Submitted to the Annals of Applied Statistics

THE SHUFFLE ESTIMATOR FOR EXPLAINABLE VARIANCE IN FMRI EXPERIMENTS

By Yuval Benjamini[*†] and Bin Yu[*]

Department of Statistics, UC 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 accounted for, 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 humans 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 authors gratefully acknowledge support from NSF grants DMS-0907632, DMS-1107000, SES-0835531 (CDI) and CCF-0939370, and ARO grant W911NF-11-1-0114.

[†]The author gratefully acknowledges support from the NSF VIGRE fellowship.

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 non-invasive, 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 allows neuroscientists to develop quantitative models, *encoding models*(1), 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 yard-stick to contrast competing models regarding the function of the neurons spanned by the voxel (2). 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"(3)), and building brain machine interfaces(4). 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 non-cognitive factors related to blood circulation, or by the measurement apparatus. Regardless of the source, noise cannot be predicted by encoding models that are deterministic functions of the inputs (5). 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 (6). See (7; 8) for examples, and (9) 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 will focus on one such indicator, the proportion of signal variance in the total variance of the measurements. We call this quantity the *explainable variance*[1], because it measures the proportion of variance that

[1]This proportion is known by other names depending on context, such as interclass correlation, effect-size, and pseudo R^2.

can be explained by a deterministic model. The comparison of explainable variance with prediction success (5; 10) informs how much room is left on this 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 (11; 12). 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 validation image set were surprisingly high given the richness of the input class, inspiring many studies of rich stimuli class encoding(13; 14). 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, reliable 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 (15). 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 (16). Ignoring the correlation would greatly bias the signal variance estimation (see Figure 7 below), and would cause us to over-estimate the explainable variance. This over-estimation of signal variance may be a contributing factor to replicability concerns raised in neuroscience (17).

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 (15), or simultaneously using restricted maximum likelihood (18). In both cases, some parametric model for the correlation is needed for the methods to be feasible, for example a fast decay (19). The problem with this type of analysis is that it is sensitive to misspecification of the correlation parameters. In fMRI, the correlation of the noise might vary with the specifics of the preprocessing method in a

way that is not easy to follow or parametrize. As we show in Section 6, if the parameterization for the correlation is too simplistic it might not capture the correlation well and over-estimate the signal, but if it is too flexible the noise might be over-estimated, and the numeric optimizations involved in estimating the correlation might fail to converge.

An alternative way (10; 20) to get around 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 can be displayed within an experimental block(9). 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 (21). 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 with accordance to 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.

This paper is organized as follows. 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.

2. Preliminaries.

2.1. *An FMRI Experiment.* In this section we describe an experiment carried out by the Gallant lab at UC Berkeley (11), in which a human subject viewed natural images while scanned by fMRI [2]. 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 non-invasive 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 (13; 14). 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 (11). This is background for our work, which begins in Section 2.2.

The data of this experiment is composed of the set of natural images, and the fMRI scans recorded for each presentation of an image. The images were sampled from a library of gray-scale photos depicting natural scenes, objects, etc. Two non-overlapping random samples were taken: 1750 images, the training sample, were used for fitting the models; and 120 images, the validation sample, were used for measuring prediction accuracy. Images were sequentially displayed in a randomized order, each image appearing multiple times. BOLD contrast, signaling neural activity, was continuously being recorded

[2]We use data from subject S1 in Kay et al.

by the fMRI machine across the full visual cortex as the subject watched the images. For each voxel, the responses were temporally discretized so that a single value (per voxel) was associated with a single image displayed.

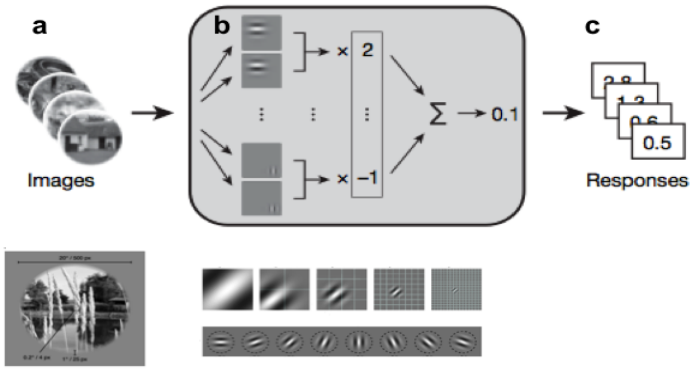

Fig 1: **Encoding models for natural images**. A cartoon depicting the encoding models used by the Gallant lab in the fMRI experiment. Each natural image (**a**) was transformed into a vector of 10409 features, each feature representing the combined energy from two Gabor filters with complementary phases. The 10409 features (**b**) spanned different combinations of spatial frequency, location in the image, and orientations. The features were shared across all voxels, but were linearly combined according to weights fit for each voxel separately, to give a single response per image and voxel (**c**). Images were adapted from Kay *et al.* (2007).

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. For more details on V1 encoding see (22). 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 ($d =$ 10409 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 based on the 1750 training images. Together, the transformation and linear weight vector result in a *prediction rule*, that maps novel images to a real-valued response per voxel. Let $\{I_i\}_{i \leq M}$ be the library of M images from which data was sampled, then denote $f^{(r)} : \{I_i\} \to \mathbb{R}$, the prediction rule corresponding to voxel r for $r = 1, ..., Q$ estimated

based on the training data.

In their paper (11), 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. We can index each displayed image by $t = 1, ..., T$, and let the *schedule* function $h(t) : \{1, ..., T\} \to \{1, ...m\}$ denote the index of the image shown at time slot t. Though a distinct measurement $Y_t^{(r)}$ was extracted for each voxel $r = 1, ..., Q$ at time slot $t = 1, ..., T$, all measurements of the same image were first averaged to reduce noise, obtaining

$$\bar{Y}_j^{(r)} = \underset{t:h(t)=j}{\text{avg}} Y_t^{(r)}, \quad j = 1, ..., m, \ r = 1, ..., Q.$$

Let $\mathbf{s} : \{1, ..., m\} \to \{1, ..., M\}$ be the validation sampling function, indexing the sampled images in the image library (the population), so that $I_{\mathbf{s}(j)}$ is the j'th image in the sample. Consider $\mathbf{s}$ is random due to the design of the experiment, which will allow us to relate the observed accuracy of the sample to the population. A single value per voxel summarizes prediction accuracy

$$Corr^2[f^{(r)}, r] := Corr_j^2(f^{(r)}(I_{\mathbf{s}(j)}), \bar{Y}_j^{(r)}),$$

where $f^{(r)}(I_{\mathbf{s}(j)})$ is the predicted value and $\bar{Y}_j^{(r)}$ the averaged observed value. Note that because $f^{(r)}$ was fitted on an independent training data set, it can be considered as fixed w.r.t. the validation sample. In Figure 2 we show examples of voxels with low, intermediate, and high prediction accuracies, and a histogram of accuracy for all 1250 voxels located within the V1 area. Finally we can drop the superscript r from now because each voxel is analyzed separately.

2.2. *Correlation in the data.* The goal of our work is to separate two factors that determine the accuracy of prediction rules: the adequacy of the chosen features and models to achieve the optimal prediction, and the measurement noise level. Explainable variance represents the accuracy possible with the optimal prediction function, unrestricted by the choice of features and models.

