Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Optimizing multivariate pattern classification in rapid event-related designs

Daniel A. Stehr [a,*], Javier O. Garcia [b], John A. Pyles [c], Emily D. Grossman [a]

[a] *University of California, Irvine, United States of America*
[b] *US DEVCOM Army Research Laboratory, United States of America*
[c] *University of Washington, Seattle, Washington, United States of America*

ARTICLE INFO

Dataset link: https://doi.org/10.6084/m9.figshare.13708654.v1

Keywords:
functional magnetic resonance imaging
multivariate pattern analysis
pattern classification
pattern decoding
rapid event-related design

ABSTRACT

Background: Multivariate pattern analysis (MVPA or pattern decoding) has attracted considerable attention as a sensitive analytic tool for investigations using functional magnetic resonance imaging (fMRI) data. With the introduction of MVPA, however, has come a proliferation of methodological choices confronting the researcher, with few studies to date offering guidance from the vantage point of controlled datasets detached from specific experimental hypotheses.
New method: We investigated the impact of four data processing steps on support vector machine (SVM) classification performance aimed at maximizing information capture in the presence of common noise sources. The four techniques included: trial averaging (classifying on separate trial estimates versus condition-based averages), within-run mean centering (centering the data or not), method of cost selection (using a fixed or tuned cost value), and motion-related denoising approach (comparing no denoising versus a variety of nuisance regressions capturing motion-related reference signals). The impact of these approaches was evaluated on real fMRI data from two control ROIs, as well as on simulated pattern data constructed with carefully controlled voxel- and trial-level noise components.
Results: We find significant improvements in classification performance across both real and simulated datasets with run-wise trial averaging and mean centering. When averaging trials within conditions of each run, we note a simultaneous increase in the between-subject variability of SVM classification accuracies which we attribute to the reduced size of the test set used to assess the classifier's prediction error. Therefore, we propose a hybrid technique whereby randomly sampled subsets of trials are averaged per run and demonstrate that it helps mitigate the tradeoff between improving signal-to-noise ratio by averaging and losing exemplars in the test set.
Comparison with existing methods: Though a handful of empirical studies have employed run-based trial averaging, mean centering, or their combination, such studies have done so without theoretical justification or rigorous testing using control ROIs.
Conclusions: Therefore, we intend this study to serve as a practical guide for researchers wishing to optimize pattern decoding without risk of introducing spurious results.

1. Introduction

Multivariate pattern analysis (MVPA or pattern decoding) has become an increasingly preferred analytical tool in functional magnetic resonance imaging (fMRI) studies. In contrast to univariate analyses, which relate the effects of experimental variables to the activity of single voxels or to the average activity within a region of interest (ROI), MVPA leverages machine learning algorithms (Hastie et al., 2001; Vapnik, 1995) to classify (or 'decode') attributes of the experimental stimuli from the distributed pattern of BOLD activity across *many* voxels (Haynes and Rees, 2006; Kriegeskorte, 2011; Pereira et al., 2009; Norman et al., 2006). Successful classification is taken as evidence that the particular collection of voxels under examination contains information relevant to the task at hand. Multivariate analyses have gained wide appeal over univariate approaches for offering improved sensitivity and, in principle, the possibility to map regions coding experimental variables in latent multidimensional spaces (Diedrichsen et al., 2013; Naselaris et al., 2011)(but see Popov et al. (2018), Davis et al. (2014)). This, in turn, greatly deepens the richness of informational content available in neural representations as measured using fMRI.

[*] Corresponding author.
E-mail address: Daniel.A.Stehr@dartmouth.edu (D.A. Stehr).

https://doi.org/10.1016/j.jneumeth.2023.109808
Received 1 September 2022; Received in revised form 28 December 2022; Accepted 29 January 2023
Available online 2 February 2023

With the adoption of multivariate methods in neuroimaging has come the development of experimental design and statistical analytic approaches optimized for machine learning classification. Using either real or simulated datasets, researchers have documented the extent to which select methodological approaches can improve the power and sensitivity of MVPA using support vector machines (SVMs). For example, because standard machine learning practice advises that statistical methods tend to perform better when trained on more observations (Hastie et al., 2001), specialized approaches have been developed for producing accurate and unbiased single-trial activation estimates from rapid event-related fMRI designs to maximize the number of training exemplars (Prince et al., 2022; Turner et al., 2012; Mumford et al., 2012). The efficiency of two such approaches (least squares single, LSS; least squares all, LSA) interact with experimental design considerations such as stimulus onset asynchrony (Mumford et al., 2014; Abdulrahman and Henson, 2016), the number of trials per run and run length (Coutanche and Thompson-Schill, 2012; Zeithamova et al., 2017). Moreover, which approach is more efficient depends, in part, on the mixing proportions of noise components, with different recommendations given when scanner noise dominates trial-level noise or vice versa (Abdulrahman and Henson, 2016). Other studies have instead focused on optimization at the level of the classifier by investigating the impacts of dimensionality reduction techniques (e.g. feature selection) (De Martino et al., 2008; Mourão-Miranda et al., 2006), choice of kernel and hyperparameter settings (LaConte et al., 2005), data partitioning schemes (Etzel et al., 2011; Varoquaux et al., 2017), or the type of performance measure chosen for model evaluation (Dinga et al., 2019).

The practical importance of these reports is that experimental design, statistical modeling and preprocessing approaches interact in complex ways to influence final classifier performance. Left without guidance, researchers must be cautious of the risk of spurious results arising by trying out a large number of processing variations directly on experimental data (Etzel et al., 2011). Though no "one size fits all" set of guidelines is likely to exist for all experimental questions and designs, the field would benefit from more systematic studies of how certain processing strategies (or unique combinations thereof) impact classification of diverse datasets.

1.1. The current investigation

In this paper, we evaluate the independent and joint effects of four data processing approaches on multivariate pattern classification using support vector machines, with the goal of providing recommendations to researchers using rapid event-related designs. We propose that even in the most idealized setting when the researcher is able to make rational and informed design decisions, fMRI data is still susceptible to multiple sources of noise that will impede classification performance. These include trial-level variations in the BOLD, run-level shifts in mean BOLD signal, and human subject movement. Therefore, the four processing approaches we investigate here were selected in a concerted effort to improve the power and sensitivity of MVPA by counteracting these common noise components.

1.1.1. Condition-wise trial averaging

The first noise component we target is trial-level noise by which we mean random deviations around a voxel's mean activation across repetitions of a given condition (Davis et al., 2014). Most MVPA studies to date perform classification on data composed of separate activation estimates for each trial of each run. Although this decision is motivated by the fact that modeling each trial individually maximizes the *quantity* of data available for training the classifier, ultimately it increases the risk of poor classification performance if the signal-to-noise ratio (SNR) at the trial level is low. In fMRI contexts, trial-level deviations from a voxel's expected activation level is assumed to be normally distributed with a mean of zero. Therefore, a simple way to improve the SNR is to simply average the separate trial estimates within each condition, run, and subject, thereby canceling out the noise. Alternatively, one could estimate the average response across trials of each type by collapsing all trials of each type into a single regressor in a traditional GLM, although it is worth noting this latter approach might offer a marginally less precise estimate of neural activity when inter-trial intervals are short (common in rapid event-related designs) and trial variability is higher than scanner noise (Abdulrahman and Henson, 2016).

In univariate analyses, trial averaging increases the spatial extent of activation as an approximate function of the square root of the number of trials averaged (Huettel and McCarthy, 2001). In multivariate contexts, the impact is less well understood. To our knowledge, only one methodological investigation has compared MVPA classification for trial-wise versus (partially) aggregated condition estimates (Zeithamova et al., 2017). The primary focus of that work was on the tradeoff between trial number and SOA, and in supplementary materials the researchers noted a small but significant increase in classification performance when repetitions of the same exemplar were modeled using a single regressor (arguably a form of averaging) despite the reduction in the total number of training exemplars. Averaging *all* trials of each type could potentially drive SNR even higher. Interestingly, a small set of MVPA studies have noted in their methods that they chose to average trial-specific estimates (producing one example per condition per run for training and testing the classifier) in an effort to improve the robustness of the signal (Nestor et al., 2013, 2011; Etzel et al., 2016; Stehr et al., 2021). This is despite the concern that averaging *all* trials of each type dramatically reduces the size of the test sets, which has the potential to increase error variance (Pereira et al., 2009). Therefore, in the current study we include a quantitative comparison of the performance of SVM classifiers built on individual-trial vs trial-averaged data along with a hybrid model in which random subsets of trials from each condition were averaged (partial averaging).

1.1.2. Run-wise mean centering

Another source of noise inherent to all fMRI studies consists in run-level shifts in mean activity, which uniformly impact all trials within each run. Run-based variance may derive from cognitive processes in the participant, such as drifts in attention or changes in physiological arousal, or may occur as an artifact of deconvolving overlapping trials in the presence of noise (Lee and Kable, 2018). Cross-validation using a leave-one-run out approach is commonly recommended for within subject analyses (Etzel, 2015), but it is possible that run-level structure in variance will impede the classifier's ability to find a stable separating hyperplane between clusters composed of trials from different runs. Therefore, run-level mean-centering of trial-specific estimates (i.e. subtracting each voxel's run-level mean from each trial estimate of that run) is recommended as a means to cancel out such run-based shifts and is documented to produce robust and significant improvements in classification accuracy across both real and simulated data (Lee and Kable, 2018). Nevertheless, run-wise mean centering does not appear to be widely adopted among researchers. It is important to note that run-based mean-centering differs from the default scaling applied by many SVM algorithms which is applied to data from all runs simultaneously and therefore cannot adequately address variance structured across separate runs. Given past work showing that different data treatment techniques often interact in complex ways (Etzel et al., 2011), we also investigate how run-wise mean centering interacts with the other techniques investigated here.

