image. Here is a brief overview; more details on these V1 encoding models can be found in Vu

---

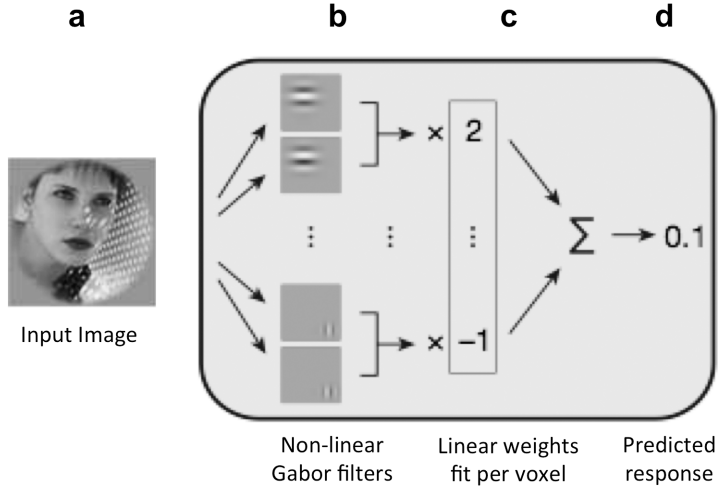[3]We use data from subject S1 in Kay et al. (2008).

FIG. 1. *A Gabor-based encoding model for natural images. A cartoon depicting the encoding models that were used for the fMRI experiment (adapted from Kay et al.). In these models, natural images* (a) *are transformed into a vector of* 10,409 *features; features* (b) *measure the energy from two linear Gabor filters—tuned for specific spatial frequency, location in the image and orientation—with complementary phases. These features are combined according to a linear weight vector* (c) *that was fit separately for each voxel. The predicted response of a voxel for an image is the weighted sum of the features representing the image* (d). *The linear weights were fit on the training data consisting of responses to* 1750 *images.*

et al. (2011). The model was based on multiple Gabor filters capturing spatial location, orientation and spatial-frequency of edges in the images (see Figure 1). Because of the tuning properties of the Gabor energy filters, this filter set is typically used for representing receptive fields of mammalian V1 neurons. Gabor filters ($d = 10,409$ filters) transformed each image into a feature vector in $\mathbb{R}^d$. For each of Q voxels of interest, a linear weight vector relating the features to the measurements was estimated from the training data. Together, the transformation and linear weight vector result in a *prediction rule* that maps novel images to a real-valued response per voxel.

In their paper, Kay et al. measured prediction accuracy by comparing observations from the validation sample with the predicted responses for those images. The validation data consisted of a total of $T = 1560$ measurements (per voxel): $m = 120$ different images, each repeated $n = 13$ times. Let $Y_t^{(r)} \in \mathbb{R}$ denote the measured fMRI activity at voxel r and time t. Repeated measurements of the same image were averaged to reduce noise:

$$\bar{Y}_j^{(r)} = \operatorname*{avg}_{t\,:\,h(t)=j} Y_t^{(r)}, \qquad j = 1, \ldots, m,$$

where $h(t) \in \{1, \ldots, m\}$ indexes the image that was shown at time t.
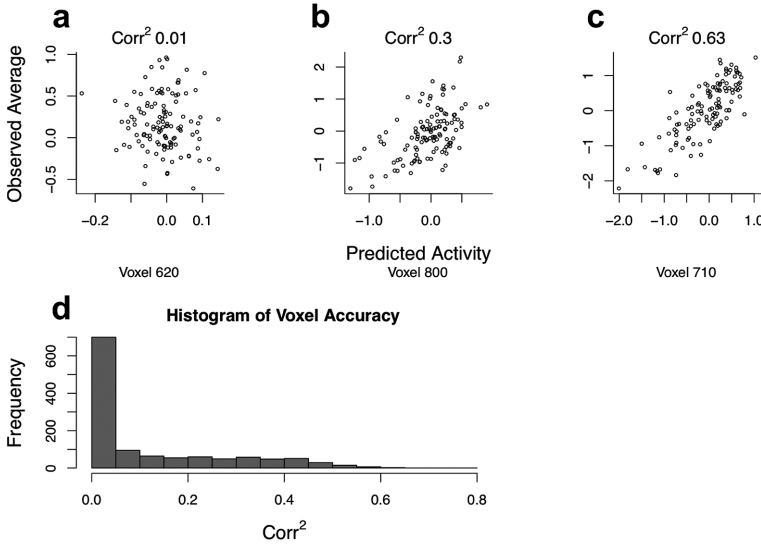
FIG. 2. *Prediction accuracy for* V1 *voxels. Predicted vs. observed average responses for three voxels in the* V1 *area, reflecting poor* (a), *medium* (b) *and high* (c) *prediction accuracies. Each point depicts the predicted response* (x*-axis*) *and the observed average response* (across all repeats) *for an image of the validation sample* (m = 120 *images*). (d) *Histogram of prediction accuracy for all* 1250 V1 *voxels.*

Let $f_1^{(r)}, \ldots, f_m^{(r)}$ be the sequence of predictions for voxel r, and $\bar{f}^{(r)}$ their average. A single value per voxel summarizes prediction accuracy. That is,

$$\text{Corr}^2\big[(f_j^{(r)})_{j \leq m}, (\bar{Y}_j^{(r)})_{j \leq m}\big] := \frac{(\sum_{j=1}^m (f_j^{(r)} - \bar{f}^{(r)})(\bar{Y}_j^{(r)} - \bar{Y}^{(r)}))^2}{\sum_{j=1}^m (f_j^{(r)} - \bar{f}^{(r)})^2 \sum_{j=1}^m (\bar{Y}_j^{(r)} - \bar{Y}^{(r)})^2}.$$

In Figure 2 we show examples of voxels with low, intermediate and high prediction accuracies, and a histogram of the accuracy for all 1250 voxels located within the V1 area. We can drop the superscript r because each voxel is analyzed separately.

2.2. *Correlation in the data.* The goal of our work is to separate the two factors that determine the accuracy of prediction rules: the adequacy of the feature set or the linear model; and the noise level. Explainable variance represents the optimal accuracy if prediction were unrestricted by the choice of features and model.

We validate our explainable variance estimators by showing the estimators account well for the differences in prediction accuracy between $Q = 1250$ voxels within the primary visual cortex (V1). This analysis depends on the assumption that the observed variation between voxels is primarily due to differences in the level of the signal-to-noise in the validation data, rather than, for example, differences in the adequacy of the feature set underlying the prediction models. Once the estimators are validated on this controlled setting, explainable variance can be

used more broadly, for example, to compare the predictability levels of different functional areas.

Since we intend to use the validation data to estimate the explainable variance, we now give a few more details on how it was collected. Recall that the validation data consisted of $m = 120$ images each repeated $n = 13$ times [see Figure 3(a)]. These data were recorded in 10 separate sessions so that the subject could rest between sessions; the fMRI was recalibrated at the beginning of each session.
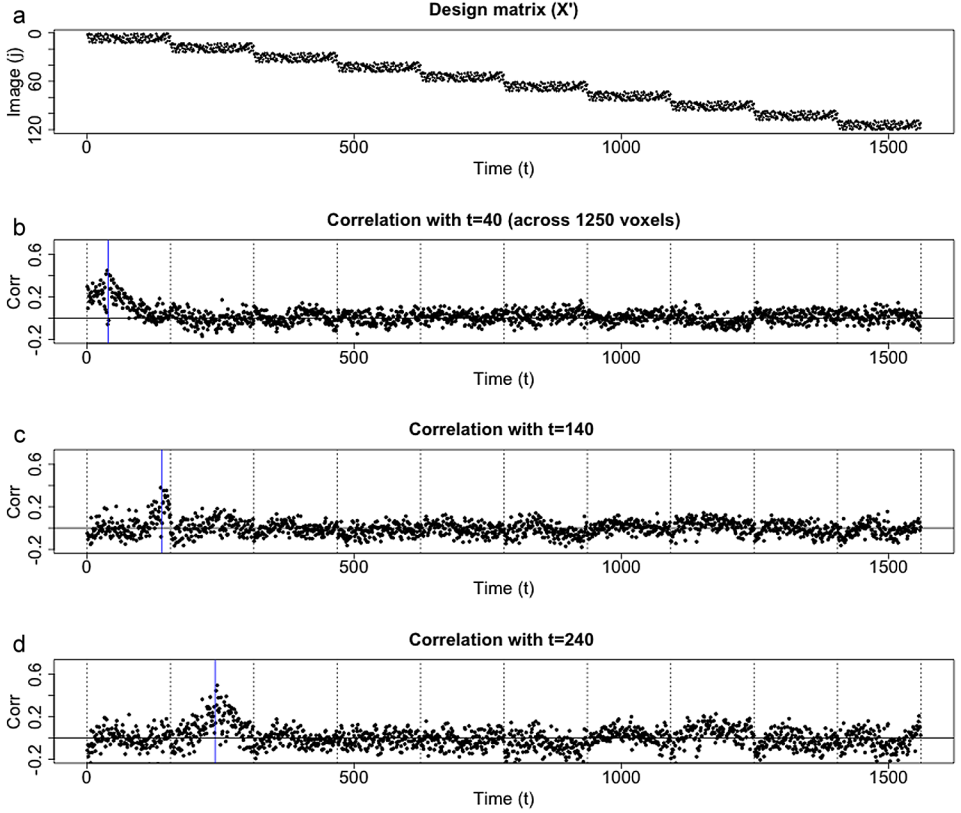

