NeuroImage 103 (2014) 130–138

Contents lists available at ScienceDirect

NeuroImage

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

The impact of study design on pattern estimation for single-trial multivariate pattern analysis

Jeanette A. Mumford [a,f,*], Tyler Davis [b], Russell A. Poldrack [c,d,e]

[a] Waisman Laboratory for Brain Imaging and Behavior, WI, USA
[b] Department of Psychology, Texas Tech University, Lubbock, TX, USA
[c] Department of Psychology, University of Texas at Austin, USA
[d] Department of Neuroscience, University of Texas at Austin, USA
[e] Imaging Research Center, University of Texas at Austin, USA
[f] Center for Investigating Healthy Minds at the Waisman Center, University of Wisconsin, Madison, WI, USA

ARTICLE INFO

Article history:
Accepted 11 September 2014
Available online 19 September 2014

Keywords:
fMRI
MVPA
Pattern similarity
Pattern classification
False positive rate

ABSTRACT

A prerequisite for a pattern analysis using functional magnetic resonance imaging (fMRI) data is estimating the patterns from time series data, which then are input into the pattern analysis. Here we focus on how the combination of study design (order and spacing of trials) with pattern estimator impacts the Type I error rate of the subsequent pattern analysis. When Type I errors are inflated, the results are no longer valid, so this work serves as a guide for designing and analyzing MVPA studies with controlled false positive rates. The MVPA strategies examined are pattern classification and similarity, utilizing single trial activation patterns from the same functional run. Primarily focusing on the Least Squares Single and Least Square All pattern estimators, we show that collinearities in the models, along with temporal autocorrelation, can cause false positive correlations between activation pattern estimates that adversely impact the false positive rates of pattern similarity and classification analyses. It may seem intuitive that increasing the interstimulus interval (ISI) would alleviate this issue, but remaining weak correlations between activation patterns persist and have a strong influence in pattern similarity analyses. Pattern similarity analyses using only activation patterns estimated from the same functional run of data are susceptible to inflated false positives unless trials are randomly ordered, with a different randomization for each subject. In other cases, where there is any structure to trial order, valid pattern similarity analysis results can only be obtained if similarity computations are restricted to pairs of activation patterns from independent runs. Likewise, for pattern classification, false positives are minimized when the testing and training sets in cross validation do not contain patterns estimated from the same run.

Introduction

Traditional data analysis approaches in functional magnetic resonance imaging (fMRI) often employ voxel-wise models to identify where in the brain aggregate activation differs between experimental conditions. A more recently developed set of analysis strategies, multivoxel pattern analysis (MVPA), often starts with similar voxel-wise activation estimates, but instead of testing for differences in aggregate activation, focuses on the information contained in the distributed patterns of activation across voxels (Kriegeskorte et al., 2008a; Kriegeskorte, 2011; Haxby et al., 2001; Carlson et al., 2003; Pereira et al., 2009; Norman et al., 2006; Davis and Poldrack, 2013; Haynes and Rees, 2006). Multivariate pattern classification and pattern similarity analyses are two of the most common MVPA strategies. Pattern classifiers test whether an activation pattern can be used to decode the mental state of the subject (Haynes and Rees, 2006; Norman et al., 2006). In pattern similarity analyses, the goal is often not to simply decode mental states, but to examine the geometric relationships between activation patterns for different conditions and stimuli in a task. To this end, pattern similarity analysis involves computing a similarity metric between pairwise activation patterns elicited for different conditions or stimuli, and testing how these pattern similarities relate to psychological states, predictions from cognitive models, or patterns elicited for the same stimuli in non-human primates (Kriegeskorte et al., 2008a; Kriegeskorte, 2011; Kriegeskorte et al., 2008b; Kriegeskorte and Kievit, 2013; Davis and Poldrack, 2013). Despite the increasing popularity of MVPA approaches to fMRI analysis, there have been few systematic studies of how study design impacts results from MVPA.