1.1.3. Motion-related nuisance regression

Human subject movement is ubiquitous in fMRI data, and such noise has been shown to drastically compromise the quality of statistical analyses (Power et al., 2014; Oakes et al., 2005). This, in turn, has sparked an extensive endeavor to estimate and model motion-related variance (Friston et al., 1996; Hajnal et al., 1994; Power, 2017). Although it is standard practice to include rigid body motion estimates (and their derivatives) as nuisance regressors in functional connectivity analyses, some researchers omit these nuisance factors in univariate statistical design matrices due to concerns that these approaches may be overly aggressive in removing task-related variance. It is unclear whether the benefits obtained by carefully accounting for human subject motion as observed in functional connectivity analyses will be mirrored in MVPA approaches. Therefore, we evaluate four common implementations of motion regression on SVM classification and compare the impact of this additional processing step against that achieved by model-free approaches (i.e. condition-wise trial averaging, run-wise mean centering).

1.1.4. *Cost parameter tuning*

It is important to note that SVM classification performance is sensitive to the cost parameter, C, which controls the bias–variance trade off in the linear boundary fit to the training feature space (Pereira et al., 2009; Cortes and Vapnik, 2015). When C is small, the classifier seeks narrow margins with few violations to the data producing solutions that have low bias but high variance. When C is large, the classifier fits the data less rigidly and is more tolerant to the number and severity of violations to the margin producing lower variance but potentially higher variance. Although the majority of MVPA studies fit SVMs with a fixed cost at a default value of 1, the current advice in the machine learning field is to optimize C in a nested cross-validated fold (a computationally more demanding approach). Because both approaches may be implemented by researchers, we evaluate the above data processing methods using classifiers that are developed using both a fixed cost value ($C = 1$) and costs tuned over wide range of possible values within nested cross-validated folds.

In what follows, we evaluate the success of each approach at mitigating variance at either the trial (condition-based trial averaging), run (run-wise mean centering), or human subject (motion-related nuisance regression) level along with their possible interactions and discuss the relative costs and benefits involved in executing each technique. We implement the manipulations on data consisting of single-trial activation estimates (derived by a LSS GLM) on two separate datasets. The first dataset consisted of real human fMRI data obtained as part of a previously published study in which participants made finger presses in response to visual discrimination judgments (Stehr et al., 2021). The second dataset consists of simulated multivoxel patterns, generated for many crossed levels of trial- and voxel-level noise, in an additional effort to better understand the underlying dynamics of each proposed method.

To preview the results, we found that trial averaging in conjunction with run-wise mean centering produced substantial and robust increases in mean cross-validated classification accuracy across both real and simulated data (an average increase of 16% classification accuracy, SE = 0.04, in real data), as compared to the most standard processing without these data treatments. However, these gains from trial averaging, in particular, were also accompanied by a sizable increase in between-subject variance, likely attributed to having a reduced number of test observations available for externally evaluating the trained classifier. Therefore, we propose a new technique in which we compute multiple averages within each condition and run by randomly sampling trials without replacement. We recommend this technique to researchers who wish to better mitigate the trade-off between the improvement in signal strength from trial averaging and the increased variance due to having a smaller test set.

2. Materials and methods

2.1. *Human participant fMRI data*

fMRI data was acquired from human participants in a study investigating the action observation network (Stehr et al., 2021). Twenty-five healthy adults, ranging in age from 21 to 42 years (mean = 24.7, sd = 3.6), participated in the experiment which was approved by the ethical review board of the University of California, Irvine. One participant was excluded from the study due to excessive head motion. Participants were scanned at the Facility for Imaging and Brain Research at the University of California, Irvine on a 3 T Siemens Prisma MRI scanner (Siemens Medical Solutions) equipped with a 32-channel receive-only phased array head coil. High resolution anatomical images were collected using a single T1-weighted magnetization prepared rapid acquisition gradient echo (MPRAGE) sequence (176 sagittal slices; 1 mm isovoxel resolution; field of view = 256 mm; TR = 2000 ms; TE = 1.99 ms; TI = 900 ms: flip angle = 9 degrees; GRAPPA acceleration factor = 2; bandwidth = 240 Hz/Px).

Functional images were acquired using a T2*-weighted gradient recalled echoplanar imaging multi-band pulse sequence (*cmrr_mbep2d_bold*) from the University of Minnesota Center for Magnetic Resonance Research (68 slices co-planar with the AC/PC; interleaved acquisition; in-plane resolution = 2 × 2 mm; 2 mm slice thickness, no gap; 106 × 106 matrix size; field of view = 212 mm; phase partial Fourier scheme of 6/8; TR = 1500 ms; TE = 30 ms; flip angle = 79 degrees; bandwidth = 2144 Hz/Px; echo spacing = 0.57 ms; excite pulse duration = 8200 microseconds; multi-band factor = 4; phase encoding direction = AP; fat saturation on; advanced shim mode on). At the beginning of each session, an additional pair of EPI images with phase-encoding directions of opposite polarity in the anterior to posterior plane were acquired to correct for susceptibility distortions in each participant's functional data.

This was an event-related study in which participants viewed short (3 s) animations of human avatars performing one of two actions. After viewing the clip, the participant's task was to press a button with the index or middle finger of their right hand reflecting a 2-alternative forced choice judgment on the action depicted. To prevent motor planning during the action vignette, stimulus–response labels were randomized across the two buttons and were not displayed until the vignette was completed. The screen cleared as soon as the participant made their response.

Trials were separated by a 3, 4.5, or 6 s inter-trial interval (ITI), pseudo-randomized within each run such that, in total, each trial lasted 10.5, 12, or 13.5 s. The onset of each trial event, including the response interval, was synchronized with the onset of volume acquisition. The experiment was organized into 8 runs containing 24 trials each for a grand total of 192 trials. Of the 24 participants included in the study, 22 completed the full 8 runs while 2 participants only completed 7 runs.

2.1.1. *Image preprocessing*

Preprocessing was conducted using BrainVoyager QX v20.6 (Goebel et al., 2006). All functional images were slice-time corrected, motion corrected to the first volume both within and between runs using rigid body transformations, linearly detrended, and temporally high pass filtered (cutoff frequency 0.01 Hz). Scans were additionally corrected for susceptibility-induced magnetic field distortions using the field map method (Jezzard and Balaban, 1995), implemented in BrainVoyager's COPE v1.0 plugin. All functional images were co-registered to each individual's T1-weighted image.

2.1.2. *ROI definition*

For the current investigation, our aim was to classify which of two buttons was depressed during the response intervals of a visual discrimination task in two regions of interest (ROIs): left somatomotor (SomMot) and right primary auditory cortex (A1), respectively.

SomMot. To identify brain areas activated by making button presses, events were modeled as a boxcar function of 200 msec duration starting from the moment the button was depressed, convolved with a two-gamma hemodynamic impulse response function (Friston et al., 1998; Glover, 1999). Fixed effects maps for individual subjects were then aligned to a template subject (a pilot participant) in surface space using

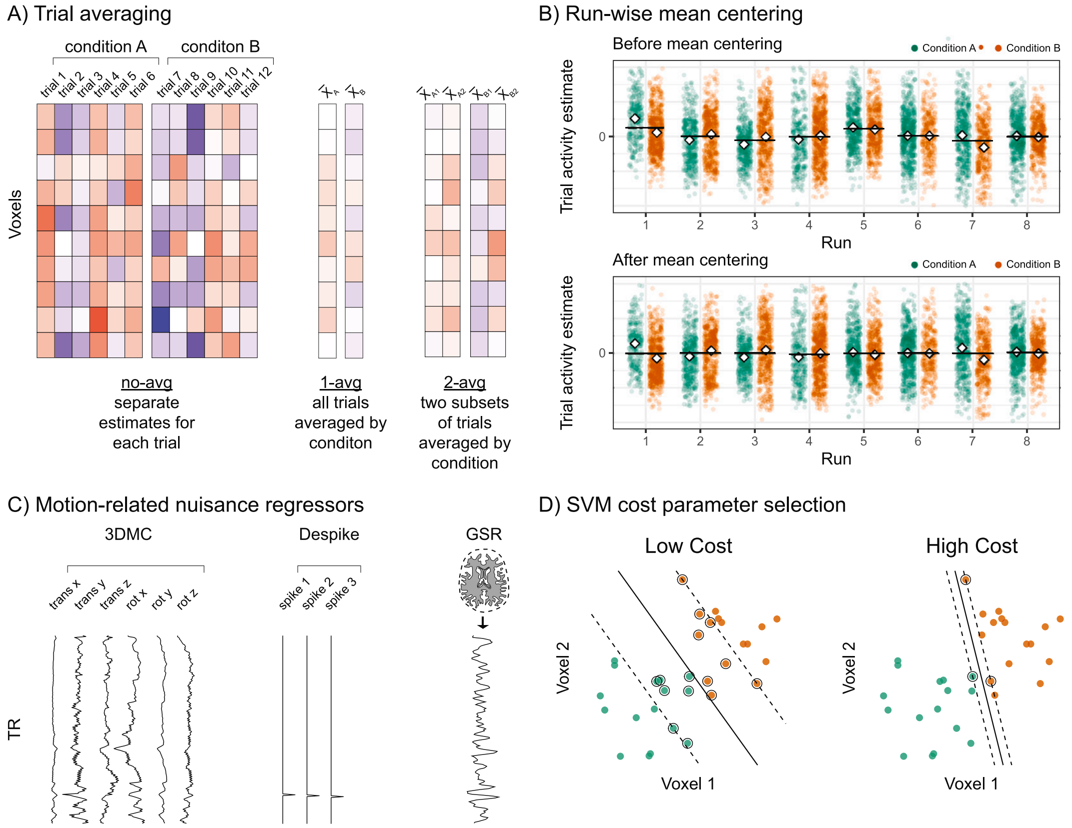

Fig. 1. (A) Method of aggregating data for classification, shown for a sample of 12 trials from a single run of human participant data (somatomotor region). (B) The effect of run-wise mean centering shown from a sample subject (somatomotor region). Diamonds represent the marginal means within runs and black horizontal bars represent the overall mean within runs. (C) Sample motion-related nuisance regressors, from a single subject and run. (D) The impact of the cost hyper-parameter, C, on the SVM decision boundary.