FIG. 3.   *Data acquisition for the validation data. The responses in the validation set were collected in* 10 *separate sessions* (blocks). *Both the design matrix for the experiment and the noise correlation are influenced by this structure. In* (a), *the transposed design matrix* X' *is shown. This matrix records which image* (*y-axis*) *is displayed at each time slot* t (*x-axis*). *Separate sets of* 12 *images were repeated* $n = 13$ *times within each block, whereas no image was recorded in more than one block. In* (b)–(d) *temporal autocorrelation is displayed, as measured between a single time point* [$t^* = 40, 140, 240$ *for* (b), (c), (d), *resp.*] *and all others points. The* $t = t^*$ *point is marked by a blue vertical line. On average, strong but nonsmooth correlations are found within the blocks, but separate blocks seem uncorrelated. Note that we calculate here the aggregate correlation of the voxels, but cannot use this same method to measure the autocorrelation of any specific voxel. Furthermore, the correlation depicted here is due to both noise and signal.*

Each session contained all presentations of 12 different images. A pseudo-random integer sequence ordered the repeats within a session.[4]

When we measure correlation across many voxels, it appears that the design of the experiment induces strong correlation in the data. To see this, in Figure 3(b)–(d) we plot the correlation between measurements at different time slots (each time slot is represented by the vector of $Q = 1250$ measurements). This plot shows qualitatively the correlation among individual voxels, including both noise driven and possibly stimuli-driven correlations. Clearly, there are strong correlations between time slots within a block, but no observable correlations between blocks. As these within-block correlation patterns do not correspond to the stimuli schedule, that is, randomized within a block, we conclude the correlations are largely due to noise. These noise correlations need to be taken into account to correctly estimate the explainable variance.

2.3. *A probability model for the measurements.* We introduce a probabilistic model for the measurements $\mathbf{Y} := (Y_t)_{t=1}^{T}$ at a single voxel. $\mathbf{Y}$ is modeled as a random effects model with additive, correlated noise [Williams (1952)]. We assume the observed response is the sum of two independent random processes: the signal process, which is random due to sampling of images into the validation set, and the noise process describing fluctuations unrelated to the stimuli. Additivity of noise is considered a good approximation for fMRI-event related designs and is commonly used [Buracas and Boynton (2002)]. The random effects model accounts for the generalization of prediction accuracy from the validation sample to the larger population of natural images.

2.3.1. *Signal.* Images are shown in a long sequence, in which each of the m images in the random sample is repeated multiple times. The order of presentation is described by the design matrix $X \in \{0, 1\}^{T \times m}$ (see Figure 3). Each row vector X_t has a single 1 entry, which identifies the image shown at time t,

$$X_{t,j} = 1 \qquad \text{if the } j\text{th image in the sample is displayed at time } t.$$

For now, the sampling of images is conveyed through their effects on the measurement. We use a homogenous set of random variables [Owen (2007)] to represent the mean responses to the images in the sample. Let A_j be the mean of the observed responses to the jth image in the sample. We assume that

$$A_j \sim (0, \sigma_A^2) \qquad \text{for } j = 1, \dots, m,$$

that is, A_j's are i.i.d. with mean 0 and variance $\sigma_A^2 \geq 0$, but are not necessarily normally distributed. In the experiment we are analyzing, the randomness originates from sampling a large (infinite) population of images. We use $\mathbf{A} := (A_j)_{j=1}^{m}$

---

[4]The pseudo-random sequence allocated spots for 13 different images; no image was shown in the last category and the responses were discarded.

for the random-effect column vector, $\bar{A} := \frac{1}{m}\sum_{j=1}^{m} A_j$ for the sample mean, and $s_A^2 := \frac{1}{(m-1)}\sum_{j=1}^{m}(A_j - \bar{A})^2$ for the sample variance of the random effects.

$X\mathbf{A}$ is the T-dimensional random signal vector, denoting the random effect at each measurement. To index the effect corresponding to time t, we use the shorthand $A_{(t)} := X_t A$.

2.3.2. *Noise.* We assume the measurement noise process $\varepsilon = (\varepsilon_t)_{t=1}^{T}$ is independent of the random signal vector $X\mathbf{A}$. We further assume the noise elements ε_t have 0 mean, $\sigma_\varepsilon^2 \geq 0$ variance, but may be autocorrelated. The unknown correlation, denoted by a matrix $\Sigma \in \mathbb{R}^{T \times T}$ ($\mathrm{diag}(\Sigma) = \mathbf{1}$), captures the slow-changing hemodynamics of the BOLD and the effects of preprocessing on the BOLD signals. Hence,

$$(2.1) \qquad \mathbb{E}_\varepsilon[\varepsilon_t] = 0, \qquad \mathrm{cov}(\varepsilon_t, \varepsilon_u) = \sigma_\varepsilon^2 \Sigma_{tu}, \Sigma_{tt} = 1,$$

or in matrix notation $\mathrm{cov}[\varepsilon] = \sigma_\varepsilon^2 \Sigma$.

2.3.3. *Model for observed responses.* We are now ready to introduce the observed response (column) vector $\mathbf{Y} \in \mathbb{R}^T$ as follows:

$$(2.2) \qquad \mathbf{Y} = X\mathbf{A} + \varepsilon$$

and for a single time slot t

$$Y_t = A_{(t)} + \varepsilon_t.$$

2.3.4. *Response covariance.* The model involves two independent sources of randomness: the image sampling, modeled by the random effects ($\mathbf{A}$), and the measurement noise ε. Assuming independence between $\mathbf{A}$ and ε, the covariance of Y_t and Y_u amounts to adding the individual covariances

$$(2.3) \qquad \begin{aligned} \mathrm{cov}_{A,\varepsilon}(Y_t, Y_u) &= \mathrm{cov}_A(A_{(t)}, A_{(u)}) + \mathrm{cov}_\varepsilon(\varepsilon_t, \varepsilon_u) \\ &= \sigma_A^2 \mathbf{1}_{(X_t = X_u)} + \sigma_\varepsilon^2 \Sigma_{tu}. \end{aligned}$$

The first term on the RHS shows that treatment (random) effects are uncorrelated if they are based on different inputs, but are identical if based on the same input, with a variance of σ_A^2. In matrix form, we get

$$(2.4) \qquad \mathbb{E}_{A,\varepsilon}[\mathbf{Y}] = 0, \qquad \mathrm{cov}_{A,\varepsilon}(\mathbf{Y}) = \sigma_A^2 X X' + \sigma_\varepsilon^2 \Sigma.$$

2.4. *Explainable variance and variance components.* We are ready to define explainable variance, a scaled version of the treatment variance σ_A^2. Explainable variance is relevant to the performance of prediction models, a property we will discuss in Section 4.

Recall that $\bar{Y}_j$ are the averaged responses per image ($j = 1, \ldots, m$ for the images in our sample), and let $\bar{Y} = \frac{1}{T} \sum_{t=1}^{T} Y_t$ be the global average response. Then the sample variance of averages is

$$(2.5) \qquad \mathrm{MS_{bet}} := \frac{1}{m-1} \sum_{j=1}^{m} (\bar{Y}_j - \bar{Y})^2.$$

The notation $\mathrm{MS_{bet}}$ refers to the mean-of-squares between treatments. Let us define the total variance $\bar{\sigma}_Y^2$ as the population mean of $\mathrm{MS_{bet}}$,

$$(2.6) \qquad \bar{\sigma}_Y^2 := \mathbb{E}_{A,\varepsilon}[\mathrm{MS_{bet}}].$$

Note that $\bar{\sigma}_Y^2$ is not strictly the variance of any particular $\bar{Y}_j$; indeed, the variance of $\bar{Y}_j$ is not necessarily equal for different j's.[5] Nevertheless, we will loosely use the term *variance* here and later, owing to the parallels between these quantities and the variances in the i.i.d. noise case.