The present work examines how study design affects estimation of the activation patterns that serve as inputs into MVPA and subsequent Type I error rates. We focus on experimental contexts in which the goal is to accurately estimate separate activation patterns from single

* Corresponding author at: Center for Investigating Healthy Minds at the Waisman Center, University of Wisconsin-Madison, 1500 Highland Ave, Madison, WI 53705.
E-mail address: jeanette.mumford@gmail.com (J.A. Mumford).

http://dx.doi.org/10.1016/j.neuroimage.2014.09.026
1053-8119/

trials within the same run. For example, if 30 exemplars of each of 2 types of stimuli are presented to a subject in a single functional run, the goal is to estimate 60 separate activation patterns that are then input data for a pattern analysis that will attempt to classify or explain the similarity relationships between these exemplars. Single run analyses are common within the pattern similarity framework and have the advantage of saving time and money while collecting data. Although previous work assessed power of single trial parameter estimators within the pattern classification framework in a between-run setting (Turner et al., 2012; Mumford et al., 2012), control of Type I error is more critical since, when not controlled, the resulting statistics are invalid.[1] In the case of pattern similarity, Type I error rates will be quantified for analyses that compare similarity distributions for different pairings of trials from the same run. For pattern classification, Type I error is assessed by testing whether or not the classification accuracy of data generated under the null hypothesis is at chance when the cross validation is performed using trials from the same run.

The primary pattern estimators we examine are the Least Squares All (LSA) and Least Squares Single (LSS) models (Turner et al., 2012; Mumford et al., 2012). Both of these models estimate patterns using a voxelwise general linear model; example design matrices for a single run that presented 5 exemplars each of 2 stimulus types are illustrated in Fig. 1. In the case of LSA, all trials are estimated simultaneously in a single model, using a separate regressor consisting of an impulse (or boxcar) function convolved with a double gamma hemodynamic response function (HRF). This is often referred to as beta-series regression (Rissman et al., 2004) and the parameters, β_1, …, β_{10}, estimate the activation magnitude for each of the 10 trials within a single voxel. These estimates are then aggregated over many voxels to comprise the activation pattern that serves as the input for MVPA. A pitfall of LSA is when trials have a short interstimulus interval (ISI), e.g., less than 3 s between the end of one stimulus and onset of the next stimulus, the regressors become highly correlated, or collinear, which inflates the variance of the resulting parameter estimates. The LSS model reduces this collinearity by using a separate model for each trial, in which the first regressor models the trial of interest and the other two regressors model the remaining trials according to trial type. For example, assuming that the exemplars were images of mammals or reptiles, the first iteration of LSS is modeling the first trial, a mammal, as the first regressor while the other two regressors model the remaining mammals and remaining reptiles, respectively. In this case, only the first parameter estimate is retained in each model and estimates the activation for that individual trial. Previously, LSS has been shown to produce higher classification accuracies than LSA for short ISIs (3–5 s) (Mumford et al., 2012). Although we focus on the LSS and LSA pattern estimators, a third model that simply takes the time point 6 s after stimulus presentation as the pattern estimate (Add6) is also considered.

We will illustrate how temporal autocorrelation and pattern estimation technique, through correlations between regressors, introduce false positive correlations between the activation patterns estimated from the same functional run of BOLD data. These false correlations can then lead to false positive comparisons in pattern similarity distributions or inflated classification accuracies. Surprisingly, even with ISIs as long as 15 s, elevated false positive rates occur in pattern similarity analyses, regardless of which of the three pattern estimators (LSS, LSA, Add6) are used, unless trials are randomly ordered with a unique randomization for each subject. Similar issues arise if the cross validation of a pattern analysis uses only trials from the same functional run. Hence, within-run pattern similarity and classification analyses are not recommended.

After characterizing the pitfalls of single run, or within-run analyses, we examine whether or not between-run analyses offer reasonable solutions. Between-run analyses require multiple functional runs of BOLD data and, in the pattern similarity setting, similarities are only computed between activation patterns that were estimated from independent runs of data. For example, if there are 60 trials per run and two runs, the pattern from the first trial of run 1 would be correlated with all other patterns from run 2, only, in a between-run similarity analysis. For classification, the test and training data sets in the cross validation would comprise patterns from different runs.