In this paper, we restrict ourselves to Q = 1250 voxels within a functionally homogeneous area, the primary visual cortex (V1). Although there is considerable variation in prediction success between voxels, the voxels should be functionally similar implying that our models should work similarly for these voxels. Thus we postulate that most of the variation in prediction success would come from varied levels of (measurement) noise at different

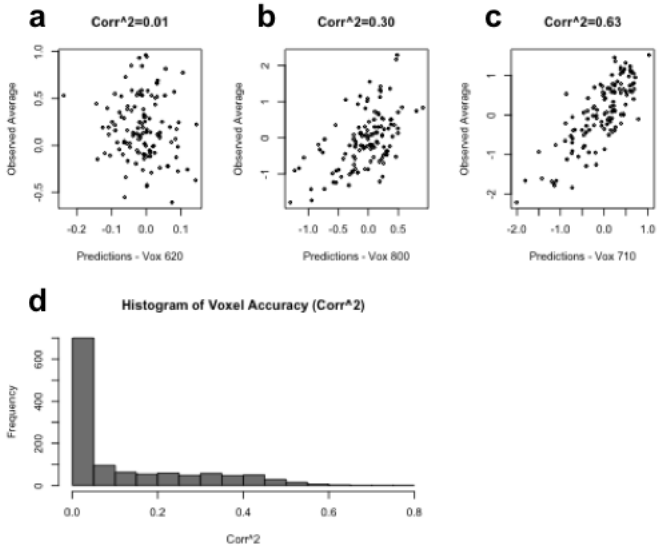

Fig 2: **Prediction accuracy for V1 voxels.** Predicted vs. observed average responses for three voxels in the V1 area, reflecting poor (**a**), medium (**b**), and high (**c**) prediction accuracy. Each point depicts the predicted response (x-axis) and the observed average response (across all repeats) for an image of the validation sample (m=120 images). (**d**) Histogram of prediction accuracy for 1250 V1 voxels.

voxels; in that case good explainable variance estimates should explain most of the voxel-to-voxel variability in prediction accuracy. Once the approach is validated on this setting, explainble variance can be used more broadly, for example to compare the predictability levels of different functional areas.

Since we intend to use the validation sample with replicates to estimate the explainable variances, 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 3a). This data was recorded in 10 separate sessions, so that the subject could rest between sessions; the fMRI was re-calibrated at the beginning of each session. Each session contained all presentations of 12 different images. A pseudo-random integer sequence ordered the repeats within a session[3].

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

When we measure correlation across many voxels, we believe that the design of the experiment induces strong correlation in the data. To see this, in Figure 3 (b-c) we plot the correlation between measurements at different time slots (each time slot is represented by the vector of Q=1250 measurements). This gives us a gross representation of the correlation for 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.

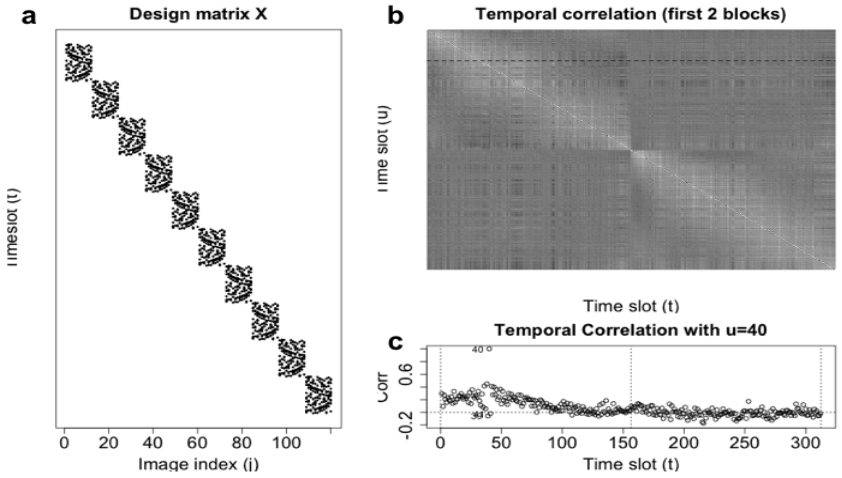

Fig 3: **Data acquisition for the validation data set.** (**a**) The design matrix X recording which image (x-axis) was displayed at each time slot t. Data was recorded in blocks of 12 unique images repeated $n = 13$ times. (**b**) The matrix of temporal correlation across all voxels, $Corr_r(Y_t^{(r)}, Y_u^{(r)})$, for first two blocks $(t, u = 1, ..., 312)$. A cross section $(u = 40)$ of this matrix marked by a dashed line is in (**c**). Strong but non-smooth correlation are found within the blocks, but separate blocks seem uncorrelated. Note that we depict the aggregate correlation of all voxels, but cannot from this infer the noise correlation of any specific voxel.

2.3. *A probability model for the measurements.* We introduce a probabilistic model for the discrete 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 (23). Additivity of noise is considered a good approximation for fMRI event related

designs and is commonly used (24). The random effects model accounts for the generalization of prediction accuracy from the validation sample to the larger population of natural images. In this section the model is carefully developed based on the fMRI experiment, and the quantities of interest for this model are defined in Section 2.4. Section 2.5 introduces algebraic tools that will be used for developing the shuffle estimator.

Let $\{I_i\}_{i=1}^M$ be the set of possible images from which we can sample. We assume each possible I_i image has a fixed mean effect $\mu_i \in \mathbb{R}$ relative to a grand mean on the fMRI response, with population quantities

$$\frac{1}{M}\sum_{i=1}^{M}\mu_i = 0, \quad \frac{1}{M}\sum_{i=1}^{M}\mu_i^2 = \sigma_\mu^2,$$

and we refer to σ_μ^2 as the *signal variance*.

2.3.1. *Random effects based on sampling.* As discussed above, m inputs are sampled randomly from $\{I_i\}_{i=1}^M$ (without replacement), denoted by the random function $\mathbf{s} : \{1, ..., m\} \to \{1, ..., M\}$. It will be useful to discuss the sampled mean effects directly. We denote $A_j = \mu_{\mathbf{s}(j)}$ for $j = 1, ..., m$, or marginally $\mathbb{P}(A_j = \mu_i) = 1/M$ foreach $i = 1, ..., M$, $j = 1, ..., m$. We use $\mathbf{A} = (A_j)_{j=1}^m$ for the random-effect column vector, and

$$\bar{A} := \frac{1}{m}\sum_{j=1}^{m} A_j \qquad s_A^2 := \frac{1}{(m-1)}\sum_{j=1}^{m}(A_j - \bar{A})^2$$

for the sample mean and sample variance of random effects. We assume M is large, which is the case for our fMRI data; so A_j's are effectively independent[4]. Then

$$\mathbb{E}_A[s_A^2] = \sigma_\mu^2$$

where $\mathbb{E}_A$ is the expectation with respect to the sampling.

Images are shown in a long sequence (which may be composed of blocks) so that each image is repeated multiple times. Our analysis will be conditioned on the schedule $h(t) : \{1, ..., T\} \to \{1, ...m\}$ defined earlier, so $h(t)$ is regarded as fixed. $h(t)$ can also be represented by fixed design matrix $X \in \mathbb{R}^{T\times m}$, with $X_{t,j} = 1$ if $h(t) = j$ and 0 otherwise. To illustrate this, consider the following toy example ($T = 5, m = 4$):

$$\begin{array}{l} h(1) = 1 \\ h(2) = 2 \\ h(3) = 3 \\ h(4) = 2 \\ h(5) = 3 \end{array} \iff X \in \mathbb{R}^{T\times m} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}.$$

[4]Both the model and the shuffle estimator can be easily adapted for sampling from a small library of images as well.