The average across repeats $\bar{Y}_j$ is composed of a signal part (A_j) and an average noise part ($\bar{\varepsilon}_j$); similarly, $\bar{Y}$ is composed of $\bar{A}$ and $\bar{\varepsilon}$. By partitioning the $\mathrm{MS_{bet}}$, and taking expectations over the sampling and the noise, we get

$$(2.7) \quad \mathbb{E}_{A,\varepsilon}[\mathrm{MS_{bet}}] = \mathbb{E}_A \left[\frac{1}{m-1} \sum_{j=1}^{m} (A_j - \bar{A})^2 \right] + \mathbb{E}_\varepsilon \left[\frac{1}{m-1} \sum_{j=1}^{m} (\bar{\varepsilon}_j - \bar{\varepsilon})^2 \right],$$

where the cross-terms cancel because of the independence of the noise from the sampling. We can call the expectation of the second term the *noise level*, or $\bar{\sigma}_\varepsilon^2$, and get the following decomposition:

$$(2.8) \qquad \bar{\sigma}_Y^2 = \sigma_A^2 + \bar{\sigma}_\varepsilon^2.$$

In other words, the signal variance σ_A^2 and the noise level $\bar{\sigma}_\varepsilon^2$ are the signal and noise components of the total variance.

Finally, we define the proportion of *explainable variance* to be the ratio

$$(2.9) \qquad \omega^2 := \sigma_A^2 / \bar{\sigma}_Y^2.$$

Explainable variance measures the proportion of variance due to signal in the averaged responses; estimating it is the goal of this work. Note that by definition, ω^2 is restricted to $[0, 1]$.

To estimate ω^2, we need estimators for $\bar{\sigma}_Y^2$ and σ_A^2. Whereas $\bar{\sigma}_Y^2$ can be directly estimated from the sample, to estimate σ_A^2 we need a method to separate the signal from the noise. In the following, we propose a method to distinguish them using their different covariance structures. We first develop some technical algebraic identities important for the estimation procedure. Some readers might prefer to skip directly to Section 3.

---

[5]In practice, this is true for the individual measurements Y_t as well. We chose $\Sigma_{tt} = 1$ for illustration reasons.

2.5. *Quadratic contrasts.* In this subsection we express MS_{bet} as a quadratic contrast of the full data vector $\mathbf{Y}$. This contrast highlights the relation between $\bar{\sigma}_Y^2$ or $\bar{\sigma}_\varepsilon^2$ with both the design XX' and the measurement correlations Σ, and produces algebraic descriptions to be used for deriving the shuffle estimator. These are simple extensions of classical treatment of variance components [Townsend and Searle (1971)].

Denote $B := XX'/n$, the $\mathbb{R}^{T \times T}$ matrix in which

$$(2.10) \qquad B_{tu} = \begin{cases} \dfrac{1}{n}, & \text{if } X_t = X_u, \\ 0, & \text{otherwise,} \end{cases}$$

where B is an averaging matrix, meaning that multiplication of a measurement vector by B replaces each element in the vector by the treatment average, as in

$$(2.11) \qquad (B\mathbf{Y})_t = \bar{Y}_{h(t)}.$$

It is easy to check that $B = B'$ and $B = B^2$. Also, let $G \in \mathbb{R}^{T \times T}$, $G_{tu} = 1/T$ for $t, u = 1, \ldots, T$, be the global average matrix, so that $(G\mathbf{Y})_t = \bar{Y}, t = 1, \ldots, T$. We can now express MS_{bet} as a quadratic expression of $\mathbf{Y}$,

$$(2.12) \qquad \text{MS}_{\text{bet}} = \frac{1}{(m-1)n} \|(B - G)\mathbf{Y}\|^2,$$

or, more generally, as a quadratic function of any input vector,

$$(2.13) \qquad \text{MS}_{\text{bet}}(\cdot) := \frac{1}{(m-1)n} \|(B - G)(\cdot)\|^2.$$

By replacing the MS_{bet} with its quadratic form, a relation is exposed between the total variance, the design and the correlation of the noise:

$$\bar{\sigma}_Y^2 = \mathbb{E}_{A,\varepsilon}\big[\text{MS}_{\text{bet}}(\mathbf{Y})\big]$$

$$= \frac{1}{(m-1)n} \mathbb{E}_{A,\varepsilon}\big[\text{tr}((B - G)(\mathbf{Y}'\mathbf{Y})(B - G))\big]$$

$$= \frac{1}{(m-1)n} \text{tr}\big((B - G)\text{cov}_{A,\varepsilon}(\mathbf{Y})\big).$$

The signal effect and noise are additive, hence,

$$\frac{1}{(m-1)n} \text{tr}\big((B - G)\text{cov}_A(\mathbf{Y})\big) + \frac{1}{(m-1)n} \text{tr}\big((B - G)\text{cov}_\varepsilon(\mathbf{Y})\big).$$

Substituting $\text{cov}_A(\mathbf{Y}) = n\sigma_A^2 B$ and $\text{cov}_\varepsilon(\mathbf{Y}) = \sigma_\varepsilon^2 \Sigma$,

$$\frac{1}{(m-1)n} \text{tr}\big((B - G)(n\sigma_A^2 B)\big) + \frac{1}{(m-1)n} \text{tr}\big((B - G)\sigma_\varepsilon^2 \Sigma\big)$$

$$= \frac{1}{(m-1)}\sigma_A^2 \text{tr}(B - G) + \frac{1}{(m-1)n}\sigma_\varepsilon^2 \text{tr}\big((B - G)\Sigma\big)$$

$$= \sigma_A^2 + \frac{1}{(m-1)n}\sigma_\varepsilon^2 \text{tr}\big((B - G)\Sigma\big).$$

DERIVATION 1. *Under the model described in Section* 2.3,

$$(2.14) \qquad \bar{\sigma}_Y^2 = \sigma_A^2 + \frac{1}{(m-1)n} \sigma_\varepsilon^2 \operatorname{tr}\big((B-G)\Sigma\big).$$

As a direct consequence of (2.8) and (2.14), we get an exact expression for the noise level

$$(2.15) \qquad \bar{\sigma}_\varepsilon^2 = \frac{1}{(m-1)n} \sigma_\varepsilon^2 \operatorname{tr}\big((B-G)\Sigma\big).$$

This expression clarifies how $\bar{\sigma}_\varepsilon^2$ depends on the design, the noise variance and the noise autocorrelation. As expected, $\bar{\sigma}_\varepsilon^2$ scales linearly with the noise variance of the individual measurements σ_ε^2. More interesting is that $\bar{\sigma}_\varepsilon^2$ depends linearly on $\operatorname{tr}(B\Sigma)$—the interplay between the design and the noise autocorrelation.

Note that if the within treatment noise is uncorrelated, this expression simplifies to a classical ANOVA result. Uncorrelated noise within treatments manifests, in a properly sorted version of Σ, as small $n \times n$ identity blocks. Therefore, $\operatorname{tr}((B-G)\Sigma) = (m-1)\sigma_\varepsilon^2$ and $\bar{\sigma}_\varepsilon^2 = \sigma_\varepsilon^2/n$. In that case $\bar{\sigma}_Y^2 = \sigma_A^2 + \sigma_\varepsilon^2/n$, and by plugging in an estimator of σ_ε^2, we can directly estimate $\bar{\sigma}_\varepsilon^2$ and σ_A^2. The estimator for ω^2 is

$$\hat{\omega}^2 = 1 - \frac{1}{F}$$

with F being the standard F statistic. This is the method-of-moments estimator described fully in Section 6.1.

On the other hand, when some correlations within repeats are greater than 0, σ_ε^2/n underestimates the noise level and inflates the explainable variance. In the next section we introduce the shuffle estimators which can deal with correlated noise.

3. Shuffle estimators for signal and noise variances. In this section we propose new estimators called the shuffle estimators for the signal and noise levels, and for the explainable variance. As in (2.8), $\bar{\sigma}_Y^2 = \sigma_A^2 + \bar{\sigma}_\varepsilon^2$, but the noise level $\bar{\sigma}_\varepsilon^2$ is a function of the (unknown) measurement correlation matrix Σ. Using shuffle estimators, we can estimate σ_A^2 and $\bar{\sigma}_\varepsilon^2$ without having to estimate the full Σ or imposing unrealistically strong conditions on it.