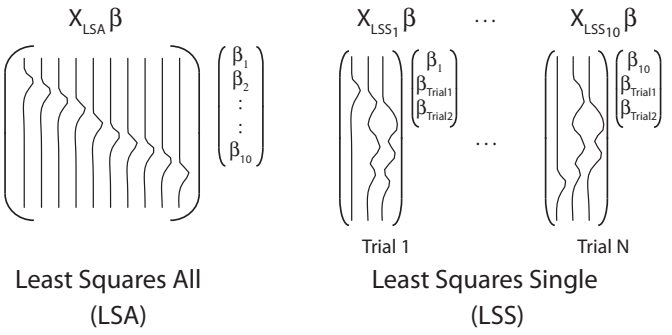

Fig. 1. Model illustration for LSA and LSS. In both cases, trial-specific activations are estimated for each of 10 trials and the model is run in a voxel-wise fashion. The left panel shows Least Squares All (LSA), which estimates all trials simultaneously in a single regression and the estimates β_1, …, β_{10} represent the activation magnitudes for each of the trials. The right panel shows Least Squares Single (LSS) where each trial's activation is estimated in a separate model where the first regressor represents the trial of interest and the two additional regressors model the remaining trials according to trial type. In this case there are 2 trial types. Only the estimates for the first parameter are retained from each model.

The following section contains a theoretical derivation of the LSS- and LSA-based pattern estimates to clearly motivate the impact of study design and temporal autocorrelation on estimated patterns. Both the Methods and Results sections start with pattern similarity analyses and end with classification analyses. Within each analysis setting, within-run approaches are studied first, followed by between-run approaches. This work can be used as a guideline when designing and implementing future MVPA analyses that will yield valid results.

Derivations

The following derivations were used to generate data for the simulation studies and help in understanding the source of invalid statistics for within-run pattern analyses. Before deriving the distributions for the LSS- and LSA-based pattern estimates, we first characterize the distribution of the BOLD time series, Y, within a single voxel. The data follow a multilevel structure where one level (Eq. (2)) describes the trial-specific activations, β, and the other level (Eq. (1)) describes how these trial-specific activations are related to the BOLD time series. Specifically,

$$Y = X_{LSA}\beta + \epsilon_Y, \quad \epsilon_Y \sim N(0, V_Y) \tag{1}$$

$$\beta = \mu + \epsilon_\beta, \quad \epsilon_\beta \sim N\left(0, V_\beta\right). \tag{2}$$

Assume that there are N_{trials} total trials presented, β is a vector of length N_{trials} with a true mean of μ, also a vector of length N_{trials}, and a covariance following the $N_{trials} \times N_{trials}$ matrix, V_β. Since V_β is the true covariance between the trials, it represents the true representational similarity covariance matrix, from which the pattern similarity correlations can be derived. The vector Y is the voxel-wise time series with length N_{tpts} and assumes a mean equal to the product of the trial-specific activation magnitudes, β, and the trial-specific regressors following the LSA design matrix (Fig. 1, left panel). The LSA design matrix consists of a single regressor per trial, where the regressor is constructed by convolving a delta or boxcar function, depending on

[1] In these previous studies, Type I error was assessed and found to be controlled, but this was not reported.

the duration of the trial, with a double gamma hemodynamic response function. This approach specifically assumes that the LSA model is a correct representation of the BOLD time series. The time points are assumed to be temporally autocorrelated following a covariance matrix, V_Y.

Combining the two levels shown in Eqs. (1) and (2) yields,

$$Y = X_{LSA}\mu + X_{LSA}\epsilon_\beta + \epsilon_Y, \tag{3}$$

from which the variance of Y is,

$$\mathrm{Var}(Y) = X_{LSA}V_\beta X'_{LSA} + V_y. \tag{4}$$

The following subsections derive the theoretical distributions of the LSA- and LSS-based similarities using this assumed distribution of Y.

Pattern distribution: LSA

The LSA-based model for estimating trial-specific activation magnitudes within a voxel, is given by $Y = X_{LSA}\beta_{LSA} + \epsilon_Y$, from which the trial-specific parameter estimates are given by

$$\hat{\beta}_{LSA} = (X'_{LSA}X_{LSA})^{-1}X'_{LSA}Y. \tag{5}$$

The estimates, $\hat{\beta}_{LSA}$, are voxelwise estimates of the trial-specific activation and are aggregated over multiple voxels to produce the input patterns for MVPA. The true similarity between the estimated patterns of all pairs of trials can be derived from Eqs. (4) and (5) and is given by,

$$\begin{aligned}\mathrm{Var}\left(\hat{\beta}_{LSA}\right) &= (X'_{LSA}X_{LSA})^{-1}X'_{LSA}\mathrm{Var}(Y)X_{LSA}(X'_{LSA}X_{LSA})^{-1}\\ &= V_\beta + (X'_{LSA}X_{LSA})^{-1}X'_{LSA}V_yX_{LSA}(X'_{LSA}X_{LSA})^{-1}.\end{aligned} \tag{6}$$