In this example, the second sampled image, or $I_{\mathbf{s}(2)}$ is shown at time slots 2 and 4. The (random) mean effect is represented by $A_2 = \mu_{\mathbf{s}(2)}$ in both cases. (These are the first 5 rows in the design matrix used for the example in Figure 4).

We denote $D = XX' \in \mathbb{R}^{T\times T}$ the matrix marking repeats of the same stimulus so for $t, u = 1, ..., T$,

$$D_{t,u} = \begin{cases} 1 & \text{if } h(t) = h(u), \\ 0 & \textit{otherwise.} \end{cases} \tag{2.1}$$

2.3.2. *Noise.* We assume the components of the measurement noise vector $\epsilon = (\epsilon_t)_{t=1}^T$ are independent of the random treatment effects $\{A_j\}_{j=1}^m$, have 0 mean, but may be correlated to capture the slow-changing dynamics of hemodynamics and effects of preprocessing on fMRI signals. Hence,

$$\mathbb{E}_\epsilon[\epsilon_t] = 0; \quad cov(\epsilon_t, \epsilon_u) = \sigma_\epsilon^2 C_{t,u} \;\; C_{t,t} = 1, \tag{2.2}$$

or in matrix notation $cov[\epsilon] = \sigma_\epsilon^2 C$ for $C \in \mathbb{R}^{T\times T}$.

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

$$\mathbf{Y} = X\mathbf{A} + \epsilon \tag{2.3}$$

and for single time slot t

$$Y_t = A_{h(t)} + \epsilon_t,$$

where $\{A_1, ..., A_m\}$ are iid samples from $\{\mu_1, ..., \mu_M\}$.

2.3.4. *Response covariance.* There are two independent sources of randomness in the model:[5] the image sampling modeled by random effects, and the measurement errors, which are of unknown form.

It's easy to see that under our independence assumption of $\mathbf{A}$ and ϵ, the covariance of Y_t and Y_u is composed of the covariance from the sampling and the covariance of the noise,

$$cov_{A,\epsilon}(Y_t, Y_u) = cov_A(A_{h(t)}, A_{h(u)}) + cov_\epsilon(\epsilon_t, \epsilon_u) = \sigma_\mu^2 1_{(h(t)=h(u))} + \sigma_\epsilon^2 C_{t,u}, \tag{2.4}$$

where the first term on the RHS reflects that treatment (random) effects have the same variance σ_μ^2 if they are based on the same input, but are uncorrelated if they are based on different inputs. In matrix form, we get:

$$\mathbb{E}_{A,\epsilon}[\mathbf{Y}] = 0; \quad cov_{A,\epsilon}(Y) = \sigma_\mu^2 D + \sigma_\epsilon^2 C. \tag{2.5}$$

[5]Throughout this paper, it is in fact enough to assume the responses are generated according to $E[Y_t|\mathbf{A}] = A_{h(t)} = \mu_{\mathbf{s}(t)}$ and $cov(Y_t, Y_u|\mathbf{A}) = \sigma_\epsilon^2 C_{t,u}$ without explicit additivity.

2.4. *Explainable variance and variance components.* Explainable variance is a measure of signal-to-noise which generalizes intraclass correlation. When noise is correlated, the ratio of signal variance and total variance no longer equals the correlation between two measurements of the same treatment. Nevertheless, explainable variance is very informative for estimating effect sizes in cases where there are no prediction models, optimizing preprocessing methods, and choosing regularization parameters. This paper was motivated by importance of explainable variance to predictive models, which we will discuss in Section 4.

Recall that $\bar{Y}_j$ are the averaged responses per image ($j = 1, ..., m$ 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

$$MS_{bet} = \frac{1}{m-1}\sum_{j=1}^{m}\left(\bar{Y}_j - \bar{Y}\right)^2. \tag{2.6}$$

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

$$\bar{\sigma}^2_Y = \mathbb{E}_{A,\epsilon}[MS_{bet}]. \tag{2.7}$$

Note that $\bar{\sigma}^2_Y$ 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[6]. Nevertheless, we will loosely use the term *variance* here and later, owing to the parallels between these quantities and the variances in the iid noise case, which are further discussed in Section 4.

$\bar{Y}_j$ is composed of a treatment part (A_j) and average noise part ($\bar{\epsilon}_j$); similarly $\bar{Y}$ is composed of $\bar{A}$ and $\bar{\epsilon}$. By partitioning the MS_{bet} and taking expectations over the sampling and the noise we get

$$\mathbb{E}_{A,\epsilon}[MS_{bet}] = \mathbb{E}_A[\frac{1}{m-1}\sum_{j=1}^{m}\left(A_j - \bar{A}\right)^2] + \mathbb{E}_\epsilon[\frac{1}{m-1}\sum_{j=1}^{m}(\bar{\epsilon}_j - \bar{\epsilon})^2], \tag{2.8}$$

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 $\bar{\sigma}^2_\epsilon$, and get the following decomposition

$$\bar{\sigma}^2_Y = \sigma^2_\mu + \bar{\sigma}^2_\epsilon \tag{2.9}$$

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

[6]In practice, this is true for the individual measurements Y_t as well. We chose $C_{t,t} = 1$ for illustration reasons.

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

$$\omega^2 = \sigma_\mu^2/\bar{\sigma}_{\bar{Y}}^2.$$

Explainable variance measures the proportion of variance due to treatment in the averaged responses, hence is an alternative to signal-to-noise measures.

Note that of the two expressions in ω^2, $\bar{\sigma}_{\bar{Y}}^2$ can be naturally estimated from the sample, while σ_μ^2 requires more work. To estimate σ_μ^2 the signal and noise need to be separated. As we see in the next section, one way to separate them is based on their different covariance structure.

2.5. *Quadratic contrasts.* In this subsection we derive MS_{bet} as a quadratic contrast of the full data vector $\mathbf{Y}$. This would highlight the relation between $\bar{\sigma}_{\bar{Y}}^2$ or $\bar{\sigma}_\epsilon^2$ with both the design $D = XX'$ and the measurement correlations C, and would produce algebraic descriptions used in Section 3. These are simple extensions of classical treatment of variance components (25).

Denote $B := XX'/n$, an $\mathbb{R}^{T\times T}$ scaled version of D, with

$$B_{t,u} = \begin{cases} \frac{1}{n} & \text{if } h(t) = h(u), \\ 0 & \textit{otherwise}. \end{cases} \tag{2.10}$$

B is an averaging matrix, because when it multiplies $\mathbf{Y}$, each element in the vector is replaced with the treatment average, that is

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

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

$$MS_{bet} = \frac{1}{(m-1)n}\|(B-G)\mathbf{Y}\|^2. \tag{2.12}$$

or more generally as a function of any input vector
$MS_{bet}(\cdot) := \frac{1}{(m-1)n}\|(B-G)(\cdot)\|^2$.

The following proposition outlines the relation between total variance, the design and the correlation of the noise.