cortex-based alignment (Frost and Goebel, 2012), which allows for group level analysis without applying nonlinear registration into atlas space. The resulting left hemisphere somatomotor ROI, contralateral to the right hand button presses, was identified using a random effects group GLM, with threshold determined using the false discovery rate $q < (1 \times 10^{-6}.)$. The identified vertices were then projected back into native volumetric coordinates to extract voxel-wise patterns specific to each participant. Across participants, SomMot ranged in size from 588 to 828 voxels (mean=691, SD=62.2).

A1. Primary auditory cortex served as a control ROI, and was identified anatomically in each individuals' native surface using Freesurfer's cortical surface atlas mapping algorithms in conjunction with the 1,000 atom resolution Schaefer atlas (Schaefer et al., 2018). This atlas emphasizes homogeneity of function within parcels coupled with high resolution "atomic" parcellation in approximately equisized units. Across participants, A1 ranged in size from 386 to 632 voxels (mean=496, SD=62.4).

2.1.3. Motion-related nuisance regressors

We evaluated the impact of 4 different types of motion-related nuisance regressors on MVPA classification results (see Fig. 1 C). Nuisance regressors included the detrended time series of the 6 rigid-body realignment parameters (R = [X Y Z pitch yaw roll]), estimated from the three-dimensional motion correction (3DMC) procedure performed during preprocessing. The 24 parameter Volterra expansion nuisance model included the 6 rigid body estimates and the preceding timepoint, as well as their first derivatives ((Friston et al., 1996): [R = [R R^2 R_{t-1} R_{t-1}^2], where t and $t-1$ refer to the current and immediately preceding timepoint).

Despiking (e.g. volume censoring) is commonly used to reduce variance accounted for by large head jerks (producing large changes in image intensity) which may not be captured well by the 3DMC or Volterra nuisance regressors (Lemieux et al., 2007; Satterthwaite et al., 2013; Yan et al., 2013). Despiking was performed by including in the model a matrix of "scan nulling" regressors (i.e. a heaviside function) targeting each corrupted timepoint identified as volumes with framewise displacement exceeding 0.5 mm (Power et al., 2012). In addition, all trials with three or more timepoints censored near the peak of the expected hemodynamic response were excluded due to increased variance in the trial-specific beta estimates when peak volumes are censored.

Global signal regression (GSR) is commonly used to remove distributed, non-neural sources of variance contaminating the images (Power, 2017). The global signal regressor was computed as the average BOLD intensity measured from the white matter and ventricles over time, using the anatomical masks derived from Freesurfer's segmentation, and the first derivative thereof.

All nuisance regressions were performed prior to estimating the trial-by-trial activation estimates, which constitute the data passed on to the classifier. The denoised timeseries was produced by regressing each voxel's timeseries onto the respective matrix of nuisance regressors and collecting the model residuals.

2.1.4. Trial-specific activation estimates

After z-scoring the pre-processed time series, trial-specific activation estimates for all voxels within the ROIs were derived by iteratively fitting a separate general linear model for each trial that included one regressor for the trial-of-interest and two nuisance regressors modeling all other trials grouped by the type of button that was pressed (the least squares separate or LSS approach (Mumford et al., 2012; Turner et al., 2012)). Trial-specific activation estimates were generated using Matlab R2018b (TheMathWorksInc, 2017) and the code used to produce them is available at: https://doi.org/10.6084/m9.figshare.13708654.v1.

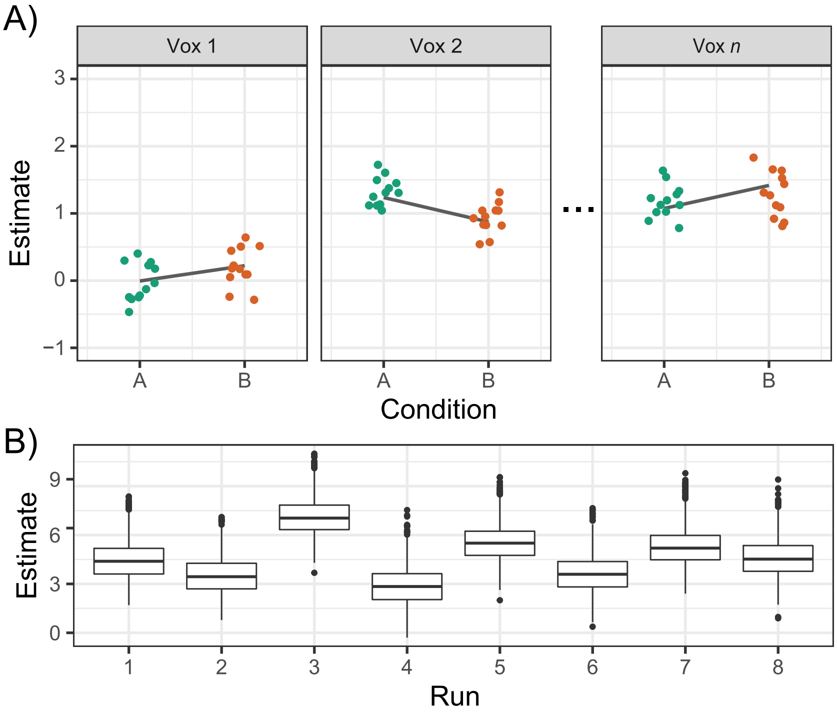

Fig. 2. (A) Data from three sample voxels illustrating how trial-specific estimates were created by first generating a line by sampling an intercept and slope parameter from a bivariate normal distribution and then adding normally distributed noise independently to each trial. Each voxel's 'ideal' line is shown for illustration only and reflects the difference between the expected responses to each type of trial in that voxel. (B) Illustration of run-level shifts in mean activity across all trial types shown from a sample of simulated data.

2.2. Simulated multi-voxel activation patterns

Activation patterns were simulated using an analytic framework that describes the observed activation on any given trial as a combination of the fixed effects of the experimental variables along with random deviations from these fixed effects at the trial, voxel and run level. To generate multivariate response patterns that properly incorporate these unique variance components, we used a multilevel modeling approach (Diedrichsen et al., 2013; Davis et al., 2014) that incorporated trial, voxel, and run-level variability. Fig. 2 conceptually illustrates how the pattern data was generated and Appendix A details the formal description. Custom R code for generating activation patterns is provided at https://doi.org/10.6084/m9.figshare.13708654.v1.

Each voxel was assigned a unique intercept and slope term (see Fig. 2 A) by sampling from a bivariate Gaussian distribution with mean intercept of one and mean slope of zero, with the intercept and slope assumed to be uncorrelated. This corresponds to a population of voxels with no effect of stimulus condition, on average. The standard deviation of the intercept term was fixed at 1 while the standard deviation of the slope term was varied over four values (0.01, 0.05, 0.1, and 0.2), simulating a range of effect sizes across the voxels but no univariate effect of the conditions (i.e. voxel's with more extreme slopes are more “informative" about the conditions of interest). Following the convention set forth by Davis and colleagues (Davis et al., 2014), we refer to the voxel-level variability in slope (i.e. effect size) by τ_{β_1} which is a variance component of the G matrix in Diedrichson's random effects model (Diedrichsen et al., 2011).

Run-specific shifts in mean activity across all conditions were constructed through independent draws from a Gaussian distribution with mean zero and standard deviation 1.5, then added to each voxel's intercept term of each run as a single constant. Each voxel's idealized response to the experimental conditions, remained consistent across runs.

Normally distributed noise was generated for each trial with mean zero and standard deviation varied over five values (0.5 to 1.5 with step size of .25). Finally, trial-by-trial activation estimates were generated by linear combination of all fixed and random effects (including trial-, voxel-, and run-related deviations).

This process was implemented for an ROI containing 200 voxels using a study design with 12 repetitions of two trial types, for a total of 24 trials per run. Each simulated study consisted of 8 runs of data which we generated 30 times, simulating 30 subjects. For ease of interpretation, subjects were not modeled as random effects though each simulated subject was generated by independent draws from the model. This simulation method captures the intuition that trials are repeated measurements across voxels and scans; and that voxels differ in how informative they are about the experimental variables.