In the special case where the BOLD time series are uncorrelated, $V_y = \sigma_y^2 I$, where σ_y^2 is the variance and I is a $Ntpts \times Ntpts$ identity matrix, this estimated variance reduces to

$$\mathrm{Var}\left(\hat{\beta}_{LSA}\right) = V_\beta + \sigma_y^2(X'_{LSA}X_{LSA})^{-1}. \tag{7}$$

Recall that V_β is the true covariance between the activation patterns, from which the pattern similarity correlations can be derived and therefore the LSA-based covariances shown above will be impacted by both the inherent temporal autocovariance of the BOLD time series, V_y, and the LSA design matrix, X_{LSA} (Eqs. (6) and (7)).

Pattern distribution: LSS

The LSS algorithm uses a separate GLM to estimate the pattern for each trial, as shown in Fig. 1. Assume that there are two trial types of interest with multiple presentations of exemplars of each trial type. For the i^{th} trial, the model is $Y = X_{LSS_i}\beta_{LSS_i} + \epsilon_i$, where X_{LSS_i} has 3 regressors: one modeling the i^{th} trial, one modeling all other type 1 exemplars and the last regressor models all other type 2 exemplars. The estimate for the first trial is given by

$$\hat{\beta}_{LSS_i,1} = c\left(X'_{LSS_i}X_{LSS_i}\right)^{-1}X_{LSS_i}Y, \tag{8}$$

where c is the row vector, [1, 0, +0]. Since each trial's estimate is obtained by multiplying the $1 \times N_{tpts}$ vector, $c\left(X'_{LSS_i}X_{LSS_i}\right)^{-1}X_{LSS_i}$, by Y, all LSS-based trial estimates can simultaneously be estimated using

$$\hat{\beta}_{LSS} = X_{LSS}Y, \tag{9}$$

where

$$X_{LSS} = \begin{pmatrix} c\left(X'_{LSS_1}X_{LSS_1}\right)^{-1}X_{LSS_1} \\ c\left(X'_{LSS_2}X_{LSS_2}\right)^{-1}X_{LSS_2} \\ \vdots \\ c\left(X'_{LSS_{Ntrials}}X_{LSS_{Ntrials}}\right)^{-1}X_{LSS_{Ntrials}} \end{pmatrix}. \tag{10}$$

Combining this with the variance of Y given in Eq. (4) yields,

$$\begin{aligned}\mathrm{Var}\left(\hat{\beta}_{LSS}\right) &= X_{LSS}\mathrm{Var}(Y)X'_{LSS}\\ &= X_{LSS}X_{LSA}V_\beta X'_{LSA}X'_{LSS} + X_{LSS}V_yX'_{LSS}.\end{aligned} \tag{11}$$

In the special case where $V_y = \sigma_y^2 I$,

$$\mathrm{Var}\left(\hat{\beta}_{LSS}\right) = X_{LSS}X_{LSA}V_\beta X'_{LSA}X'_{LSS} + X_{LSS}X'_{LSS}\sigma_y^2. \tag{12}$$

Just as in the case of LSA, the LSS pattern estimate's covariance is not equal to the true covariance associated with the true similarity, V_β. Instead, both the design matrix used to obtain the estimate and the covariance of the fMRI time series have an impact on the estimated pattern similarity matrices.

Methods

Pattern similarity: within-run

The impact of pattern estimator and study design, assuming no temporal autocorrelation, can be studied using Eqs. (7) and (12). The ISI and stimulus order will drive different effects in the similarity estimates. We assumed two trial types of interest ($t1$ and $t2$) for which multiple exemplars of each (either 22 or 42) were presented during the study, resulting in similarity matrix composed of within-trial-type similarities, comparing patterns from two exemplars of the same type ($wt1$ or $wt2$) or between-trial-type similarities, comparing different trial types ($bt1t2$). The hypothesis of interest is whether within-trial-type similarities differ from each other or from between-trial-type similarities. This is tested by computing the mean for each of the 3 similarity types, within-subject, and then using a paired t-test across subjects to test the three hypothesis of interest: $wt1$–$wt2$, $wt1$–$bt1t2$ and $wt2$–$bt1t2$. The time series variance, σ_y^2, was set to 1 and the matrices, X_{LSA} and X_{LSS} were varied according to different lengths of interstimulus interval (ISI) and trial presentation ordering.