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

$$\bar{\sigma}_{\bar{Y}}^2 = \sigma_\mu^2 + \frac{1}{(m-1)n}\sigma_\epsilon^2 tr\left((B-G)C\right) \tag{2.13}$$

PROOF.

$$\begin{aligned}
\bar{\sigma}_Y^2 = \mathbb{E}_{A,\epsilon}[MS_{bet}(\mathbf{Y})] &= \tfrac{1}{(m-1)n}\mathbb{E}_{A,\epsilon}[tr\left((B-G)(\mathbf{Y}'\mathbf{Y})(B-G)\right)] \\
&= \tfrac{1}{(m-1)n} tr\left((B-G)cov_{A,\epsilon}(\mathbf{Y})(B-G)\right) \\
&= \tfrac{1}{(m-1)n} tr\left((B-G)cov_{A}(\mathbf{Y})\right) + \tfrac{1}{(m-1)n} tr\left((B-G)cov_{\epsilon}(\mathbf{Y})\right) \\
&= \tfrac{1}{(m-1)n} tr\left((B-G)(n\sigma_\mu^2 B)\right) + \tfrac{1}{(m-1)n} tr\left((B-G)\sigma_\epsilon^2 C\right) \\
&= \tfrac{1}{(m-1)}\sigma_\mu^2 tr(B-G) + \tfrac{1}{(m-1)n}\sigma_\epsilon^2 tr((B-G)C) \\
&= \sigma_\mu^2 + \tfrac{1}{(m-1)n}\sigma_\epsilon^2 tr\left((B-G)C\right).
\end{aligned}$$

□

From (2.9,2.13) we get an exact expression for the noise level

$$\bar{\sigma}_\epsilon^2 = \tfrac{1}{(m-1)n}\sigma_\epsilon^2 tr\left((B-G)C\right). \tag{2.14}$$

Obviously, $\bar{\sigma}_\epsilon^2$ scales with the noise variance of the individual measurements σ_ϵ^2. Moreover, $\bar{\sigma}_\epsilon^2$ depends on the relation between the design and the measurement correlation C. Note that if there are no correlation within repeats, then $tr\left((B-G)C\right) = (m-1)\sigma_\epsilon^2$ and $\bar{\sigma}_\epsilon^2 = \sigma_\epsilon^2/n$. In that case $\bar{\sigma}_Y^2 = \sigma_\mu^2 + \sigma_\epsilon^2/n$, and by plugging in an estimator of σ_ϵ^2, we can directly estimate $\bar{\sigma}_\epsilon^2$ and σ_μ^2. This gives us an estimator for ω^2 if noise is uncorrelated

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

for F the standard F statistic. This is 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 level of the noise and inflates the explainable variance. In the next section we introduce the shuffle estimators which can deal with non-0 correlations in the 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 level, and for the explainable variance. As in (2.9), $\bar{\sigma}_Y^2 = \sigma_\mu^2 + \bar{\sigma}_\epsilon^2$, but the noise variance $\bar{\sigma}_\epsilon^2$ is a function of the (unknown) measurement correlation matrix C. Using shuffle estimators we can estimate σ_μ^2 and $\bar{\sigma}_\epsilon^2$ without having to estimate the full C 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 with accordance to symmetries that are based on 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 variance due to signal will be reduced in the shuffled data. We derive an unbiased estimator for signal variance σ_μ^2 based on this reduction in variance, and use the plug-in estimators for $\bar{\sigma}_\epsilon^2$ and ω^2.

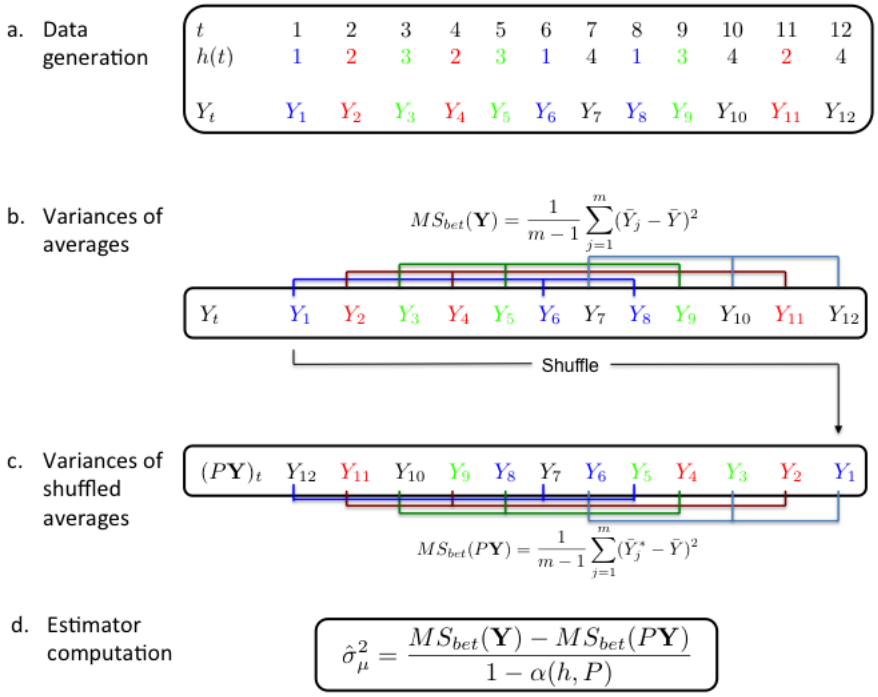

Fig 4: **Cartoon of the shuffle estimator.** (**a**) Data is generated according to schedule $h(t)$, with each color representing repeats of a different image. (**b**) Repeats of each image are averaged together and the sample variance is computed on these averages. (**c**) Data is shuffled by P, in 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 $\alpha(h,P) = \frac{1}{m-1} tr\left((B-G)PBP'\right)$. (**d**) The shuffle estimator for signal variance is the difference between the two sample variances, after correction of $1-\alpha(h,P)$.

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

Recall (2.14),

$$\bar{\sigma}^2_\epsilon = \frac{1}{(m-1)n} tr((B-G)(\sigma^2_\epsilon C)),$$

where $\sigma^2_\epsilon C = cov_\epsilon[\mathbf{Y}]$ as before. Let $P \in \mathbb{R}^{T\times T}$ be a permutation matrix. Then

DEFINITION 2. *P is noise conserving w.r.t h, if*

$$tr\left((B-G)P\sigma^2_\epsilon CP'\right) = tr\left((B-G)\sigma^2_\epsilon C\right). \tag{3.1}$$

Equivalently,

$$tr\left((B-G)cov_\epsilon[P\cdot\mathbf{Y}]\right) = tr\left((B-G)\sigma^2_\epsilon C\right).$$

Although we define the noise conserving property based on the covariance, replacing the covariance with the correlation matrix C would not change the permutation class.

Noise conservation is a property that depends on the interplay between the design B and the noise covariance C. Let us take a look at important cases.

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

$$h(t) = h(u) \Rightarrow h(g_P(t)) = h(g_P(u)), \quad \forall t, u. \tag{3.2}$$

It is easy to show that for trivial P, $MS_{bet}(P\mathbf{Y}) = MS_{bet}(\mathbf{Y})$.

3.1.2. *Noise conserving permutations based on symmetries of C.* A useful class of non-trivial noise conserving permutations is the class of symmetries in the correlation matrix C: a symmetry of C is a permutation P such that $PC = C$. If P is a symmetry of C, 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 $C = I$ where all responses are exchangeable. Hence any permutation is noise-conserving.

2. **Stationary time series** Neuroscience data is typically 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 C is a Toeplitz matrix parameterized by $\{\rho_d\}_{d=0}^{T-1}$, the set of correlation values $C_{t,u} = \rho_d, \quad \text{where } d = |t-u|$. 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.
   Another family of noise conserving permutations are the shift operator $(P\mathbf{Y})_t = (\mathbf{Y})_{t+k}$ (up to edge effects).