The key idea is to artificially create a second data vector that will have similar noise patterns as our original data (see Figure 4). We do this by permuting, or *shuffling*, the original data along symmetries that we identify in the data collection. In Section 3.1 we formalize the definition of such permutations that conserve the noise correlation, and give plausible examples for neuroscience measurements. In Section 3.2 we compare the variance of averages (MS_bet) of the original data [Figure 4(b)], with the same contrast computed on the shuffled data (c). Because repeated measures for the same image are shuffled into different categories, the
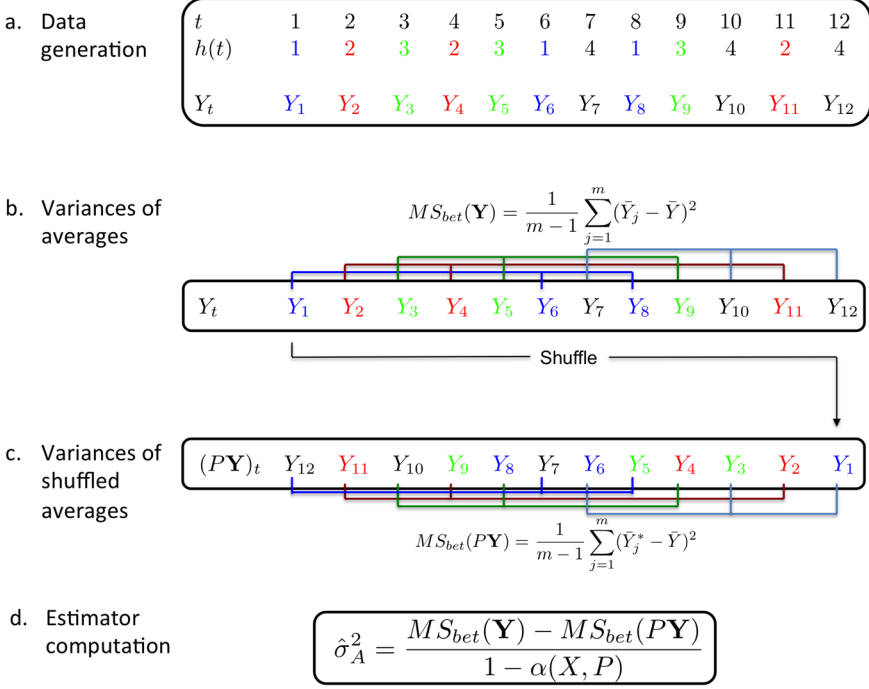
FIG. 4. *Cartoon of the shuffle estimator.* (a) *Data is generated according to a predetermined design, with each color representing repeats of a different image.* (b) *Repeats of each image are averaged together and the sample variance is computed on these averages.* (c) *Data is shuffled by P, in this example reversing the order. Now measurements which do not originate from the same repeat are averaged together ($\bar{Y}_j^*$'s), and the sample variance of the new averages is computed. These averages should have a lower variance in expectation, and we can calculate the reduction amount $\alpha = \frac{1}{m-1} \operatorname{tr}((B - G)PBP')$, where $B = XX'/n$.* (d) *The shuffle estimator for signal variance is the difference between the two sample variances, after correction of $1 - \alpha$.*

variance due to signal will be reduced in the shuffled data. We derive an unbiased estimator for signal variance σ_A^2 based on this reduction in variance, and use the plug-in estimators to estimate $\bar{\sigma}_\varepsilon^2$ and ω^2.

3.1. *Noise-conserving permutation for $\mathbf{Y}$ with respect to design X.* A prerequisite for the shuffle estimator is to find a permutation that will conserve the noise contribution to $\bar{\sigma}_Y^2$. We will call such permutations noise-conserving w.r.t. to X.

Recall (2.15),

$$\bar{\sigma}_\varepsilon^2 = \frac{1}{(m-1)n} \operatorname{tr}\big((B - G)(\sigma_\varepsilon^2 \Sigma)\big),$$

where $\sigma_\varepsilon^2 \Sigma = \operatorname{cov}_\varepsilon[\mathbf{Y}]$ as before. Let $P \in \mathbb{R}^{T \times T}$ be a permutation matrix. Then we have the following:

DEFINITION 1. P is noise conserving w.r.t. X, if

$$(3.1) \qquad \mathrm{tr}\big((B-G)P\sigma_\varepsilon^2 \Sigma P'\big) = \mathrm{tr}\big((B-G)\sigma_\varepsilon^2\Sigma\big).$$

Equivalently,

$$\mathrm{tr}\big((B-G)\operatorname{cov}_\varepsilon[P\cdot\mathbf{Y}]\big) = \mathrm{tr}\big((B-G)\operatorname{cov}_\varepsilon[\mathbf{Y}]\big).$$

Although we define the noise-conserving property based on the covariance, replacing the covariance with the correlation matrix Σ would not change the class of noise-conserving permutations.

We consider first two important cases of noise conservation. Though most useful noise-conserving permutations will be derived from assumptions regarding the noise correlation Σ, the noise-conservation property can be derived from the design X as well. This can be seen in the first example.

3.1.1. *Trivial noise-conserving permutations.*   A permutation P that simply relabels the treatments is not a desirable permutation, even though it is noise conserving. We call such permutations trivial:

DEFINITION 2.   A permutation P, associated with permutation function $g_P : \{1,\ldots,T\} \to \{1,\ldots,T\}$, is trivial if

$$(3.2) \qquad X_t = X_u \quad \Rightarrow \quad X_{g_P(t)} = X_{g_P(u)} \qquad \forall t,u.$$

It is easy to show that a trivial P does nothing, that is, $\mathrm{MS}_{\mathrm{bet}}(P\mathbf{Y}) = \mathrm{MS}_{\mathrm{bet}}(\mathbf{Y})$.

3.1.2. *Noise-conserving permutations based on symmetries of Σ.*   A useful class of nontrivial noise-conserving permutations is the class of symmetries in the correlation matrix Σ: a symmetry of Σ is a permutation P such that $P\Sigma P' = \Sigma$. If P is a symmetry of Σ, then P is noise conserving regardless of the design. Here are three important general classes of symmetries which are commonly applicable in neuroscience:

1. *Uncorrelated noise.* The obvious example is the uncorrelated noise case $\Sigma = I$ where all responses are exchangeable. Hence, any permutation is noise conserving.
2. *Stationary time series.* Neuroscience data typically are recorded in a long sequence containing a large number of serial recordings at constant rates. It is natural to assume that correlations between measurements will depend on the time passed between the two measures, rather than on the location of the pair within the sequence. We call this the *stationary time series*. Under this model Σ is a Toeplitz matrix parameterized by $\{\rho_d\}_{d=0}^{T-1}$, the set of correlation values $\Sigma_{t,u} = \rho_d$, where $d = |t-u|$. Though the correlation values ρ_d's are related, this parameterization does not enforce any structure on them. This robustness

is important in the fMRI data we analyze. For this model, a permutation that *reverses* the measurement vector is noise conserving

$$(P\mathbf{Y})_t = \mathbf{Y}_{T+1-t}.$$

This is the permutation we use on our data in Section 6.

The shift operators of the form $(P\mathbf{Y})_t = (\mathbf{Y})_{t+k}$ define a transformation that, up to edge effects, can be considered noise conserving.

3. *Block effect models.* Another important case is when measurements are collected in distinct sessions or blocks. Measurements from different blocks are assumed independent, but measurements within the same block may be correlated, perhaps because of calibration of the measurement equipment. We index the block assignment of time t with $\beta(t)$. A simple parameterization for noise correlation would be to let $\Sigma_{t,u} = \zeta(\beta(t), \beta(u))$ depend only on the block identity of measurements t and u. We call this the *block* structure. Under the block structure, any permutation P (associated with function g_P) that maintains the identity of blocks, meaning

$$(3.3) \qquad \beta(t) = \beta(u) \quad \Rightarrow \quad \beta(g_P(t)) = \beta(g_P(u)),$$

would be noise conserving w.r.t. any X.

The scientist is given much freedom in choosing the permutation P, and should consider both the variance of the estimator and the estimator's robustness against plausible noise-correlation structures. Establishing criteria for choosing the permutation P is the topic of current research.

3.2. *Shuffle estimators.* We can now state the main results. From the following lemma we observe that every noise-conserving permutation establishes a linear mean equation with two parameters: σ_A^2 and $\bar{\sigma}_\varepsilon^2$. The coefficient of $\bar{\sigma}_\varepsilon^2$ is 1, whereas the coefficient for σ_A^2 is the constant

$$(3.4) \qquad \alpha = \alpha(X, P) = \frac{1}{m-1} \operatorname{tr}((B - G)(PBP')),$$

which depends only on the design $B = XX'/n$ and the permutation P—both known to the scientist. The size of α reflects how well P mixes the treatments; the greater the mix, the smaller α.

LEMMA 1. *If P is a noise-conserving permutation for* $\mathbf{Y}$, *then*

(a) $\mathbb{E}_{A,\varepsilon}[\mathrm{MS}_{\mathrm{bet}}(P\mathbf{Y})] = \alpha\sigma_A^2 + \bar{\sigma}_\varepsilon^2$,
(b) $\alpha \leq 1$, *and the inequality is strict iff P is nontrivial.*

The proof of (a) involves derivations as in Section 2.5 [e.g., equation (2.14)], and the proof of (b) further requires the Cauchy–Schwarz inequality. Both are found in the supplementary materials [Benjamini and Yu (2013)].