The ISIs were generated from a truncated exponential distribution, where the rate parameter, λ, was set to 1.5 and the upper truncation limit was set to 3 s. This created ISIs ranging between 0 and 3 s with a mean of 1.03 s. The ISI range was shifted by adding a scalar value, k, so the ISI ranged between k and $k + 3$ with a mean of $k + 1.03$. A shorter ISI with mean of 3.03 s ($k = 2$) and a longer ISI with mean of 7.03 s ($k = 6$), were used in the simulations. Later on these simulations are extended to include temporal covariance estimates based on real data with 225 time points (TR = 2s), which limits the number of trials that can be presented within a single run and so for the sake of continuity, runs with at most 225 TRs were used in all simulations. When using an average ISI of 3.03 s, 42 exemplars of each trial type (84 trials) were used and with an ISI of 7.03 s, 22 exemplars of each trial type were used. The following trial orderings were considered: blocked (all exemplars of trial 1, followed by all trial 2), alternating (type 1, type 2, type 1, type 2, …), and randomly ordering trials within a scanning run. In the blocked and alternating trial orderings, the starting trial type was randomly chosen for each run. For each simulated subject, assuming that the true similarity was the identity (V_β), a design matrix was randomly generated and Eqs. (7) and (12) were used to compute the similarity matrices.

To compute Type 1 error rates, 10,000 data sets of 30 subjects were randomly generated, using a different set of ISIs and randomly ordered

trials, when applicable, for each subject. Within each set of 30 subjects the p-values for the three pairwise comparisons described above were computed and used to estimate the Type I error across the 10,000 simulations.

Temporal covariance will also impact the estimated similarity according to Eqs. (6) and (11). To produce realistic simulations, we used temporal covariances based on real resting state fMRI data. The values used for V_y were estimated from 198 resting state data sets for the same ROI, a randomly chosen $7 \times 7 \times 7$ voxel cube in standard MNI space. Results were similar for different ROIs of the same size. The Zhang data from the 1000 Functional Connectomes project were used (http://www.nitrc.org/projects/fcon_1000/). There were 198 subjects (122 female) between 18–26 years of age. The TR was 2 s long and the time series contained 225 time points. FMRI data processing was carried out using FEAT (FMRI Expert Analysis Tool) Version 6.00, part of FSL (FMRIB's Software Library, www.fmrib.ox.ac.uk/fsl). The following pre-statistics processing was applied; motion correction using MCFLIRT (Jenkinson et al., 2002); non-brain removal using BET (Smith, 2002); spatial smoothing using a Gaussian kernel of FWHM 5 mm; multiplicative mean intensity normalization of the volume at each time point (i.e., global signal regression); highpass temporal filtering (Gaussian-weighted least-squares straight line fitting, with sigma = 50.0 s). Image registration to high resolution structural space images was carried out using FLIRT (Jenkinson and Smith, 2001; Jenkinson et al., 2002). The BBR (boundary based registration) algorithm was used to register the BOLD data to the subject's structural image (Greve and Fischl, 2009) and a 12 degree of freedom affine transformation was used to register the subject's structural image to MNI space. The 6 motion parameters, their squared values and the derivatives of each were modeled as nuisance regressors and resulting residual images were used to compute within-subject temporal covariance estimates for a randomly placed cubic ROI of dimension $7 \times 7 \times 7$ voxel^3 ($2 \times 2 \times 2$ mm^3 voxels). The ROI's center was located in the Right Putamen (MNI coordinates: 18 mm, 12 mm, −6 mm).

Estimated similarity matrices were generated using Eqs. (6) and (11), again assuming that V_β was an identity matrix. The value of V_y for each simulated subject was a temporal covariance estimate from a randomly chosen single resting state data set. The design matrix was varied as in the previous simulation by using average ISIs of 3.03 or 7.03 s and the three trial orderings of interest (blocked, alternating and random). Type I error rates were also derived as described in the previous section, across 10,000 simulated data sets with 30 subjects each. Importantly, in the randomly ordered trial simulation, the trial order was randomly generated for each subject. Thus we added an additional simulation where the same design setup was used for all subjects within a single simulation to study whether Type I error rates for randomly ordered trials required a different randomization for each subject to provide valid results.