3. **Independent blocks** 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 to let $C_{t,u} = \zeta(\beta(t), \beta(u))$ 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 session structure, meaning

$$\beta(t) = \beta(u) \Rightarrow \beta(g_P(t)) = \beta(g_P(u)) \tag{3.3}$$

   would be noise-conserving w.r.t. any h.

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 mean-equation with two parameters: σ_μ^2 and $\bar{\sigma}_\epsilon^2$. The coefficient for σ_μ^2

is $\alpha(h,P) := \frac{1}{m-1} tr\left((B-G)(PBP')\right)$, which only depends on parameters known to the scientist.

LEMMA 4. *If P is a noise-conserving permutation for $\mathbf{Y}$, then*

1. $\mathbb{E}_{A,\epsilon}[MS_{bet}(P\mathbf{Y})] = \alpha(h,P)\sigma_\mu^2 + \bar{\sigma}_\epsilon^2$.
2. $\alpha(h,P) \leq 1$, *and the inequality is strict iff P is non-trivial.*

PROOF. 1. Using similar algebra as in Proposition 1, the expectation $\mathbb{E}_{A,\epsilon}[MS_{bet}(P\mathbf{Y})]$ can be partitioned into a term depending on the sampling covariance $cov_A(P\mathbf{Y})$ and a term depending on the noise covariance $cov_\epsilon(P\mathbf{Y})$. Since P is noise-conserving, for the noise term:

$$cov_\epsilon(P\mathbf{Y}) = \bar{\sigma}_\epsilon^2.$$

As for the sampling:

$$cov_A(P\mathbf{Y}) = P cov_A(\mathbf{Y})P' = \sigma_\mu^2 P(nB)P'.$$

Hence,

$$\frac{1}{(m-1)n}\sigma_\mu^2 tr\left((B-G)(P(nB)P')\right) = \alpha(h,P)\sigma_\mu^2.$$

2. In Proposition 1 we saw that the sampling component for the unpermuted vector $cov_A(\mathbf{Y})$ is σ_μ^2. Hence for P the identity matrix $I \in \mathbb{R}^{T\times T}$ we have $\alpha(I,h) = 1$. For all other P's, note that the global mean term (G) is unaffected by the permutation ($PG = G$) or the averaging ($BG = G$), so it remains unchanged.

   From the Cauchy-Schwartz inequality,

$$tr\left(B(PBP')\right) \leq tr\left(BB\right) = tr(B)$$

   as P is unitary and B a projection Recall that P is trivial if P reorders measurements within categories and renames categories. It is easy to check that $PBP' = B$ iff P is trivial. For trivial P's, we again get equations similar to Proposition 1, so $\alpha(P,h) = 1$.
   For any **non-trivial** permutation $B \neq PBP'$, in which case the CS-inequality is strict resulting in $\alpha(h,P) < 1$.

□

As can be seen in the proof, α depends only on B and P which are both known:

$$\alpha(h,P) = \frac{1}{m-1} tr\left((B-G)(PBP')\right) \tag{3.4}$$

It reflects how well P "mixes" the treatments; the greater the mix, the smaller α.

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

$$\begin{cases} \mathbb{E}_{A,\epsilon}[MS_{bet}(\mathbf{Y})] = \sigma_\mu^2 + \bar{\sigma}_\epsilon^2 \\ \mathbb{E}_{A,\epsilon}[MS_{bet}(P\mathbf{Y})] = \alpha(h, P)\sigma_\mu^2 + \bar{\sigma}_\epsilon^2 \end{cases} \tag{3.5}$$

can be solved.

This leads to our main point, defining the shuffle estimator for σ_μ^2 based on (3.5), and the estimator for $\bar{\sigma}_\epsilon^2$ based on its complement to MS_{bet}:

DEFINITION 5. *Let P be a non-trivial noise conserving permutation for $\mathbf{Y}$. Then the shuffle estimators for the signal variance ($\hat{\sigma}_\mu^2$) and noise level ($\hat{\bar{\sigma}}_\epsilon^2$) are*

$$\hat{\sigma}_\mu^2 := \frac{MS_{bet}(\mathbf{Y}) - MS_{bet}(P\mathbf{Y})}{1 - \alpha(h, P)}, \tag{3.6}$$

$$\hat{\bar{\sigma}}_\epsilon^2 := MS_{bet}(\mathbf{Y}) - \hat{\sigma}_\mu^2. \tag{3.7}$$

LEMMA 6. *If P is a non-trivial noise-conserving permutation for $\mathbf{Y}$, then*

1. $\mathbb{E}[\hat{\sigma}_\mu^2] = \sigma_\mu^2$
2. $\mathbb{E}[\hat{\bar{\sigma}}_\epsilon^2] = \bar{\sigma}_\epsilon^2$

In practice, we prefer the restricted shuffle estimators

$$(\hat{\sigma}_\mu^2)_+ = \max\{\hat{\sigma}_\mu^2, 0\} \quad (\hat{\bar{\sigma}}_\epsilon^2)_+ = \min\{MS_{bet}(\mathbf{Y}), \hat{\bar{\sigma}}_\epsilon^2\} \tag{3.8}$$

which have lower MSEs but are no longer unbiased.

Finally, we would like to estimate the explainable variance $\omega^2 = \sigma_\mu^2/\bar{\sigma}_Y^2$. We use the plug in estimator,

$$\hat{\omega}^2 = \frac{(\hat{\sigma}_\mu^2)_+}{MS_{bet}(\mathbf{Y})}.$$

Note that $\hat{\omega}^2$ is restricted between 0 and 1.

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 *et al.* (2000) (5) show that explainable variance upper bounds the accuracy of prediction on the sample when noise is iid. We generalize their results for arbitrary noise correlation and account for generalization from sample to population[7]. As shown in Lemma 7, the noise level $\bar{\sigma}_\epsilon^2$ 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.

First let us recall the setup. Let f be a prediction function that predicts a real-valued response to any possible image (out of a population of M):

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

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 validation set. Recall $\mathbf{s} : \{1, ..., m\} \to \{1, ..., M\}$ is the random sampling function. For the j'th sampled image, the predicted response $f(I_{\mathbf{s}(j)})$ is compared with the average observed response for that image $\bar{Y}_j$. We consider two common accuracy measures: mean squared prediction error ($MSPE[f]$) and the squared correlation ($Corr^2[f]$), defined

(4.2)

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

(4.3)

$$Corr^2[f] := Corr_j^2(f(I_{\mathbf{s}(j)}), \bar{Y}_j) = \frac{\left(\frac{1}{m-1} \sum_{j=1}^{m} (f(I_{\mathbf{s}(j)}) - \bar{f}_{\mathbf{s}})(\bar{Y}_j - \bar{Y}) \right)^2}{\frac{1}{m-1} \sum_{j=1}^{m} (f(I_{\mathbf{s}(j)}) - \bar{f}_{\mathbf{s}})^2 \sum_{j=1}^{m} (\bar{Y}_j - \bar{Y})^2},$$

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; details can be found in the appendix.