The consequence of part (b) is that for any nontrivial P, we get a second mean equation, which is linearly independent from the equation based on the original data (because $\alpha < 1$). In other words, for a nontrivial P, the equation set

$$(3.5) \qquad \begin{cases} \mathbb{E}_{A,\varepsilon}[\mathrm{MS}_{\mathrm{bet}}(\mathbf{Y})] = \sigma_A^2 + \bar{\sigma}_\varepsilon^2, \\ \mathbb{E}_{A,\varepsilon}[\mathrm{MS}_{\mathrm{bet}}(P\mathbf{Y})] = \alpha\sigma_A^2 + \bar{\sigma}_\varepsilon^2 \end{cases}$$

has a unique solution. Solving the two equations above, we arrive at an unbiased estimator for the signal variance.

DEFINITION 3. The shuffle estimator for the signal variance is defined as

$$(3.6) \qquad \hat{\sigma}_A^2 = \frac{\mathrm{MS}_{\mathrm{bet}}(\mathbf{Y}) - \mathrm{MS}_{\mathrm{bet}}(P\mathbf{Y})}{1 - \alpha}.$$

Finally, we can plug in $\hat{\sigma}_A^2$ and $\hat{\bar{\sigma}}_Y^2 = \mathrm{MS}_{\mathrm{bet}}(\mathbf{Y})$ to get the *shuffle estimator* for the explainable variance ω^2:

$$\hat{\omega}^2 = \frac{\hat{\sigma}_A^2}{\mathrm{MS}_{\mathrm{bet}}(\mathbf{Y})} = \frac{1}{1 - \alpha}\left(1 - \frac{\mathrm{MS}_{\mathrm{bet}}(P\mathbf{Y})}{\mathrm{MS}_{\mathrm{bet}}(\mathbf{Y})}\right).$$

REMARKS.

1. An unbiased estimator of the noise level $\bar{\sigma}_\varepsilon^2$ can be derived from equations (3.6) and (2.8),

$$\hat{\bar{\sigma}}_\varepsilon^2 = \mathrm{MS}_{\mathrm{bet}} - \hat{\sigma}_A^2.$$

2. Because $\hat{\sigma}_A^2$ estimates a nonnegative quantity, it is preferable to restrict $\hat{\sigma}_A^2$ to nonnegative values by taking $\hat{\sigma}_{A+}^2 = \max\{0, \hat{\sigma}_A^2\}$.

3. The estimator is consistent under proper decay of the dependance. This statement is conditional on the asymptotic setup: explainable variance typically changes as the number of measurements (T) increases. Nevertheless, the shuffle estimator is consistent for a sequence of data sets (indexed by $k = 1, 2, \ldots$) of growing sizes $[T(k) \to \infty]$ for which total variance and explainable variance converge if (a) the number of treatments $m(k)$ grows to ∞ and (b) the dependence decays. For $\mathbf{Y}$ distributed as a multivariate gaussian,[6] a sufficient condition for (b) can be given in terms of eigenvalues of Σ:

   (b*) The largest $m - 1$ eigenvalues $\lambda_{(1)}(k), \ldots, \lambda_{(m-1)}(k)$ of the noise correlation matrix $\Sigma(k)$ satisfy

$$\frac{1}{n^2(m-1)^2}\sum_{i=1}^{m-1}\lambda_{(i)}^2(k) \to 0 \qquad \text{as } k \to \infty.$$

---

[6]More general SLLN conditions for the weakly dependent random variables $\bar{Y}_1^2(k), \ldots, \bar{Y}_{m(k)}^2(k)$ can be found in Lyons (1988).

Conditions (a) and (b*) assure that $\text{var}(\text{MS}_{\text{bet}}(\mathbf{Y}_k)) \to 0$ as $k \to \infty$. The proof relies mainly on the expression for the variance of a quadratic contrasts, as found, for example, in Searle (1971). For the proof of these results refer to the supplementary materials [Benjamini and Yu (2013)].

4. Evaluating prediction for correlated responses. Although there are many uses for estimating the explainable variance, we focus on its role in assessing prediction models. Roddey, Girish and Miller (2000) show that explainable variance upper bounds the accuracy of prediction on the sample when noise is i.i.d. We generalize their results for arbitrary noise correlation and account for generalization from sample to population.[7] As shown in Lemma 2, the noise level $\bar{\sigma}_\varepsilon^2$ is the optimal expected loss under mean square prediction error (MSPE) loss, and the explainable variance ω^2 approximates the accuracy under squared-correlation Corr^2 utility.

A more explicit notation setup is needed for studying the relation between predictions and signal responses. Let f be a prediction function that predicts a real-valued response to any possible image I, out of a large population of M images,

$$(4.1) \qquad\qquad f : \{I_i\}_{i=1}^M \to \mathbb{R}.$$

We will assume f does not depend on the sample we are evaluating, meaning that it was fit on separate data. We usually think of f as using some aspects of the image to predict the response, although we do not restrict it in any parametric way to the image.

Prediction accuracy is measured only on the m images sampled for the (nonoverlapping) validation set. Let $\mathbf{s} : \{1, \ldots, m\} \to \{1, \ldots, M\}$ be the random sampling function, and $I_{\mathbf{s}(j)}$ the random image chosen for the jth sample image.

Furthermore, let us introduce notation relating the random effects to the image sampling. For this, assume each image is associated with a mean activation quantity μ_i, so that $\sum \mu_i = 0$ and $\sum \mu_i^2 = \sigma_A^2$. Then the random effects A_j defined before can now be described $A_j = \mu_{\mathbf{s}(j)}$.

To evaluate prediction accuracy, the predicted response $f(I_{\mathbf{s}(j)})$ is compared with the averaged (observed) response for that image $\bar{Y}_j$. We consider two common accuracy measures: mean squared prediction error ($\text{MSPE}[f]$) and the squared correlation ($\text{Corr}^2[f]$), defined

$$(4.2) \qquad \text{MSPE}[f] := \frac{1}{m-1} \sum_{j=1}^m (f(I_{\mathbf{s}(j)}) - \bar{Y}_j)^2,$$

$$(4.3) \qquad \begin{aligned} \text{Corr}^2[f] &:= \text{Corr}_j^2\big(f(I_{\mathbf{s}(j)}), \bar{Y}_j\big) \\ &= \frac{(1/(m-1) \sum_{j=1}^m (f(I_{\mathbf{s}(j)}) - \bar{f}_{\mathbf{s}})(\bar{Y}_j - \bar{Y}))^2}{1/(m-1) \sum_{j=1}^m (f(I_{\mathbf{s}(j)}) - \bar{f}_{\mathbf{s}})^2 \sum_{j=1}^m (\bar{Y}_j - \bar{Y})^2}, \end{aligned}$$

---

[7]While these results may have been proved before, we have not found them discussed in similar context.

where $\bar{f}_{\mathbf{s}}$ denotes the average of the predictions for the sample.

We will state and discuss the results relating the explainable variance to optimal prediction; the proof can be found in the supplementary materials [Benjamini and Yu (2013)].

LEMMA 2.   *Let* $f^*:\{I_i\}_{i=1}^M \to \mathbb{R}$ *be the prediction function that assigns for each stimulus* I_i *its mean effect* μ_i, *or* $f^*(I_i) = \mu_i$. *Under the model described in Section* 2.3,

(a) $f^* = \arg\min_f \mathbb{E}_{A,\varepsilon} [\text{MSPE}[f]]$;
(b) $\bar{\sigma}_{\varepsilon}^2 = \mathbb{E}_{A,\varepsilon} [\text{MSPE}[f^*]] (= \min_f \mathbb{E}_{A,\varepsilon} [\text{MSPE}[f]]$ *by* (a));
(c) $\omega^2 \approx \mathbb{E}_{A,\varepsilon}[\text{Corr}^2[f^*]]$ *with a bias term smaller than* $\frac{1}{m-1}$.

Under our random effects model, the best prediction (in MSPE) is obtained by the mean effects, or f^*. More important to us, the accuracy measures associated with the optimal prediction f^* can be approximated by signal and noise levels: $\bar{\sigma}_{\varepsilon}^2$ for MSPE$[f^*]$ and ω^2 for Corr$^2[f^*]$.