An additional simulation addressed specifically the impact of trial averaging and mean-centering on situations where all voxels entirely lack information relative to the conditions being decoded. In these simulations a 200 voxel ROI was generated such that each voxel had a mean slope of zero with zero variability ($\tau_{\beta_1} = 0$). Thirty studies containing thirty participants each were simulated for the same five levels of trial-level variability as in the other simulations, with all other parameters constant.

2.3. Classification

MVPA was performed on the acquired and simulated data using a linear support vector machine with default scaling (all voxels and observations standardized to zero mean and unit variance) using custom R code relying on the e1071 package (Meyer et al., 2018). The labels classified in the human subject fMRI data corresponded to which of two response buttons were pressed during the experiment. In the simulated data, the SVM algorithm classified labels corresponding to the two simulated trial types — type A and type B. All analyses were performed within subject using 8 fold leave-one-run-out cross validation. Within each fold, predictions were made on samples from the left out run and final classification performance was computed by taking the mean across all folds.

2.3.1. Trial averaging by condition and run

We investigated the impact of three methods of aggregating event-related data for multivariate analyses (see Fig. 1A). The first and most commonly used method involves training and testing the classifier on data composed of activation patterns observed on each individual trial. In this analysis, the number of samples (training and testing, combined) passed to the classifier is equal to the number of trials within the current data partition (i.e. exchangeability block, Winkler et al. (2014)). Due to unavoidable temporal dependencies between adjacent trials and the tendency for trial estimates drawn from within the same run to be more similar than trial estimates drawn from different runs (Pereira et al., 2009; Etzel et al., 2009) it is ideal to partition data such that all the trials from a single run are excluded from training and used for subsequent validation, which eliminates the potential for within-scan bias during classifier training.

The second approach we investigated involves averaging all trial-specific estimates of each type within run (e.g. averaging all Type A trials within run into a single, average sample). We refer to this method as “1-Avg" because it results in one averaged observation per class within each run. Averaging trial-specific estimates by run has the potential to reduce trial-variability that could be a major source of noise limiting classifier performance, but comes at the expense of greatly reducing the number of training and testing examples supplied to the classifier. Reducing training observations has the potential to impoverish the fit of the decoder to the data whereas reducing test observations impacts the precision with which the prediction error can be estimated within each cross-validated fold thereby increasing between-subject variance of the final classification accuracy

In a third approach we introduce a hybrid model to strike a better balance between the opposing effects of improving signal-to-noise ratio (SNR) by trial averaging and maintaining a sufficiently large number of test samples. In this “2-Avg" approach, we randomly sampled (without replacement) half of the trials from each condition within a run, then averaged each group of trials separately thus producing two averaged activity estimates per condition, per run. For our datasets, both of which involve two trial types, this results in a test set of four observations. To reduce sampling error, this process was iterated ten times within each fold and the resulting classification accuracies were averaged.

2.3.2. Run-wise mean centering

Run-wise mean centering (see Fig. 1B) was performed by subtracting each voxel’s run-level mean beta estimate, for all trial types, from the estimates within that run (Lee and Kable, 2018; Etzel et al., 2011; Pereira et al., 2009). Though we subsequently refer to this operation as “mean centering", it should be noted that this type of mean centering is distinct from the default mean centering performed by the majority of SVM algorithms that operate on all samples pooled across the acquisition scans.

2.3.3. Cost tuning

Many MVPA studies fit linear SVM classifiers to multi-voxel response patterns using a fixed cost parameter, C, of 1. However, optimizing C (see Fig. 1D) by minimizing the cross-validated test-error has been shown to improve the predictive power of a classifier (Hastie et al., 2001). On both datasets, we compare the benefits of tuning the cost parameter over 12 values from the more liberal 2^{-12} to the more rigid 2^1 compared to using a fixed $C = 1$. This was achieved using a nested cross-validated fold in which an inner second level-split was generated leaving one run of the original training data out and used to evaluate the performance of each value of C. This was repeated for all folds of the nested loop, and the lowest value of C maximizing predictive accuracy of the inner cross-validation test data was then applied to the training and testing data in the external loop.

2.3.4. Statistical analyses

Statistical significance of classification results at the group level using only the most common processing decisions (no trial averaging, no mean centering, no motion-related nuisance regression, no cost tuning) was evaluated in each ROI using non-parametric permutation tests. In these tests, labels of button responses were permuted within individual participants 1,000 times each and classifiers were trained and testing using the same procedures outlined above. This yielded the expected distribution of classification accuracy for each participant under the null hypothesis. Significance ($p < 0.05$, one tailed) was ascertained from group-level null distributions constructed using a bootstrap procedure in which a single sample was drawn from each participant’s null distribution (iterated 1,000 times and sampled with replacement).

To statistically evaluate the individual and joint impact of the four methodological decisions on MVPA decodability, we constructed multilevel linear models (MLMs) for both the real and simulated datasets. For the human fMRI data, MLMs were created using as dependent variable each participant’s cross-validated classification accuracy. Fixed factors included the type of ROI (somatomotor versus the control region) and the four methodological treatments (type of motion-related nuisance regression, presence or absence of trial averaging, presence or absence of mean centering, and cost parameter choice). Participants were modeled as a random effect (random intercepts) and data from each ROI was explicitly nested within each participant to account for shared variance (a nested random effects structure).

The data grouping structure of the simulated data diverged considerably from the human participant data in the sense that ‘subjects’ were generated by independent draws from the model for a variety of levels of trial-level noise and voxel-level variability in effect size. We refer to each unique combination of trial-level noise and voxel-level variability as a ‘dataset’, with each dataset containing activation patterns from 30 simulated subjects. The random effects structure of the MLM applied to the simulated data was therefore specified to include a random intercept for each dataset (to account for differences in baseline classification accuracy across different parameter settings) with simulated subjects nested within datasets. The fixed effects included the same methodological factors that were tested in the human fMRI data with the exception of motion-related nuisance regression, which was not evaluated because the data were simulated at the level of trial activation estimates rather than timeseries.

All statistical analyses were conducted in R using the ‘nlme’ package (Pinheiro et al., 2019). The significance of each factor (or interaction) was assessed using Likelihood Ratio Tests comparing each model to reduced models lacking the variable (or interaction) in question.

3. Results

3.1. Human participant data

To establish a baseline for comparing the impact of the four methodological approaches on mean classification accuracy, we begin by reporting group-level results in each ROI using only the most common processing combinations: no motion-related nuisance regression besides that typically deployed during preprocessing, training and testing on separate estimates for each trial, no run-wise mean centering of trial estimates, and training the SVM with a fixed cost value of 1.

Across all 24 subjects, the left somatomotor region (SomMot) classified the button pressed (button 1 versus button 2) with a mean accuracy of 56.90% (SE = 1.36), which non-parametric permutation tests revealed to be significantly higher than that expected by chance ($p_{perm} < 0.001$, mean of null distribution = 50.10%, SD = 0.88). The ROI serving as a control region (primary auditory cortex or A1) classified the type of button pressed with a mean accuracy of 51.89% (SE = 0.82) which is 5.01% lower that obtained in SomMot and yet still higher than

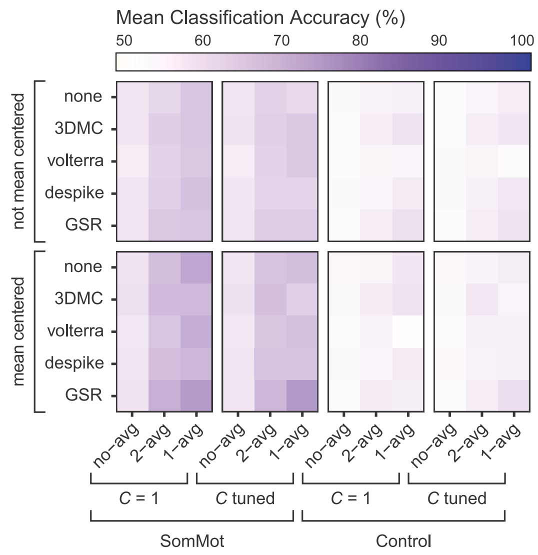

Fig. 3. Average classification accuracy for all combination of methodological decisions grouped by ROI (SomMot = somatomotor; Control = primary auditory cortex).

that expected by chance (p_{perm} = 0.029, mean of null distribution = 50.17%, SD = 0.90).

Fig. 3 displays mean classification accuracies for all combinations of methodological approaches applied to the human fMRI data. Initial inspection of the data revealed that trial averaging, run-wise mean centering, and certain types of motion-related nuisance regression lead to higher across-subject variability in classification accuracies (i.e. heteroscedasticity), which is a problem for lineal models. For a discussion of measures taken to address violations to the assumption of equal variances, please see Appendix B.

The multilevel linear model (MLM), including all methodological factors as well as the type of ROI as fixed factors, revealed a significant main effect of ROI on classification accuracies ($\chi^2(10) = 12.208, p < 0.001$) with SomMot classifying the type of button pressed with significantly higher accuracy than the control region (b = 5.434, SE = 1.458, $t(23)$ = 3.727, p = 0.001).

However, the main effect of ROI was qualified by a significant three-way interaction between ROI, the method of trial averaging, and the presence or absence of within-run mean centering ($\chi^2(65) = 19.947, p < 0.001$). To interpret this interaction, planned contrasts compared classification accuracies for the two methods of computing condition-based trial averages (2-avg and 1-avg) to the results obtained by classifying data consisting of separate estimates for each trial (no-avg). Fig. 4 shows all parameter estimates along with 95% confidence intervals.