LEMMA 7. *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,*

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

(a) $f^* = \arg\min_f \mathbb{E}_{A,\epsilon}[MSPE[f]]$;

(b) $\bar{\sigma}^2_\epsilon = \mathbb{E}_{A,\epsilon}[MSPE[f^*]] = \min_f \mathbb{E}_{A,\epsilon}[MSPE[f]]$;

(c) $\omega^2 \approx \mathbb{E}_{A,\epsilon}[Corr^2[f^*]]$ *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: $\bar{\sigma}^2_\epsilon$ 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 yard stick 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^2_μ. 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 non-zero σ^2_μ, 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, $k = 1, ..., B$ blocks), and an iid Gaussian term (e_t, $t = 1, ..., T$)

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

A_j, b_k and e_t are sampled from centered normal distributions with variances $(\sigma^2_\mu, \sigma^2_b, \sigma^2_e)$. We used $\sigma^2_b = 0.5, \sigma^2_e = 0.7$, and varied the signal level $\sigma^2_\mu = 0, 0.1, ..., 0.9$. We used $m = 120$, $n = 15$, with all presentations of every 5 stimuli composing a blocks ($B = 20$ 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 ($\alpha(h, P) = 0.115$). The results are shown in Figure 5 (a).

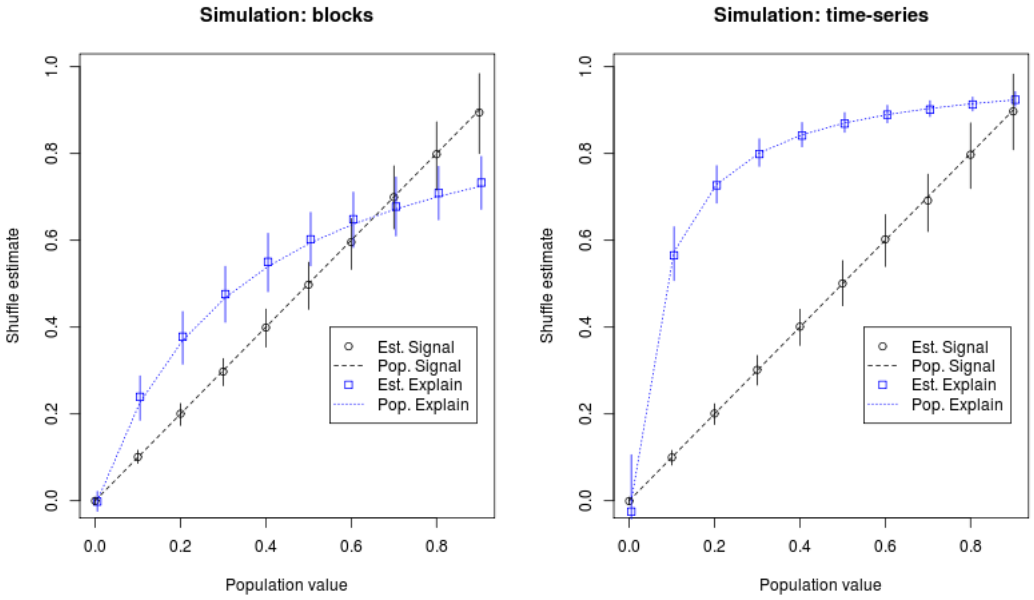

Fig 5: **Simulations for the block and time-series** (**a**) Simulation results comparing shuffle estimates for signal variance σ_μ^2 (black) and explainable variance ω^2 (blue) to the true population values (dashed line). Noise correlation followed an independent block structure: noise within blocks was correlated, and between blocks was independent. The x-axis represents the true signal variance σ_μ^2 of the data, and the y-axis marks the average of the estimates and [0.25,0.75] quantile range. (**b**) Similar plot for data generated under a stationary time-series model.

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 C, where C is an exponentially decaying covariance with a nugget,

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

Then $Y = A_{(t)} + e_t$ with the random effects $A_{(t)}$ sampled from $\mathcal{N}(0, \sigma_\mu^2)$ for $\sigma_\mu^2 = 0, 0.1, ...0.9$. 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

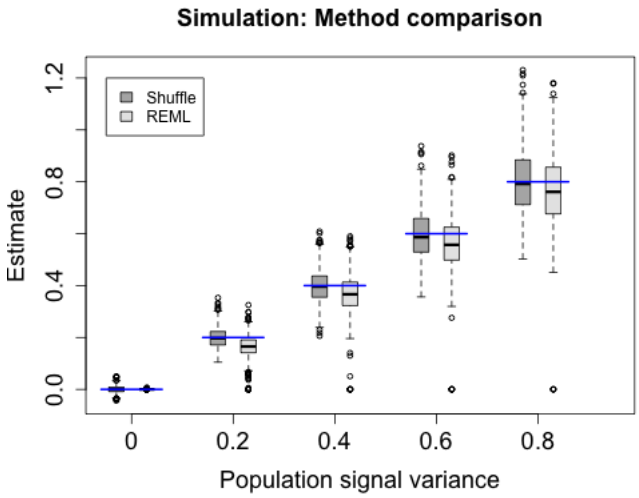

Fig 6: **Comparison of methods on simulation**. Each pair of box-plots represents the estimated signal variance $\hat{\sigma}^2_\mu$ using the shuffle estimator (dark gray) and REML (light gray) for 1000 simulations. The blue horizontal line represents the true value of σ^2_μ. The REML estimator assumes 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.

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 $\alpha(h, P) = 0.064$.

5.3. *Comparison to REML.* In Figure 6 we used time-series data to compare σ^2_μ estimates based on the shuffle estimators to those obtained by an REML estimator with the correct parametrization for the noise correlation matrix. We used `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 $\sigma^2_\mu = 0, 0.2, 0.4, 0.6, 0.8$, 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 are slightly better in both bias and in variance.

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 postulate that most of the variation in prediction accuracy would come from varied levels of noise at different voxels; in that case good explainable variance estimates should explain most of the voxel-to-voxel variability in prediction accuracy.

Prediction accuracy values for these 1250 voxels are compared to explainable variance estimates for each voxel, as generated by the shuffle estimator. 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.* Several methods are compared for estimating the explainable variance ($\omega^2 = \sigma^2_\mu/\bar{\sigma}^2_Y$). The methods differ in how σ^2_μ is estimated; all methods use the sample averages variance $MS_{bet}(\mathbf{Y})$ for $\bar{\sigma}^2_Y$, and plug in the two estimates into ω^2. We estimate ω^2 separately for each voxel ($r = 1, ..., 1250$). The methods we compare are

1. The shuffle estimators estimator. We assume time-series stationarity within each block, and independence between the blocks, so choose a 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. $\alpha(h, P) = 0.17$. We use the restricted estimator
$$\hat{(\sigma^2_\mu)}_+ = \max\left\{\frac{MS_{bet}(\mathbf{Y}) - MS_{bet}(P\mathbf{Y})}{1-\alpha(h,P)}, 0\right\}$$
for signal variance, and the explainable variance is obtained by plugging the estimate of σ^2_μ into $\hat{\omega}^2 = \hat{\sigma^2_\mu}/MS_{bet}$.
2. An estimator ($\tilde{\omega}^2$) unadjusted for correlation. We use the mean-square within ($MS_{wit} = \frac{1}{(m-1)n}\sum_{j=1}^{m}\sum_{t:h(t)=j}(Y_t - \bar{Y}_j)^2$) contrast to estimate the noise variance σ^2_ϵ, scale by $1/n$ to estimate the noise level $\bar{\sigma}^2_\epsilon$,

and remove the scaled estimate from MS_{bet}, $\tilde{\sigma}^2_\mu = MS_{bet} - MS_{wit}/n$. Explainable variance is obtained by plug in estimator $\tilde{\omega}^2 = \tilde{\sigma}^2_\mu / MS_{bet}$.

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 $C_{t,t+d} = \lambda_2 exp(-d/\lambda_1) + 1_{(d=0)}(1-\lambda_2)$ where the rate of decay λ_1 and nugget effect λ_2 where 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 $\epsilon_t = \eta_t + \sum_{k=1}^{3} a_k \epsilon_{t-k}$. This models allows for non-monotone correlations.