The main consequence of this lemma is that the researcher does not need a "good" prediction function to estimate the "predictability" of the response. Prediction is upper-bounded by ω^2, a quantity which can be estimated without setting a specific function in mind. Moreover, when a researcher does want to evaluate a particular prediction function f, $\hat{\omega}^2$ can serve as a yardstick with which f can be compared. If Corr$^2[f] \approx \hat{\omega}^2$, the prediction error is mostly because of variability in the measurement. Then the best way to improve prediction is to reduce the noise by preprocessing or by increasing the number of repeats. On the other hand, if Corr$^2[f] \ll \hat{\omega}^2$, there is still room for improvement of the prediction function f.

5. Simulation.   We simulate data with a noise component generated from either a block structure or a times-series structure, and compute shuffle estimates for signal variance and for explainable variance. For a wide range of signal-to-noise regimes, our method produces unbiased estimators of σ_A^2. These estimators are fairly accurate for sample sizes resembling our image-fMRI data, and the bias in the explainable variance ω^2 is small compared to the inherent variability. These results are shown in Figure 5. In Figure 6 we show that under nonzero σ_A^2, the shuffle estimates have less bias and lower spread compared to the parametric model using the correctly specified noise correlation.

5.1. *Block structure.*   For the block structure we assumed the noise is composed of an additive random block effect constant within blocks (b_k, $k = 1, \ldots, B$ blocks), and an i.i.d. Gaussian term (e_t, $t = 1, \ldots, T$)

$$Y_t = A_{(t)} + b_{\beta(t)} + e_t.$$

A_j, b_k and e_t are sampled from centered normal distributions with variances $(\sigma_A^2, \sigma_b^2, \sigma_e^2)$. We used $\sigma_b^2 = 0.5, \sigma_e^2 = 0.7$, and varied the signal level $\sigma_A^2 =$
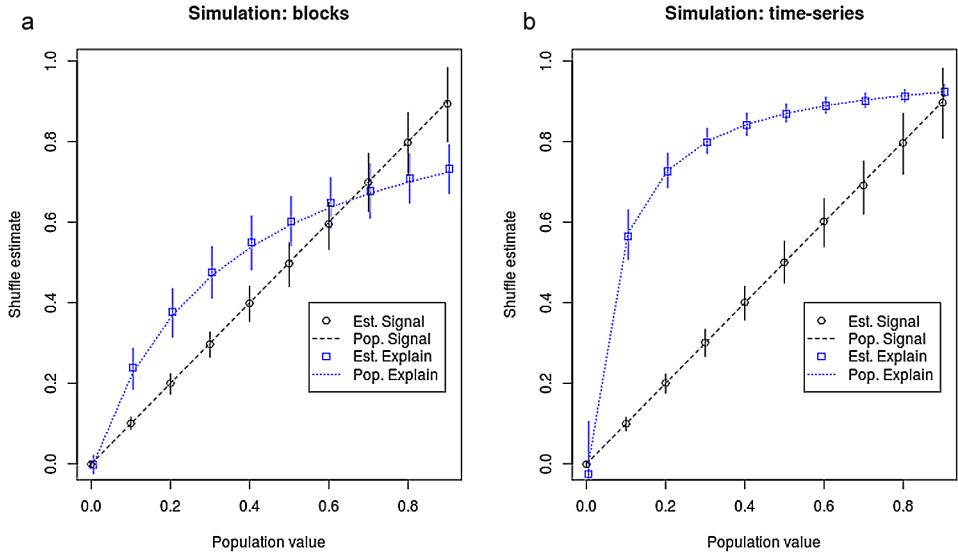
FIG. 5. *Simulations for the block and time-series.* (a) *Simulation results comparing shuffle estimates for signal variance* σ_A^2 (*black*) *and explainable variance* ω^2 (*blue*) *to the true population values* (*dashed line*). *Noise correlation followed an independent block structure*: *noise within blocks was correlated*, *and between blocks was independent. The x-axis represents the true signal variance* σ_A^2 *of the data*, *and the y-axis marks the average of the estimates and* $[0.25, 0.75]$ *quantile range*. (b) *A similar plot for data generated under a stationary time-series model.*
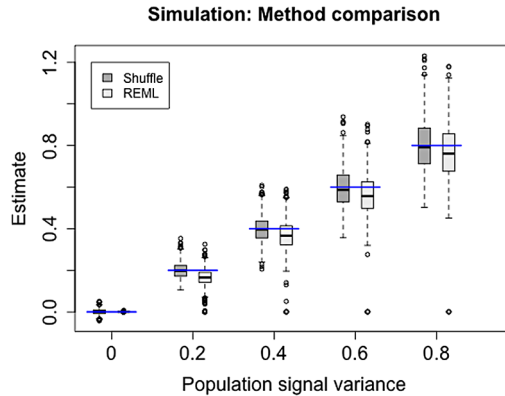

FIG. 6. *Comparison of methods on simulation. Each pair of box-plots represents the estimated signal variance* σ_A^2 *using the shuffle estimator* (*dark gray*) *and REML* (*light gray*) *for* 1000 *simulations. The blue horizontal line represents the true value of* σ_A^2. *The REML estimator assumes the correct model for the noise, while the shuffle estimator only assumes a stationary time series. When there is no signal, REML outperforms the shuffle estimators, but in all other cases it is both biased and has greater spread.*

$0, 0.1, \ldots, 0.9$. We used $m = 120$, $n = 15$, with all presentations of every 5 stimuli composing a block ($B = 20$ blocks). For each of these scenarios we ran 1000 simulations, sampling the signal, block and error effects. MS_{bet} was estimated the usual way, and P was chosen to be a random permutation within each block ($\alpha = 0.115$). The results are shown in Figure 5(a).

5.2. *Time-series model.* For the time-series model we assumed the noise vector $e \in \mathbb{R}^T$ is distributed as a multivariate Gaussian with mean 0 and a covariance matrix Σ, where Σ is an exponentially decaying covariance with a nugget,

$$\Sigma_{tu} = \rho_{|t-u|} = \lambda_1 \cdot \exp\{-|t-u|/\lambda_2\} + (1-\lambda_1)1_{(t=u)}.$$

Then $Y = A_{(t)} + e_t$ with the random effects $A_{(t)}$ sampled from $\mathcal{N}(0, \sigma_A^2)$ for $\sigma_A^2 = 0, 0.1, \ldots, 0.9$. We used $m = 120$, $n = 15$, and the parameters for the noise were $\lambda_1 = 0.7$ and $\lambda_2 = 30$, meaning $\rho_{125} \approx 0.01$. The schedule of treatments was generated randomly. For each of these scenarios we ran 1000 simulations, sampling the signal and the noise. In Figure 5(b) we estimated the shuffle estimator with P the reverse permutation ($g_P(t) = T + 1 - t$), resulting in $\alpha = 0.064$.

5.3. *Comparison to REML.* In Figure 6 we used time-series data to compare σ_A^2 estimates based on the shuffle estimators to those obtained by an REML estimator with the correct parametrization for the noise correlation matrix. We used the nlme package in R to fit a repeated measure analysis of variance for the exponentially decaying correlation of noise with a nugget effect. The comparison included 1000 simulations for $\sigma_A^2 = 0, 0.2, 0.4, 0.6, 0.8$ and a noise model identical to Section 5.2.

5.4. *Results.* Figure 5 describes the performance of shuffle estimates on two different scenarios: block correlated noise (a), and stationary time-series noise (b). For signal variance (black) the shuffle estimator gives unbiased estimates. The shuffle estimator for explainable variance is not unbiased, but the bias is negligible compared to the variability in the estimates. In Figure 6 we compare the signal variance estimates based on the shuffle estimator (dark gray) with estimates based on REML (light gray). The estimates based on the shuffle have no bias, while those based on REML underestimate the signal. The variance of the REML estimates is slightly larger, due in part to rare events (between 1%–0.5% of runs) in which the estimated signal variance was effectively 0—perhaps indicating a problem with the optimization.

6. Data. We are now ready to evaluate prediction models using the shuffle estimates for explainable variance. Prediction accuracy was measured for encoding models of 1250 voxels within the primary visual cortex (V1). Because V1 is functionally homogenous, encoding models for voxels within this cortical area should

work similarly. As observed in Figure 2, there is large variation between prediction accuracies for the different voxels. We try to explain this observed variation as a result of variation in the explainable variance. To do this, prediction accuracy values for these 1250 voxels are compared to the explainable variance estimates generated by the shuffle estimator for each voxel. We also compare the accuracy values to alternative estimates for explainable variance, using the method of moments for uncorrelated noise, and REML under several parameterizations for the noise.