First, fixed effect parameter estimates revealed that classification accuracies were higher in both the 2-avg and 1-avg conditions compared to no-avg (2-avg vs no-avg: b = 4.180, SE = 0.235, $t(2813)$ = 17.761, p < 0.001; 1-avg vs no-avg: b = 5.124, SE = 0.430, $t(2813)$ = 11.930, p < 0.001). Additionally, classification accuracies in the 2-avg and 1-avg conditions were significantly higher when the data was also mean centered within each run (2-avg vs no-avg with mean centering: b = 1.795, SE = 0.500, $t(2801)$ = 3.591, p < 0.001; 1-avg vs no-avg with mean centering: b = 2.072, SE = 1.030, $t(2801)$ = 2.011, p = 0.044). Furthermore, this increase in classification accuracies by averaging, when coupled with run-wise mean centering, was found to exist only in the SomMot region (2-avg vs no-avg with mean centering in SomMot vs control: b = 3.664, SE = 0.960, $t(2780)$ = 3.818, p < 0.001; 1-avg vs no-avg with mean centering in SomMot vs control: b = 5.250, SE = 2.034, $t(2780)$ = 2.581, p = 0.010).

In order to better understand the effect of trial averaging within each ROI, two separate MLMs were constructed using only data from each ROI. Both models revealed that averaging trials together within each run improved classification accuracies over not averaging any trial estimates (SomMot: $\chi^2(10) = 377.355, p < 0.001$; Control: $\chi^2(10) = 66.382, p < 0.001$). However, trial averaging improved decodability in SomMot (2-avg vs no-avg: b = 6.025, SE = 0.325, $t(1401)$ = 18.540, p < 0.001; 1-avg vs no-avg: b = 7.120, SE = 0.570, $t(1401)$ = 18.540, p < 0.001) considerably more than it did in the control region (2-avg vs no-avg: b = 2.312, SE = 0.328, $t(1414)$ = 7.058, p < 0.001; 1-avg vs no-avg: b = 3.155, SE = 0.638, $t(1414)$ = 4.946, p < 0.001).

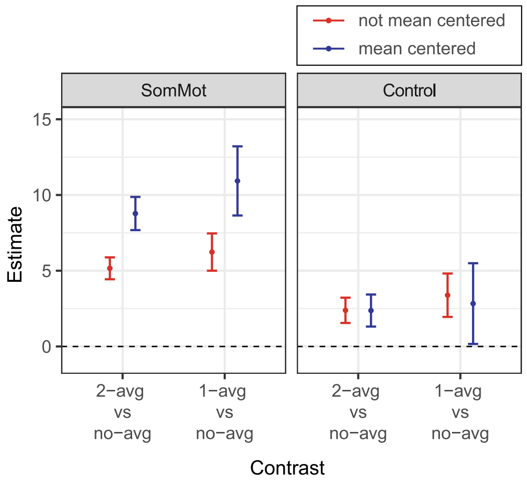

Fig. 4. Fixed effect parameter estimates from multilevel linear models (MLMs) showing the interaction between ROI, trial-averaging technique and within-run mean centering in the human subject fMRI dataset. The 95% confidence intervals were computed for the contrasts comparing the two conditions where trials were averaged within runs (2-avg and 1-avg) versus data comprising a separate estimate for each trial. Parameter estimates above zero indicate that averaging trials by run produced higher accuracies than training/testing the classifier on individual trial estimates. Estimates were computed from four separate MLMs fixing the level of ROI and mean centering.

As shown in Fig. 4, confidence intervals were wider for the 1-avg condition compared to the 2-avg condition, reflecting larger between subject variation in classification accuracies when trial estimates for an entire run are averaged into a single exemplar per condition. This was true of both ROIs, indicating that the increase in variance is linked to having fewer observations rather than the presence or absence of true signal embedded in the data.

The impact of the type of motion-related nuisance regression (i.e. data cleaning step) applied prior to pattern estimation can be seen by comparing the rows of Fig. 3 within each panel, with parameter estimates shown in Fig. 5. There was a significant main effect of the type of data cleaning step applied ($\chi^2(14) = 17.190, p = 0.002$) as well as a significant interaction between the data cleaning step and the amount of trial averaging applied within run ($\chi^2(34) = 19.423, p = 0.010$).

This interaction was broken down by comparing classification accuracy resulting from each data cleaning step to accuracies obtained from using no motion-related nuisance regression separately for each of the two contrasts on the trial averaging level (2-avg vs no-avg and 1-avg vs no-avg). These contrasts revealed that applying global signal regression (GSR) to the timeseries before extracting trial estimates significantly improved classification accuracy in the 2-avg condition compared to no-avg (b = 2.222, SE = 0.729, $t(2795)$ = 3.049, p = 0.002) as well as in the 1-avg condition compared to no-avg (b = 2.881, SE = 1.347, $t(2795)$ = 2.140, p = 0.033).

No other data cleaning steps significantly differed by the type of trial averages computed. Parameter estimates from the main effect of data cleaning step revealed that, averaged across all other factors, using the Volterra expansion as nuisance regressor significantly lowered classification accuracy (b = –0.800, SE = 0.285, $t(2815)$ = –2.8100, p = 0.005).

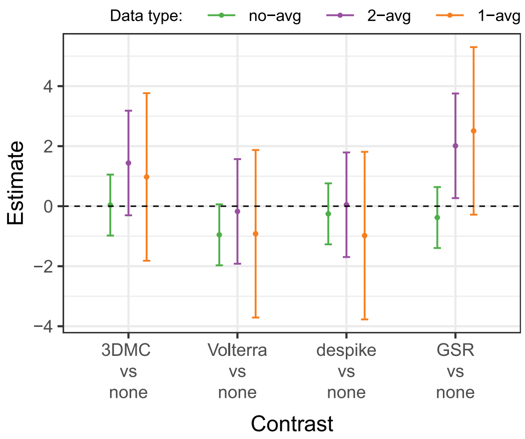

Fig. 5. The interaction between type of motion-related nuisance regression (data cleaning) and trial averaging within the human subject fMRI dataset. Results show parameter estimates with 95% confidence intervals from a multilevel linear model. Contrasts were set on the type of data cleaning step applied by comparing each data cleaning step to using no nuisance regression at all. Contrasts on the type of trial averaging method compared each method to the baseline approach using a separate activation estimate for each trial. Therefore, estimates above zero indicate that the given data cleaning step produced higher classification accuracies for the given trial averaging method versus using no trial averaging.

The impact of cost parameter selection can be assessed by comparing the first and last three columns within each group of ROIs in Fig. 3. Overall, choosing a fixed cost value of one versus tuning the cost parameter did not impact mean classification accuracies nor interact with any of the other three processing decisions (all *p's n.s.*).

3.2. Simulated data

Fig. 6 shows mean classification accuracies from simulated pattern data for all combinations of methodological factors (level of trial averaging, presence or absence of within-run mean centering, and cost parameter selection). These methodological approaches were applied to several simulated datasets generated with varying levels of trial-level variability (σ^2) and voxel-level variability in effect of experimental conditions (slope or τ_{β_1}).

Overall, mean classification accuracies varied with both trial- and voxel-level variability. Classification accuracy increased when trial-level variability decreased, indicating that more consistent patterns across trials improved decoding. Moreover, classification accuracy increased as voxel-level variability in the mean effect of the experimental conditions increased, consistent with reports that increased variance in the spatial patterns, even when the fixed effect size is zero, improves classifier performance (Davis et al., 2014).

An MLM was conducted to determine which, if any, of the data processing choices impacted mean classification accuracies across all combinations of model parameter settings used to generate the pattern data. The MLM included the cross-validated classification accuracies as the dependent measure and included a random intercept for each combination of trial variability and voxel-level variability in slope (e.g. the dataset) with simulated participants nested within datasets. As such, it evaluated the independent and joint effects of the processing methods on classification accuracies across all datasets. The analysis yielded many significant main effects and interactions, therefore we focus on the highest order interaction which occurred between all three processing choices ($\chi^2(20) = 67.104, p < 0.001$). Fig. 7 displays this interaction graphically by plotting the MLM parameter estimates along with confidence intervals for the two contrasts on trial averaging from four simpler MLMs holding mean centering and cost choice constant.

Results from the full MLM revealed that both methods of trial averaging improved classification accuracies over using separate trial estimates (avg-2 vs no-avg: b = 5.173, SE = 0.182, t(8248) = 28.368, p < 0.001; avg-1 vs no-avg: b = 5.369, SE = 0.249, t(8248) = 21.567, p < 0.001). Furthermore, this improvement from averaging trials (both 2-avg and 1-avg) was significantly higher when the data were also mean centered within runs (avg-2 vs no-avg: b = 4.842, SE = 0.358, t(8248) = 13.524, p < 0.001; avg-1 vs no-avg: b = 8.530, SE = 0.588, t(8248) = 14.497, p < 0.001). Finally, tuning the cost parameter improved classification accuracies for both trial averaging methods but much less so when the data had been mean centered within each run (avg-2 vs no-avg: b = −5.189, SE = 0.701, t(8239) = −7.397, p < 0.001; avg-1 vs no-avg: b = −4.865, SE = 1.167, t(8239) = −4.169, p < 0.001).