   We use the `nlme` package in R to estimate the signal variance of this model using restricted maximum likelihood (18) (REML), and use the plug-in estimator for the explainable variance.

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 $Corr^2[f]$ and $\hat{\omega}^2$ is very linear ($r = 0.9$). 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. Although there is still room for improving prediction for some voxels, the Gabor models are not far from performing optimally on these recordings.

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 high variance.

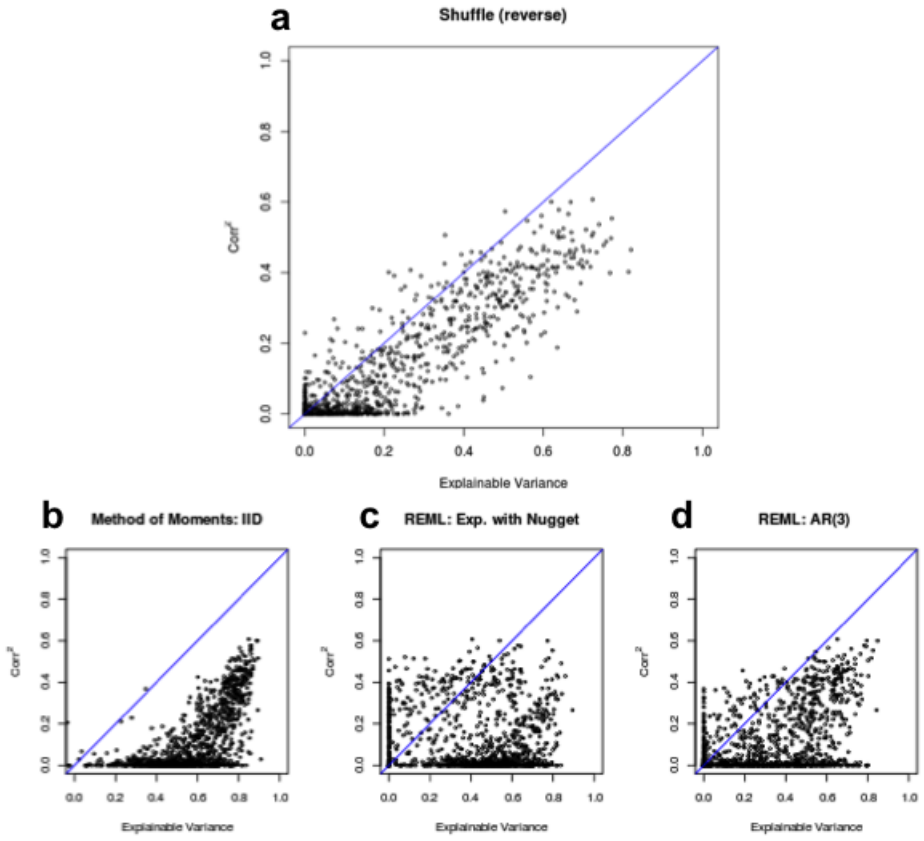

Fig 7: **Optimal vs. observed prediction accuracy.** The estimated optimal prediction is compared with observed prediction ($Corr^2$), each point representing a V1 response. The optimal prediction estimated by (**a**) shuffle estimators accounting for stationary noise distributions; (**b**) Method of 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.

7. Discussion. We have presented the shuffle estimator, a resampling-based estimator for the explainable variance in a random-effects additive model with auto-correlated 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 un-

biased 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 indicators of the explainable signal variance at each of the voxels. Shuffle estimates of the explainable variance explained most of the variation 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 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.

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

8. Acknowledgements. We are grateful to An Vu and members of Jack Gallant's laboratory for access to the data and models and for helpful discussions about the method, and to Terry Speed and Philip Stark for suggestions that greatly improved this paper.

References.

[1] Dayan, P., Abbott, L., and Abbott, L. (2001) Theoretical neuroscience: Computational and mathematical modeling of neural systems, The MIT Press, .

[2] Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., Gallant, J. L., and Rust, N. C. (2005) *The Journal of Neuroscience* **25(46)**, 10577 –10597.

[3] Nishimoto, S., Vu, A., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J. (2011) *Current Biology*.

[4] Shoham, S., Paninski, L. M., Fellows, M. R., Hatsopoulos, N. G., Donoghue, J. P., and Normann, R. A. (2005) *IEEE Transactions on Biomedical Engineering* **52(7)**, 1312–1322.

[5] Roddey, J. C., Girish, B., and Miller, J. P. (2000) *Journal of Computational Neuroscience* **8(2)**, 95–112 PMID: 10798596.

[6] Josephs, O., Turner, R., and Friston, K. (1997) *Human brain mapping* **5(4)**, 243–248.

[7] Pasupathy, A. and Connor, C. (1999) *Journal of Neurophysiology* **82(5)**, 2490.

[8] Haefner, R. and Cumming, B. (2008) *Advances in neural information processing systems* pp. 1–8.

[9] Huettel, S. (2011) *NeuroImage* **(0)**, –.

[10] Sahani, M. and Linden, J. F. (2003) *Advances in neural information processing systems* pp. 125–132.

[11] Kay, K. N., Naselaris, T., Prenger, R. J., and Gallant, J. L. (2008) *Nature* **452(7185)**, 352–355.

[12] Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., and Gallant, J. L. (2009) *Neuron* **63(6)**, 902–915.

[13] Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., Knight, R. T., and Chang, E. F. (2012) *PLoS Biol* **10(1)**, e1001251.

[14] Pereira, F., Detre, G., and Botvinick, M. (2011) *Frontiers in Human Neuroscience* **5(00072)**, 0.

[15] Scheffe, H. (1959) The analysis of variance, volume **72**, Wiley-Interscience, .

[16] Fox, M. D. and Raichle, M. E. (2007) *Nat Rev Neurosci* **8(9)**, 700–711.

[17] Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009) *Perspectives on Psychological Science* **4(3)**, 274–290.

[18] Laird, N., Lange, N., and Stram, D. (1987) *Journal of the American Statistical Association* **82(397)**, 97–105.

[19] Woolrich, M., Ripley, B., Brady, M., and Smith, S. (2001) *Neuroimage* **14(6)**, 1370–1386.

[20] Hsu, A., Borst, A., and Theunissen, F. E. (2004) *Network: Computation in Neural Systems* **15(2)**, 91–109.

[21] Edgington, E. and Onghena, P. (2007) Randomization tests, volume **191**, CRC Press, .

[22] Vu, V., Ravikumar, P., Naselaris, T., Kay, K., Gallant, J., and Yu, B. (2011) *The Annals of Applied Statistics* **5(2B)**, 1159–1182.

[23] Williams, R. M. (1952) *Biometrika* **39(1/2)**, 151–167.

[24] Buracas, G. T. and Boynton, G. M. (2002) *NeuroImage* **16(3, Part A)**, 801 – 813.

[25] Townsend, E. C. and Searle, S. R. (1971) *Biometrics* pp. 643–657.

9. Appendix.

9.1. *Lemma* (7). 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 experimental conditions and model described above,

(a) $f^* = \arg\min_f \mathbb{E}_{A,\epsilon}[MSPE[f]]$;

(b) $\bar{\sigma}_\epsilon^2 = \mathbb{E}_{A,\epsilon}[MSPE[f^*]] = \min_f \mathbb{E}_{A,\epsilon}[MSPE[f]]$;

(c) $\omega^2 \approx \mathbb{E}_{A,\epsilon}[Corr^2[f^*]]$ with a bias term smaller than $\frac{1}{m-1}$.