6.1. *Methods.* We estimate the explainable variance of voxels ($\omega^2 = \sigma_A^2 / \bar{\sigma}_Y^2$) with several different methods. The methods differ in how σ_A^2 is estimated; all methods use the sample averages variance $\text{MS}_{\text{bet}}(\mathbf{Y})$ for $\bar{\sigma}_Y^2$ and plug in the two estimates into ω^2. We estimate ω^2 separately for each voxel ($r = 1, \ldots, 1250$). The methods we compare are as follows:

1. The shuffle estimator. We assume the noise follows a stationary time-series model within each block and is independent between the blocks. We therefore choose a permutation P that reverses the order of the measurements, $(P\mathbf{Y})_t = \mathbf{Y}_{T+1-t}$. Because the size of the blocks is identical, reversing the order of the data vector is equivalent to reversing the order within each block, $\alpha = 0.14$. Specifically, the estimator is restricted to be positive:

$$\hat{\sigma}_{A+}^2 = \max \left\{ \frac{\text{MS}_{\text{bet}}(\mathbf{Y}) - \text{MS}_{\text{bet}}(P\mathbf{Y})}{1 - \alpha}, 0 \right\}$$

   for signal variance, and $\hat{\omega}^2 = \hat{\sigma}_{A+}^2 / \text{MS}_{\text{bet}}$ for the explainable variance.

2. An estimator ($\tilde{\omega}^2$) unadjusted for correlation. We use the mean square within ($\text{MS}_{\text{wit}} = \frac{1}{(m-1)n} \sum_{j=1}^{m} \sum_{t : h(t)=j} (Y_t - \bar{Y}_j)^2$) contrast to estimate the noise variance σ_ε^2, scale by $1/n$ to estimate the noise level $\bar{\sigma}_\varepsilon^2$, and remove the scaled estimate from MS_{bet},

$$\tilde{\sigma}_A^2 = \text{MS}_{\text{bet}} - \text{MS}_{\text{wit}}/n.$$

   Explainable variance is obtained by plug-in estimator $\tilde{\omega}^2 = \tilde{\sigma}_A^2 / \text{MS}_{\text{bet}}$.

3. Estimators based on a parametric noise model.

   • We assume the noise is generated from an exponentially decaying correlation matrix, with a nugget effect. This means

$$\Sigma_{t,t+d} = \lambda_2 \exp(-d/\lambda_1) + 1_{(d=0)}(1 - \lambda_2),$$

   where the rate of decay λ_1 and nugget effect λ_2 were additional parameters. If $\lambda_2 = 0$, this is equivalent to the AR(1) model.

   • Alternatively, we assume the noise is generated from an AR(3) process or $\varepsilon_t = \eta_t + \sum_{k=1}^{3} a_k \varepsilon_{t-k}$. This models allows for nonmonotone correlations.

   We use the `nlme` package in R to estimate the signal variance of this model using restricted maximum likelihood [REML, e.g., Laird, Lange and Stram (1987)], and use the plug-in estimator for the explainable variance.
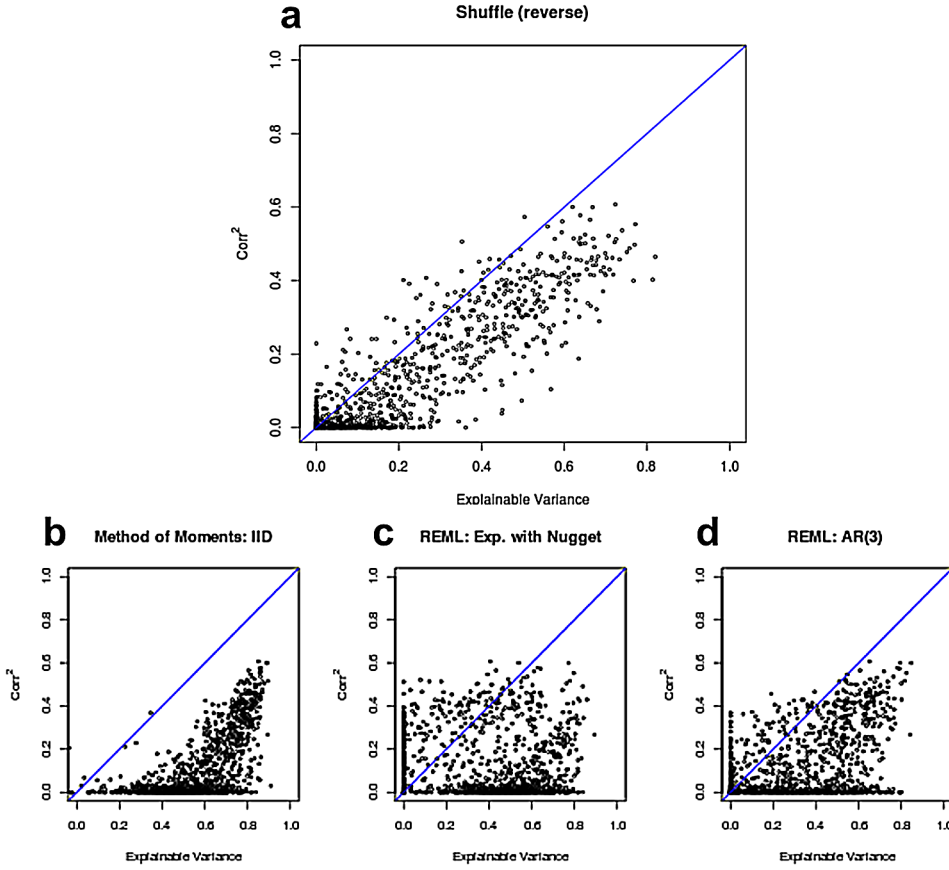
FIG. 7. *Optimal vs. observed prediction accuracy. The estimated optimal prediction is compared with observed prediction* (Corr^2), *each point representing a V1 response. The optimal prediction estimated by* (a) *shuffle estimators accounting for stationary noise distributions*; (b) *Method of moments estimator assuming independent noise*; (c) *REML estimator assuming exponential decay of noise with nugget within blocks*; *and* (d) *REML estimator assuming an* AR(3) *model for the noise correlation within blocks. The* $x = y$ *is plotted in blue.*

6.2. *Results.* In Figure 7 we compare the prediction accuracy of the voxels to estimates of the explainable variance. Each panel has 1250 points representing the 1250 voxels: the x coordinate is the estimate of explainable variance for the voxel, and the y coordinate is Corr$^2[f]$ for the Gabor-based prediction rule. The large panel shows the shuffle estimators for explainable variance. The relation between $y = $ Corr$^2[f]$ and $\hat{\omega}^2$ is very linear ($r = 0.9$). The estimated slope and intercepts (using least squares) for this relation are $y \approx 0.66 \cdot \hat{\omega}^2 - 0.009$. Note that almost all voxels for which accuracy is close to random guessing (Corr$^2[f] < 0.05$) could be identified based on low explainable variance without knowledge of the specific feature set.

When we try to repeat this analysis with other ω^2 estimators, explainable variance estimates are no longer strongly related with the prediction accuracy. When correlation in the noise is ignored (b), signal strength is greatly overestimated. In particular, some of the voxels for which prediction accuracy is almost 0 have very high estimates of explainable variance (as high as $\tilde{\omega}^2 = 0.8$). In contrast to the shuffle estimates, it is hard to learn from these explainable variance estimates about the prediction accuracy for a voxel.

This incompatibility of prediction accuracy and explainable variance estimates is also observed when the estimates are based on maximum likelihood methods that parameterize the noise matrix. For the AR(3) model in (d), we see variability between explainable variance estimates for voxels with given prediction accuracy level. The smaller model (c) seems to suffer from both overestimation of signal and from high variance.

7. Discussion. We have presented the shuffle estimator, a resampling-based estimator for the explainable variance in a random-effects additive model with autocorrelated noise. Rather than parameterize and estimate the correlation matrix of the noise, the shuffle estimator treats the contribution of the noise to the total variance as a single parameter. Symmetries in the data collection process indicate those permutations which, when applied to the original data, would not change the contribution of the noise. An unbiased estimator of the signal variance is derived from differences between the total variance of the original data vector and the shuffled vector. The resulting estimate of signal variance is plugged in as the enumerator for the explainable variance ratio estimate.

For a brain-encoding experiment, we have shown that the strong correlation present in the fMRI measurements greatly compromises classical methods for estimating explainable variance. We used prediction accuracy measures of a well-established parametric model for voxels in the primary visual cortex as correlates of signal variance at each of the voxels. Shuffle estimates explained most of the variation in prediction accuracy between voxels, even though they were blind to features of the image. Other methods did not do well: methods that ignored noise correlation seem to greatly overestimate the explainable variance, while methods that estimated the full correlation matrix were considerably less informative with regards to prediction accuracy. We consider this convincing evidence that the shuffle estimators for explainable variance can be used reliably even when no gold-standard prediction model is present.