3.3. Simulations with zero effect size for all voxels

To evaluate whether trial averaging or within-run mean centering has the potential to artificially inflate mean classification performance, we conducted simulations to quantitatively evaluate the impact of these methodological choices on a true null condition in which all the voxels in the pattern are drawn from the same distribution that is uninformative about the conditions. To that end, multivariate response patterns were generated that contained a mean effect size of zero and zero variability in effect size across voxels ($\tau_{\beta_1} = 0$). Thirty studies containing thirty participants each were simulated for the same five levels of trial-level variability as in the other simulations, with all other parameters constant. Results of the classifications can be seen in supplementary Figs. 1 and 2.

An MLM was constructed with dependent variable set to the mean classification accuracy across all 30 studies and random intercepts for each study and each level of trial-level noise variability. This analysis revealed that mean classification accuracy did not vary significantly with trial averaging ($\chi^2(11) = 2.54, p = 0.280$), presence or absence of within-run mean centering ($\chi^2(12) = 2.170, p = 0.140$), or the decision to use a fixed versus tuned cost value ($\chi^2(13) = 0.001, p = 0.981$).

4. Discussion

We evaluated the impact of four methodological approaches on MVPA-decoded classification accuracies in both real and simulated fMRI data. These methodological considerations were selected, in part, because of their common use in fMRI univariate and functional connectivity analyses, with the potential benefits when implementing them for multivariate pattern analysis unclear. This analysis is intended to serve as a practical guide for researchers wishing to optimize multivariate classification analyses without the risk of introducing spurious results by testing each method directly on experimental hypotheses of interest.

Some general observations across these analyses warrant attention. First, methodological approaches leading to large improvements in SVM classifier performance did so in the context of both real and simulated datasets. In this analysis, that is most prominently the case with run-wise trial averaging coupled with mean-centering. The benefits of these approaches for the classification of both real and simulated data is evidence that the potential to improve classification is not limited to highly specific characteristics of either dataset. With that said, future studies should test the effectiveness of these methods across a wider range of experimental designs, regions of interest and types of classifiers.

Secondly, the methods producing significant improvements often interacted in complex ways, highlighting the complex dynamics inherent to SVM analyses of multivariate pattern data. For example, the benefit of global signal regression for classification accuracy was only apparent for trial averaged data, with no improvement observed for MVPA conducted on individual trial exemplars. It is precisely these interactions that motivated this evaluation of processing pipelines.

Lastly, while classifying button presses in real human participant data, the improvements brought about by these decisions were much

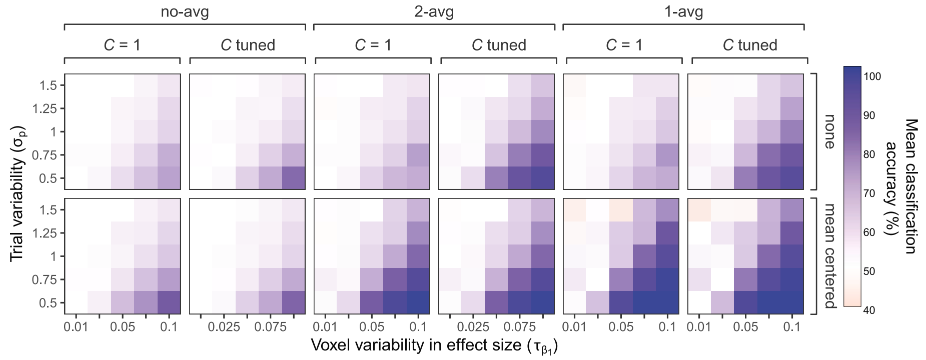

Fig. 6. Average classification accuracy for each combination of methodological factors applied to simulated pattern data. Pattern data was simulated for several combinations of trial-by-trial variability, σ, and voxel-level variability in the mean difference between trials of each type, τ_{β_1}. Each colored square displays mean cross-validated classification accuracy computed across 30 different simulations.

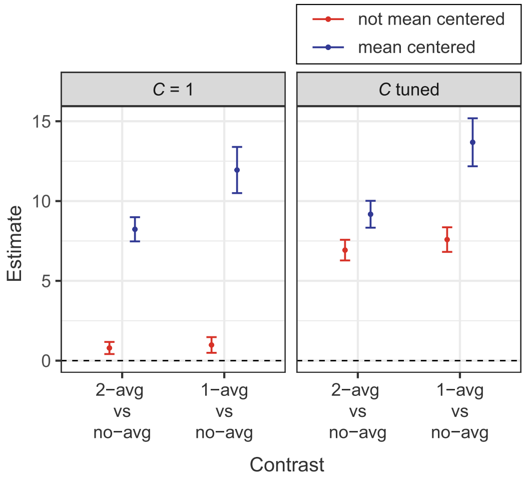

Fig. 7. The three way interaction between trial averaging, cost tuning, and mean centering present in the simulated pattern data. Fixed effect parameter estimates and 95% confidence intervals were computed from four multilevel linear models contrasting the trial averaging method versus using separate trial estimates while fixing the method of mean centering and cost parameter selection method. An estimate above zero indicates that the trial averaging technique deployed improved mean classification accuracy versus training/testing on separate trial activation estimates.

larger in a region of interest in which we had strong *a priori* expectations for highly accurate classification (somatomotor) versus a control region (primary auditory cortex). Analysis of simulated multivoxel activation patterns which lacked any informative content reinforce that trial averaging and mean centering do not artificially inflate mean classification performance in the absence of real signal. This is very reassuring, as one does not want to unintentionally introduce bias to the classification algorithm, as has been observed with some feature-reduction approaches (Ambroise and McLachlan, 2002).

4.1. Trial averaging

It is generally advised to use as many observations for training the classifier as possible (Pereira et al., 2009; Etzel et al., 2009). Therefore classifying based on separate estimates for each trial may be thought to give better results because it maximizes the training set size. Alternatively, averaging trials by condition and run could reduce uniformly distributed trial-level noise, thus enhancing the discriminability of the multivariate patterns by improving the signal-to-noise ratio (SNR). We found that reducing noise by trial averaging produced one of the largest gains in classification accuracies among the methods we tested and this result was consistent for both real and simulated data.

Given the trade-offs anticipated from trial averaging (reduced number of training/test exemplars versus trial-level noise reduction), two findings from this analysis are particularly surprising. The first is the magnitude of improvement induced by trial averaging. In the human participant data, when classifying button presses in somatomotor cortex, the improvement in mean classification accuracy brought about by averaging all trials of each type within runs was 6.3% and when coupled with within-run mean centering (discussed below) the improvement climbed to 10.9%.

Another important finding is that trial averaging causes a marked increase in the between-subject variability of the classification accuracies. One possible explanation for the increased variance may be the reduced size of the test set used to assess the prediction error of the classifier at each split of cross-validation. When estimating classification accuracy using the more traditional trial-based approach, the algorithm is tested on an entire run of samples, which in this study consisted of 24 exemplars (twelve from each condition). When all trial estimates are averaged within run to one per condition (1-avg), the cross-validated test error is evaluated with only two observations per split, constraining the test error to only a few possible values. It has been theoretically shown that with training sets of the same size, having more data for validation decreases the variance of the estimated accuracy (Arlot and Celisse, 2010). Therefore, we conclude that it is important to strike a balance between maintaining a large enough test set to yield a stable estimator of classifier performance and reducing trial-level noise through trial averaging.

As expected, doubling the number of items in the test set nearly halved the between-subject variability in classification accuracy (the 2-avg condition compared to the 1-avg condition). However, this was also associated with a reduction in mean classification accuracy, which we interpret as due to a higher SNR from having fewer trials included in each average. Therefore, when planning an MVPA study, researchers should carefully weigh any knowledge they have about the amount of trial-level noise inherent to the region(s) under study versus the increased test-set variance brought about by limiting that noise through averaging trials of variously sized subsets.

4.2. Run-wise mean centering

It is widely recognized that each scan in a session is associated with a unique shift in the mean MR signal across all trial types. These shifts may reflect the cognitive state of the participant, such as drifts in attention and changes in physiological arousal, or the state of the MR hardware (i.e. thermal noise, scanner drift). Whereas condition-based trial averaging was used to reduce trial-by-trial variability, the variance component that run-wise mean centering aims to reduce is run-level variation in the baseline activation for all trials within each run.

The mechanism by which this improves classification is intuitive: Since the cross-validation procedure for most MVPA studies is partitioned on runs, training a classifier using exemplars from run-shifted distributions introduces artificial clusters within the training data. This in turn, should be anticipated to impair the classifier's ability to find a stable separating hyperplane between blocks of training data from different runs or to generalize to test data from new runs. Our data confirm this hypothesis in both real and simulated datasets, replicating other studies (Lee and Kable, 2018; Etzel et al., 2011; Pereira et al., 2009).

Furthermore, we show that mean centering interacts with the method of trial averaging. When training and testing using separate trial estimates, mean centering did not make a significant difference to mean classification accuracy. This may be because the increased variance associated with the noisy trial exemplars in effect masks the partitioning effect of run-wise variance. However, with the inclusion of run-wise trial averaging, trial variance is reduced and large improvements are seen when mean centering is included.

4.3. Cost selection

Tuning the SVM cost parameter, C, within a nested cross-validated loop is a computationally intensive process, particularly when conducted over many regions of interest (as in a searchlight MVPA analysis) or when implemented as part of permutation testing. Consistent with previous analyses (Varoquaux et al., 2017) our results also show that tuning C versus using a fixed value of 1 depends on the statistical structure of the underlying dataset.