PROOF. (a) + (b)
First, for any image in the sample, we compare the prediction with the expected average given the sampling,

$$MSPE[f] = \frac{1}{m-1}\sum_j^m \left((f(I_{\mathbf{s}(j)}) - \mathbb{E}_\epsilon[\bar{Y}_j]) + (\mathbb{E}_\epsilon[\bar{Y}_j] - \bar{Y}_j)\right)^2. \tag{9.1}$$

From our model, $\mathbb{E}_\epsilon[\bar{Y}_j] = A_j$. Substituting this into 9.1 and taking expectation over the noise, we get

$$\mathbb{E}_\epsilon[MSPE[f]] = \frac{1}{m-1}[\mathbb{E}_\epsilon \sum_j^m \left(f(I_{\mathbf{s}(j)}) - A_j\right)^2 + \mathbb{E}_\epsilon \sum_j^m \left(A_j - \bar{Y}_j\right)^2 + \mathbb{E}_\epsilon \sum_j^m \left(f(I_s(j)) - A_j\right)\left(A_j - \bar{Y}_j\right)].$$

Recall that $\bar{Y}_j - A_j$ is $\bar{\epsilon}_j$ for each j, with $\mathbb{E}_\epsilon[\bar{\epsilon}_j] = 0$ and a sample variance $\bar{\sigma}_\epsilon^2$. Therefore

$$\mathbb{E}_\epsilon[MSPE[f]] = \frac{1}{m-1}[\sum_j^m \left(f(I_{\mathbf{s}(j)} - A_j\right)^2] + \bar{\sigma}_\epsilon^2. \tag{9.2}$$

By also taking an expectation over the sampling

$$\mathbb{E}[MSPE[f]] = \mathbb{E}_A\mathbb{E}_\epsilon[MSPE[f]] = \frac{m}{m-1}\frac{1}{M}\sum_i^M[(f(I_i) - \mu_i)^2] + \bar{\sigma}_\epsilon^2. \tag{9.3}$$

□

PROOF. (c)
Since the optimal f^* maps each image I_i to its mean-effect μ_i, for the sampled image it maps the random effect:

$$f^*(I_{\mathbf{s}(j)}) = \mu_{\mathbf{s}(j)} = A_j.$$

Hence $Corr^2[f^*] = Corr_j^2(A_j, \bar{Y}_j)$, or in extended form

$$Corr_j^2(A_j, \bar{Y}_j) = \frac{\left(\frac{1}{m-1}\sum_{j=1}^{m}(A_j - \bar{A})(\bar{Y}_j - \bar{Y})\right)^2}{\left(\frac{1}{m-1}\sum_{j=1}^{m}(A_j - \bar{A})^2\right)\left(\frac{1}{m-1}\sum_{j=1}^{m}(\bar{Y}_j - \bar{Y})^2\right)}. \tag{9.4}$$

Recall that $\bar{Y}_j = A_j + \bar{\epsilon}_j$. Equation (9.4) becomes

$$\frac{\left(\frac{1}{m-1}\sum_{j=1}^{m}\left[(A_j - \bar{A})(\bar{A}_j - \bar{A}) + (A_j - \bar{A})(\bar{\epsilon_j} - \bar{\epsilon})\right]\right)^2}{\left(\frac{1}{m-1}\sum_{j=1}^{m}(A_j - \bar{A})^2\right)\left(\frac{1}{m-1}\sum_{j=1}^{m}(\bar{Y}_j - \bar{Y})^2\right)}. \tag{9.5}$$

Let

$$s_A^2 = \frac{1}{m-1}\sum_{j=1}^{m}(A_j - \bar{A})^2; \quad \bar{s}_\epsilon^2 = \frac{1}{m-1}\sum_{j=1}^{m}(A_j - \bar{A})^2; \quad MS_{bet} = \frac{1}{m-1}\sum_{j=1}^{m}(\bar{Y}_j - \bar{Y})^2;$$

represent the sample variance of the treatment effects, averaged noise, and average measurements respectively. Moreover, let

$$r = \frac{\sum_{j=1}^{m}(A_j - \bar{A})(\bar{\epsilon}_j - \bar{\epsilon})}{(m-1)\, s_A \cdot \bar{s}_\epsilon}$$

be the empirical correlation of the treatment effects and the averaged noise.

Substituting into Equation 9.5 results in:

$$\frac{\left(s_A^2 + s_A\, \bar{s}_\epsilon\, r\right)^2}{s_A^2\, MS_{bet}} = \frac{s_A^2 + 2 s_A\, \bar{s}_\epsilon\, r + \bar{s}_\epsilon^2\, r^2}{MS_{bet}}.$$

By taking expectations over A and ϵ and approximating the expectations of the ratio with the ratio of the expectations, we get:

$$\begin{aligned}
\mathbb{E}_{A,\epsilon}[Corr^2[f^*]] =& \mathbb{E}_{A,\epsilon}\left[\frac{s_A^2 + 2 s_A\, \bar{s}_\epsilon\, r + \bar{s}_\epsilon^2\, r^2}{MS_{bet}}\right] \\
=& \mathbb{E}_{A,\epsilon}\left[\frac{s_A^2}{MS_{bet}}\right] + \mathbb{E}_{A,\epsilon}\left[\frac{2 s_A\, \bar{s}_\epsilon\, r + \bar{s}_\epsilon^2\, r^2}{MS_{bet}}\right] \\
\approx& \frac{\mathbb{E}_{A,\epsilon}\left[s_A^2\right]}{\mathbb{E}_{A,\epsilon}\left[MS_{bet}\right]} + \frac{\mathbb{E}_{A,\epsilon}\left[2 s_A\, \bar{s}_\epsilon\, r + \bar{s}_\epsilon^2\, r^2\right]}{\mathbb{E}_{A,\epsilon}\left[MS_{bet}\right]} \\
=& \omega^2 + \frac{\mathbb{E}_{A,\epsilon}\left[2 s_A\, \bar{s}_\epsilon\, r\right] + \bar{s}_\epsilon^2\, r^2\right]}{\bar{\sigma}_Y^2}
\end{aligned}$$

Since the mean effects A_j's and averaged noise $\bar{\epsilon}_j$'s are independent, $\mathbb{E}[r] = 0$. Hence

$$\mathbb{E}_{A,\epsilon}[Corr^2[f^*]] \approx \omega^2 + \bar{\sigma}_\epsilon^2 \frac{\mathbb{E}\left[r^2\right]}{\bar{\sigma}_Y^2}.$$

Under mild conditions and m large enough $\sqrt{m-1}\, r_{A,\bar{\epsilon}} \approx \mathcal{N}(0,1)$. We get a bias on the order of $\frac{\bar{\sigma}_\epsilon^2}{\bar{\sigma}_Y^2}\frac{1}{m-1} < \frac{1}{m-1}$. Note that unless $\sigma_\mu^2/\bar{\sigma}_Y^2 \approx 0$, the bias is negligible compared to the deviation of $\frac{2 s_A\, \bar{s}_\epsilon\, r}{MS_{bet}}$ which is of order $\frac{1}{\sqrt{m-1}}$.

□

Yuval Benjamini
Department of Statistics
E-mail: yuvalb@stat.berkeley.edu

Bin Yu
Department of Statistics
E-mail: binyu@stat.berkeley.edu