Explainable variance is an assumption-less measure of signal, in that it makes no assumptions about the structure of the mean function that relates the input image to response. We find it attractive that the shuffle estimator for explainable variance similarly requires only weak assumptions for the correlation of the noise. This makes the shuffle estimator a robust tool, which can be used at different stages of the processing of an experiment: from optimizing of the experimental protocol, through choosing the feature space for the prediction models, to fitting the prediction models.

Choice of permutations. The bias and variance of the shuffle estimator depend on the permutation underlying the shuffle. Different permutations pose different assumptions on the noise correlation as well as provide a different mix of treatments corresponding to different αs. We recommend the permutation be chosen prior to the analysis, based on the expected noise structure and the mixing constants (αs), to minimize the risk of data snooping. Optimally, the experiment could be designed so that a specific permutation—perhaps the reverse permutation—will mix treatments well, resulting in a low α. In our experiment, the reverse and other regular permutations such as shifts had low αs because the design was generated using an irregular pseudo-random sequence. Moreover, when several noise-conserving permutations exist and have similar αs, it may be preferable to average the corresponding shuffle estimators to reduce the variance of the estimators.

In cases were no symmetry permutations are useable, a wider class of "almost noise-conserving" permutations can be considered. To give concrete examples, consider the following two permutations: A cyclic left-shift permutation so that $(P_1\mathbf{Y})_t = \mathbf{Y}_{t+1}$ for $t = 1, \ldots, T - 1$, and $(P_1\mathbf{Y})_T = \mathbf{Y}_1$; and a permutation of odd and even channels $(P_2\mathbf{Y})_{2s-1} = \mathbf{Y}_{2s}$, $(P_2\mathbf{Y})_{2s} = \mathbf{Y}_{2s-1}$, for $s = 1, \ldots, T/2$. Neither is an exact symmetry of a covariance matrix that represents a stationary time series. In P_1, the first measurement in each block is not correlated to the sequence. P_2 is even farther from symmetry, in that the medium and long-range correlations are conserved but the local structure is scrambled.

Nevertheless, the shuffle estimates from either P_1 or P_2 produce, when compared to predictions, population results that are similar to those observed for the reverse permutation. The new estimates compare in both linearity, with $r = 0.905$ between $\text{Corr}^2[f^*]$ and $\hat{\omega}^2$ for both P_1 and P_2 compared to $r = 0.9$ for the reverse permutation, as well in the slope (0.65 for P_1, 0.68 for P_2) of the linear trend and its intercept (0 for both). This does not imply that any permutation would work well; indeed, shuffling with a random permutation, implicitly assuming i.i.d. noise, results in a similar bias as that observed in the method of moments estimator shown in Figure 7(b). Note that for any candidate P and possible Σ, the degree to which noise is conserved can be explicitly measured by comparing $\text{tr}(\Sigma(B - G))$ to $\text{tr}(P'\Sigma P(B - G))$.

The shuffle estimators may be useful for applications outside of neuroscience. These estimators can be used to estimate the variance associated with the treatments of an experiment, conditioned on the design, whenever measurement noise is correlated. Spatial correlation in measurements arise in many different domains, from agricultural experiments to DNA microarray chips. Shuffle estimators could provide an alternative to parametric fitting of the noise contributions for these applications.

Future research should be directed at expressing the variance of the shuffle estimator for a candidate permutation, as well as at developing optimal ways to combine information from multiple noise-conserving permutations. More generally, shuffle estimators are a single example of adapting relatively new nonparametric

approaches from hypothesis testing into estimation; we see much room for expanding the use of permutation methods for creating robust estimators for experimental settings.

SUPPLEMENTARY MATERIAL

Supplementary material (DOI: 10.1214/13-AOAS681SUPP; .pdf). We provide proofs for Lemmas 1 and 2 in the supplementary, as well as describe and prove conditions for consistency of the shuffle estimator.

REFERENCES

BENJAMINI, Y. and YU, B. (2013). Supplement to "The shuffle estimator for explainable variance in fMRI experiments." DOI:10.1214/13-AOAS681SUPP.

BURACAS, G. T. and BOYNTON, G. M. (2002). Efficient design of event-related fMRI experiments using M-sequences. *NeuroImage* **16** 801–813.

CARANDINI, M., DEMB, J. B., MANTE, V., TOLHURST, D. J., DAN, Y., OLSHAUSEN, B. A., GALLANT, J. L. and RUST, N. C. (2005). Do we know what the early visual system does? *J. Neurosci.* **25** 10577–10597.

DAYAN, P. and ABBOTT, L. F. (2001). *Theoretical Neuroscience*: *Computational and Mathematical Modeling of Neural Systems*. MIT Press, Cambridge, MA. MR1985615

FOX, M. D. and RAICHLE, M. E. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat. Rev. Neurosci.* **8** 700–711.

HAEFNER, R. and CUMMING, B. (2008). An improved estimator of variance explained in the presence of noise. *Adv. Neural Inf. Process. Syst.* **21** 585–592.

HSU, A., BORST, A. and THEUNISSEN, F. E. (2004). Quantifying variability in neural responses and its application for the validation of model predictions. *Network*: *Comput. Neural Syst.* **15** 91–109.

HUETTEL, S. A. (2012). Event-related fMRI in cognition. *NeuroImage* **62** 1152–1156.

JOSEPHS, O., TURNER, R. and FRISTON, K. (1997). Event-related fMRI. *Hum. Brain Mapp.* **5** 243–248.

KAY, K. N., NASELARIS, T., PRENGER, R. J. and GALLANT, J. L. (2008). Identifying natural images from human brain activity. *Nature* **452** 352–355.

LAIRD, N., LANGE, N. and STRAM, D. (1987). Maximum likelihood computations with repeated measures: Application of the EM algorithm. *J. Amer. Statist. Assoc.* **82** 97–105. MR0883338

LYONS, R. (1988). Strong laws of large numbers for weakly correlated random variables. *Michigan Math. J.* **35** 353–359. MR0978305

NASELARIS, T., PRENGER, R. J., KAY, K. N., OLIVER, M. and GALLANT, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron* **63** 902–915.

NISHIMOTO, S., VU, A. T., NASELARIS, T., BENJAMINI, Y., YU, B. and GALLANT, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Curr. Biol.* **21** 1641–1646.

OWEN, A. B. (2007). The pigeonhole bootstrap. *Ann. Appl. Stat.* **1** 386–411. MR2415741

PASLEY, B. N., DAVID, S. V., MESGARANI, N., FLINKER, A., SHAMMA, S. A., CRONE, N. E., KNIGHT, R. T. and CHANG, E. F. (2012). Reconstructing speech from human auditory cortex. *PLoS Comput. Biol.* **10** e1001251.

PASUPATHY, A. and CONNOR, C. E. (1999). Responses to contour features in macaque area V4. *J. Neurophysiol.* **82** 2490–2502.

PEREIRA, F., DETRE, G. and BOTVINICK, M. (2011). Generating text from functional brain images. *Front. Human Neurosci.* **5**.

RODDEY, J. C., GIRISH, B. and MILLER, J. P. (2000). Assessing the performance of neural encoding models in the presence of noise. *J. Comput. Neurosci.* **8** 95–112.

SAHANI, M. and LINDEN, J. F. (2003). How linear are auditory cortical responses? *Adv. Neural Inf. Process. Syst.* **15** 125–132.

SCHEFFÉ, H. (1959). *The Analysis of Variance*. Wiley, New York. MR0116429

SEARLE, S. R. (1971). *Linear Models*. Wiley, New York. MR0293792

SHOHAM, S., PANINSKI, L. M., FELLOWS, M. R., HATSOPOULOS, N. G., DONOGHUE, J. P. and NORMANN, R. A. (2005). Statistical encoding model for a primary motor cortical brain-machine interface. *IEEE Trans. Biomed. Eng.* **52** 1312–1322.

TOWNSEND, E. C. and SEARLE, S. R. (1971). Best quadratic unbiased estimation of variance components from unbalanced data in the 1-way classification. *Biometrics* **27** 643–657.

VU, V. Q., RAVIKUMAR, P., NASELARIS, T., KAY, K. N., GALLANT, J. L. and YU, B. (2011). Encoding and decoding V1 fMRI responses to natural images with sparse nonparametric models. *Ann. Appl. Stat.* **5** 1159–1182. MR2849770

VUL, E., HARRIS, C., WINKIELMAN, P. and PASHLER, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science* **4** 274–290.

WILLIAMS, R. M. (1952). Experimental designs for serially correlated observations. *Biometrika* **39** 151–167. MR0051496

WOOLRICH, M. W., RIPLEY, B. D., BRADY, M. and SMITH, S. M. (2001). Temporal autocorrelation in univariate linear modeling of fMRI data. *Neuroimage* **14** 1370–1386.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
USA
E-MAIL: yuvalben@stanford.edu

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
367 EVANS HALL
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: binyu@stat.berkeley.edu