Cost tuning did not have a significant impact on classification performance using the human participant data in either ROI. In contrast, cost tuning significantly interacted with trial averaging and mean centering in the simulated datasets such that cost tuning improved classification accuracies when trials were run-averaged and mean centered prior to classification. One explanation for this finding is that setting C high, such as when $C = 1$, leads to a higher likelihood of overfitting the classifiers (Hastie et al., 2001), a significant disadvantage when the classifier is trained and tested on data composed of blocks with distinct shifts in mean activity. Thankfully, our results show that cost tuning can be omitted from MVPA pipelines without penalty by simply mean centering the data within each run prior to classification, which is a computationally simpler and faster operation.

4.4. Motion-related nuisance regression

It has long since been recognized that head movements severely compromise the quality of fMRI data (Friston et al., 1996; Hajnal et al., 1994), sparking many endeavors to denoise the BOLD signal through reference time series capturing motion-related fluctuations (Caballero-Gaudes and Reynolds, 2017). These reference signals are sometimes added as nuisance regressors to the design matrix that is fit to the voxel time series and therefore constitute additional data cleaning above and beyond the volume registration performed during normal preprocessing. Though it is now standard to use such nuisance regressors to denoise BOLD data in preparation of functional connectivity analyses, no studies to date have examined their impact on multivariate decoders.

Prior to estimating trial-specific activation estimates, we denoised the raw timeseries using four different types of motion-related nuisance regressors: the 6 rigid-body realignment parameters (3DMC), the 24 parameter Volterra expansion, a despiking model using an FD threshold of 0.5 mm, and the average signal from the white matter and ventricles (GSR). Denoising the data using GSR led to a significant increase in classification accuracies but only for data that had undergone trial averaging (both 1-avg and 2-avg conditions). Also, including the full Volterra expansion of the rigid body realignment parameters as nuisance regressors significantly reduced accuracies irrespective of whether trial averaging, run-wise mean centering, or cost tuning was applied.

4.5. Conclusions

Though hard and definitive guidelines regarding the tested methods cannot be drawn for all designs and tasks, the current investigation reveals that across real and simulated datasets MVPA-decodability can be significantly improved through trial averaging, mean centering, and inclusion of Global Signal Regression.

CRediT authorship contribution statement

Daniel A. Stehr: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Javier O. Garcia:** Methodology, Writing – review & editing. **John A. Pyles:** Investigation, Supervision, Funding acquisition. **Emily D. Grossman:** Conceptualization, Methodology, Investigation, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data (in the form of LSS beta estimates for each ROI) and code for simultating beta patterns and performing classifications using all of the methods described in this paper is available at: https://doi.org/10.6084/m9.figshare.13708654.v1.

Acknowledgment

The human participant fMRI data supporting these analyses is based upon work funded by the National Science Foundation under grant BCS-1658560 to EG and and BCS-1636038 JP.

Appendix A. Formal framework for multi-voxel simulations

There is wide agreement that BOLD fMRI data contains multiple sources of variability, including trial-, voxel-, and run-level variability (Friston et al., 1994). To generate multivariate response patterns that properly incorporate all these unique variance components, we used a multilevel modeling approach (for similar models, see Diedrichsen et al., 2013; Davis et al., 2014).

The first level of the model is given by Eq. (A.1) and describes how activation in voxels, regardless of the type of condition, varies randomly from trial to trial.

$$\begin{aligned} A_{tvs} &= \alpha_{vs} + X_s \beta_{vs} + \epsilon_{tvs}, \\ e_{tvs} &\sim \mathcal{N}(0, \sigma^2) \end{aligned} \tag{A.1}$$

Here, the data are summary statistics (e.g. LSS beta coefficients) representing the activation, A_{tvs}, observed on trial t, voxel v, and scan s. The variable X is an N_{trials} x $N_{covariates}$ design matrix and the observed activation is represented as a linear combination of the baseline activation (or intercept), α_{vs}, plus the product of the beta coefficients, β_{vs}, and X, plus trial-specific deviations, ϵ_{tvs} (A.1). These trial-level errors are assumed to follow a normal distribution with mean zero and variance σ^2.

The voxel-level model (A.2), describes how the multivariate patterns constitute repeated measurements across voxels that vary in two important respects: firstly, voxels vary in their mean baseline activation across trials of all types and secondly they vary in the effect of the experimental conditions.

$$\begin{aligned} \alpha_{vs} &= \alpha_s + \epsilon_{\alpha vs}, \\ \beta_{vs} &= \beta_s + \epsilon_{\beta vs}, \\ \begin{pmatrix} \epsilon_{\alpha vs} \\ \epsilon_{\beta vs} \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} \mu_{\alpha s} \\ \mu_{\beta s} \end{pmatrix}, \begin{pmatrix} \tau_\alpha^2 & \rho\tau_\alpha\tau_\beta \\ \rho\tau_\alpha\tau_\beta & \tau_\beta^2 \end{pmatrix} \right) \end{aligned} \tag{A.2}$$

This level of the model characterizes the entire population of voxels as having a mean baseline activity in each scan, α_s, and a mean effect of the experimental contrast in each scan, β_s. Voxel-specific deviations to each of these summary statistics are allowed by the inclusion of error terms, $\epsilon_{\alpha vs}$ and $\epsilon_{\beta vs}$ respectively. The regression parameters in Eq. (A.1), α_{vs} and β_{vs}, are therefore not fixed but conceptualized as random variables with a multivariate Gaussian probability distribution across voxels and scans. This probability distribution is summarized by the mean vector of coefficients, $\mu_{\alpha s}$ and $\mu_{\beta s}$, and the variance–covariance matrix. This matrix contains the between-voxel variances in both baseline, τ_α^2, and effect of the experimental contrast, τ_β^2, as well as their covariance, $\rho\tau_\alpha\tau_\beta$. The parameters τ_α and τ_β are of particular importance as they inherently model voxel-level variability and fit the common understanding that in any ROI there are, to greater or lesser extent, mixtures of both task-relevant and task-irrelevant voxels.

The third, and final, level of our model (A.3) accounts for the finding that there are often signal-related shifts in the mean activity of all trials within each run. There may be many causes of these run-level shifts including drifts in attention, changes in physiological arousal, or between-run differences in proportions of trial types.

$$\begin{aligned} \alpha_s &= \gamma + \epsilon_{\alpha s}, \\ \epsilon_{\alpha s} &\sim \mathcal{N}(0, \omega^2) \end{aligned} \tag{A.3}$$

The variance component of interest in this model corresponds to run-by-run variability in the mean activity of all trials across voxels. Like other levels, this is implemented by an error term, $\epsilon_{\alpha s}$, which quantifies each run's deviation from the expected value of all runs, α_s. These errors are also assumed to be normally distributed with mean zero and variance ω^2.

The combined equation (A.4) for activation A on trial t in voxel v for scan s is therefore a combination of fixed effects of the experimental variables as well as trial-, voxel-, and scan-level random effects:

$$A_{tvs} = \gamma + \epsilon_{\alpha s} + \epsilon_{\alpha vs} + X_s\beta_s + X_s\epsilon_{\beta vs} + \epsilon_{tvs} \tag{A.4}$$

In our own simulations, since we coded the two conditions in the design matrix as −0.5 and 0.5 (deviation coding scheme) this meant that the voxel-specific intercept represented that voxel's average (baseline) activation for trials of both conditions and the voxel-specific slope represented the effect of the experimental variables within that voxel. By sampling these slopes from a distribution with a mean slope of zero and varying the standard deviation of slopes, we simulated a context in which there is a range of effect sizes across the voxels (i.e. voxel's with more extreme slopes are more "informative" about the conditions of interest) but no univariate effect of the conditions.

Appendix B. Addressing violations to the assumption of equal variances in linear models

Human fMRI data

Prior to fitting the multi-level linear models (MLMs) on classification accuracies from the human fMRI data, Levene's test revealed significant departures from the assumption of equal variances between groups (i.e. heteroscedasticity) for the fixed factors of: trial averaging method ($F(2, 2864) = 333.01, p < 0.001$); run-wise mean centering ($F(2, 2865) = 127.21, p < 0.001$); and motion-related nuisance regression approach ($F(4, 2862) = 2.86, p = 0.022$). Therefore, the heteroscedasticity was included in the model by means of a variance function allowing different variances per stratum, computed as the ratio of each variance to a reference level. Specifying unique variances for fully crossed levels of trial averaging, mean centering and data cleaning was not computationally feasible due to the sheer number of levels and model convergence issues. Therefore, we specified unique variances for the two most critically heteroscedastic factors based on the magnitude of the F statistic from Levene's test: trial averaging method and run-wise mean centering.

Simulated data

For the simulated data, Levene's test revealed significant heteroscedasticity for the factors of trial averaging ($F(2, 8997) = 887.61, p < 0.001$), presence or absence of within-run mean centering ($F(1, 8998) = 867.54, p < 0.001$), and the cost selection method ($F(1, 8998) = 61.454$, p< 0.001). Therefore, unique variances were modeled for all three heteroscedastic factors.

Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jneumeth.2023.109808.

References

Abdulrahman, H., Henson, R.N., 2016. Effect of trial-to-trial variability on optimal event-related fMRI design: Implications for Beta-series correlation and multi-voxel pattern analysis. NeuroImage 125, 756–766.

Ambroise, C., McLachlan, G.J., 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl. Acad. Sci. USA 99 (10), 6562–6566.

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Stat. Surv. 4, 40–79.

Caballero-Gaudes, C., Reynolds, R.C., 2017. Methods for cleaning the BOLD fMRI signal. NeuroImage 154 (2016), 128–149.

Cortes, C., Vapnik, V., 2015. Natural vibration response based damage detection for an operating wind turbine via Random Coefficient Linear Parameter Varying AR modelling. Mach. Learn. 20 (3), 273–297.

Coutanche, M.N., Thompson-Schill, S.L., 2012. The advantage of brief fMRI acquisition runs for multi-voxel pattern detection across runs. NeuroImage 61 (4), 1113–1119.

Davis, T., LaRocque, K.F., Mumford, J.A., Norman, K.A., Wagner, A.D., Poldrack, R.A., 2014. What do differences between multi-voxel and univariate analysis mean? How subject-voxel-, and trial-level variance impact FMRI analysis. NeuroImage 97, 271–283.

De Martino, F., Valente, G., Ashburner, J., Goebel, R., Formisano, E., De Martino, F., 2008. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. NeuroImage 43 (1), 44–58.

Diedrichsen, J., Ridgway, G.R., Friston, K.J., Wiestler, T., 2011. Comparing the similarity and spatial structure of neural representations: A pattern-component model. NeuroImage 55 (4), 1665–1678.

Diedrichsen, J., Wiestler, T., Ejaz, N., 2013. A multivariate method to determine the dimensionality of neural representation from population activity. NeuroImage 76, 225–235.

Dinga, R., Penninx, B.W., Veltman, D.J., Schmaal, L., Marquand, A.F., 2019. Beyond accuracy: Measures for assessing machine learning models, pitfalls and guidelines. 743138, bioRxiv.

Etzel, J.A., 2015. MVPA permutation schemes: Permutation testing for the group level. In: Proceedings - 2015 International Workshop on Pattern Recognition in NeuroImaging. PRNI 2015, pp. 65–68.

Etzel, J.A., Cole, M.W., Zacks, J.M., Kay, K.N., Braver, T.S., 2016. Reward motivation enhances task coding in frontoparietal cortex. Cerebral Cortex 26 (4), 1647–1659.

Etzel, J.A., Gazzola, V., Keysers, C., 2009. An introduction to anatomical ROI-based fMRI classification analysis. Brain Res. 1282, 114–125.

Etzel, J.A., Valchev, N., Keysers, C., 2011. The impact of certain methodological choices on multivariate analysis of fMRI data with support vector machines. NeuroImage 54 (2), 1159–1167.

Friston, K.J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M.D., Turner, R., 1998. Event-related fMRI: Characterizing differential responses. NeuroImage 7 (1), 30–40.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.P., Frith, C.D., Frackowiak, R.S., 1994. Statistical parametric maps in functional imaging: A general linear approach. Hum. Brain Mapp. 2 (4), 189–210.

Friston, K.J., Williams, S., Howard, R., Frackowiak, R.S., Turner, R., 1996. Movement-related effects in fMRI time-series. Magn. Reson. Med. 35 (3), 346–355.

Frost, M.A., Goebel, R., 2012. Measuring structural-functional correspondence: Spatial variability of specialised brain regions after macro-anatomical alignment. NeuroImage 59 (2), 1369–1381.

Glover, G.H., 1999. Deconvolution of impulse response in event-related BOLD fMRI. NeuroImage 9, 416–429.

Goebel, R., Esposito, F., Formisano, E., 2006. Analysis of FIAC data with BrainVoyager QX: From single-subject to cortically aligned group GLM analysis and self-organizing group ICA. Hum. Brain Mapp. 27 (5), 392–401.

Hajnal, J.V., Myers, R., Oatridge, A., Schwieso, J.E., Young, I.R., Bydder, G.M., 1994. Artifacts due to stimulus correlated motion in functional imaging of the brain. Magn. Reson. Med. 31 (3), 283–291.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer.

Haynes, J.D., Rees, G., 2006. Decoding mental states from brain activity in humans. Nat. Rev. Neurosci. 7 (7), 523–534.

Huettel, S.A., McCarthy, G., 2001. The effects of single-trial averaging upon the spatial extent of fMRI activation. NeuroReport 12 (11), 2411–2416.

Jezzard, P., Balaban, R.S., 1995. Correction for geometric distortion in echo planar images from B0 field variations. Magn. Reson. Med. 34 (1), 65–73.

Kriegeskorte, N., 2011. Pattern-information analysis: From stimulus decoding to computational-model testing. NeuroImage 56 (2), 411–421.

LaConte, S., Strother, S., Cherkassky, V., Anderson, J., Hu, X., 2005. Support vector machines for temporal classification of block design fMRI data. NeuroImage 26 (2), 317–329.

Lee, S., Kable, J.W., 2018. Simple but robust improvement in multivoxel pattern classification. PLoS One 13 (11), 1–15.

Lemieux, L., Salek-Haddadi, A., Lund, T.E., Laufs, H., Carmichael, D., 2007. Modelling large motion events in fMRI studies of patients with epilepsy. Magn. Reson. Imaging 25 (6), 894–901.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2018. E1071: Misc functions of the department of statistics. In: Probability Theory Group (Formerly: E1071). TU Wien..

Mourão-Miranda, J., Reynaud, E., McGlone, F., Calvert, G., Brammer, M., 2006. The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. NeuroImage 33 (4), 1055–1065.

Mumford, J.A., Davis, T., Poldrack, R.A., 2014. The impact of study design on pattern estimation for single-trial multivariate pattern analysis. NeuroImage 103, 130–138.

Mumford, J.A., Turner, B.O., Ashby, F.G., Poldrack, R.A., 2012. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. NeuroImage 59 (3), 2636–2643.

Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fMRI. NeuroImage 56 (2), 400–410.

Nestor, A., Behrmann, M., Plaut, D.C., 2013. The neural basis of visual word form processing: A multivariate investigation. Cerebral Cortex 23 (7), 1673–1684.

Nestor, A., Plaut, D.C., Behrmann, M., 2011. Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. Proc. Natl. Acad. Sci. USA 108 (24), 9998–10003.

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V., 2006. Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. Trends in Cognitive Sciences 10 (9), 424–430.

Oakes, T.R., Johnstone, T., Ores Walsh, K.S., Greischar, L.L., Alexander, A.L., Fox, A.S., Davidson, R.J., 2005. Comparison of fMRI motion correction software tools. NeuroImage 28 (3), 529–543.

Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: A tutorial overview. NeuroImage 45 (1 Suppl), S199–S209.

Pinheiro, J., Bates, D., Debroy, S., Sarkar, D., Team, R.C., 2019. nlme: Linear and nonlinear mixed effects models.

Popov, V., Ostarek, M., Tenison, C., 2018. Practices and pitfalls in inferring neural representations. NeuroImage 174 (2017), 340–351.

Power, J.D., 2017. A simple but useful way to assess fMRI scan qualities. NeuroImage 154 (2016), 150–158.

Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., 2012. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. NeuroImage 59 (3), 2142–2154.

Power, J.D., Mitra, A., Laumann, T.O., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., 2014. Methods to detect, characterize, and remove motion artifact in resting state FMRI. NeuroImage 84, 320–341.

Prince, Jacob S., Charest, Ian, Kurzawski, Jan W., Pyles, John A., Tarr, Michael J., Kay, Kendrick N., 2022. Improving the accuracy of single-trial fMRI response estimates using GLMsingle. eLife (ISSN: 2050084X) 11, 1–28. http://dx.doi.org/10.7554/eLife.77599.

Satterthwaite, T.D., Elliott, M.A., Gerraty, R.T., Ruparel, K., Loughead, J., Calkins, M.E., Eickhoff, S.B., Hakonarson, H., Gur, R.C., Gur, R.E., Wolf, D.H., 2013. An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. NeuroImage 64 (1), 240–256.

Schaefer, A., Kong, R., Gordon, E.M., Laumann, T.O., Zuo, X.-N., Holmes, A.J., Eickhoff, S.B., Yeo, B.T.T., 2018. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. Cerebral Cortex 28 (9), 3095–3114.

Stehr, D.A., Zhou, X., Tisby, M., Hwu, P.T., Pyles, J.A., Grossman, E.D., 2021. Top-down attention guidance shapes action encoding in the pSTS. Cerebral Cortex 31 (7), 3522–3535.

TheMathWorksInc, 2017. Matlab.

Turner, B.O., Mumford, J.A., Poldrack, R.A., Ashby, F.G., 2012. Spatiotemporal activity estimation for multivoxel pattern analysis with rapid event-related designs. NeuroImage 62 (3), 1429–1438.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, New York.

Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., Thirion, B., 2017. Assessing and tuning brain decoders: Cross-validation, Caveats, and Guidelines. NeuroImage 145 (2015), 166–179.

Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., Nichols, T.E., 2014. Permutation inference for the general linear model. NeuroImage 92, 381–397.

Yan, C.G., Cheung, B., Kelly, C., Colcombe, S., Craddock, R.C., Martino, A.Di., Li, Q., Zuo, X.N., Castellanos, F.X., Milham, M.P., 2013. A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics. NeuroImage 76, 183–201.

Zeithamova, D., de Araujo Sanchez, M.A., Adke, A., 2017. Trial timing and pattern-information analyses of fMRI data. NeuroImage 153 (2016), 221–231.