One last model was considered to see if issues that arose with LSS and LSA within the blocked and alternating trial orders could be resolved. This model simply takes the time point 6 s after the stimulus was presented as the pattern estimate and is referred to as the Add6 model (Mumford et al., 2012). This model is most related to the LSA approach, but instead of using HRF convolution to model the time course, the trial specific regressor would be a boxcar delayed 6 s from the trial onset with a width of 1 TR, which is designed to capture the activation peak for a trial. This model was compared to LSS and LSA with an ISI of 15 s with 12 exemplars of each trial type.

Pattern similarity: between-run

Lastly, we studied the Type I error rates when similarities were computed between-run. Each subject had 2 separate runs of data for which patterns were estimated and similarities were constructed only between pairs of patterns from different runs. In this case, Eq. (2) was used to simulate activation magnitudes for 500 voxels and the values for β were used to simulate time series following Eq. (1). The true similarity covariance, V_β, was set to the identity matrix and the mean trial activation, μ, was set to a vector of zeros. As described in the previous section, temporal covariances were derived from resting state data and used for V_Y. Patterns were then estimated from the simulated time series using the LSA or LSS models. The two runs, within-subject, either had the same trial ordering and ISIs, or different. The mean ISI value was 3.03 s, using the above described truncated exponential distribution.

Classification analysis: within- and between-run CV

Simulations were also run to quantify the impact of model- and data-based noise on classification analyses. Time series data were generated and patterns were simulated as described in the between-run analysis above. Simulated data using a mean ISI of 3.03 s and 7.03 s were created. The patterns were then used in a support vector classification analysis where a 2-fold cross validation (CV) was used to compute classification accuracy, within-subject. The svm function in the e1071 library in R was with a linear kernel and cost parameter of 1. The three types of trial orderings described earlier were used: blocked, alternating and random. Three types of within-subject data subsets were used in a 2-fold CV: Within-run (WR), between-run using the same stimulus ordering and ISIs (BR(Same)) and between-run with different ISIs and stimulus order, in the case of randomly ordered trials, were used (BR(Diff)). In the WR CV, trials were randomly split into 2 groups, and in the BR settings, trials were grouped according to run in the CV. In the BR case, different estimates of temporal covariance, V_Y, were used for each run. Data for 1000 subjects were generated and the distributions of the classification accuracies were used to compare the methods. Additionally, the within-run CV results when the same trial order randomization was used across subjects were studied. In this case the results are based on a data set of 30 subjects, which is representative of results found across many randomly generated data sets.

Code for running all the simulations described here and for reproducing the figures and tables within the results section can be found at https://bitbucket.org/jmumford/scripts_mumford_davis_poldrack_ni_2014/src.

Results

Pattern similarity: within-run

The first simulation focused on the impact of pattern estimator model on similarity estimates, assuming that data are not temporally correlated. Fig. 2 illustrates the true and estimated similarities using LSS or LSA. Trials were blocked (e.g. 42 exemplars of trial type 1 followed by 42 of trial type 2) with an average ISI of 3.03 s. The rows and columns of the similarity matrices are ordered according to the temporal order of the trials such that the magnitude of the cell in the 3rd row and 4th column is the correlation between the 3rd and 4th trials. The first off-diagonal of the similarity matrix corresponds to correlations in patterns of temporally adjacent trials, or the lag 1 similarities, and the following off-diagonals are lag 2, 3, etc. The LSA similarity matrix has a strong negative lag 1 correlation, followed by alternating positive and negative correlations over subsequent lags. With LSA (Fig. 1) each trial's regressor is highly correlated with its temporal neighbors' regressors resulting in collinearity, which does not impact the fit of the model, but produces highly variable parameter estimates. Therefore, one parameter estimate will be elevated and, to preserve the model fit, the collinear counterparts' parameter estimates will be pushed in the opposite direction. This causes a negative lag 1 correlation and at lag 2 the two trials will be pushed in the same direction by their common collinear neighbor, causing a positive correlation.

With LSS there are two patterns in the similarity matrix: two blocks along the diagonal and strong positive correlations for early lags. The

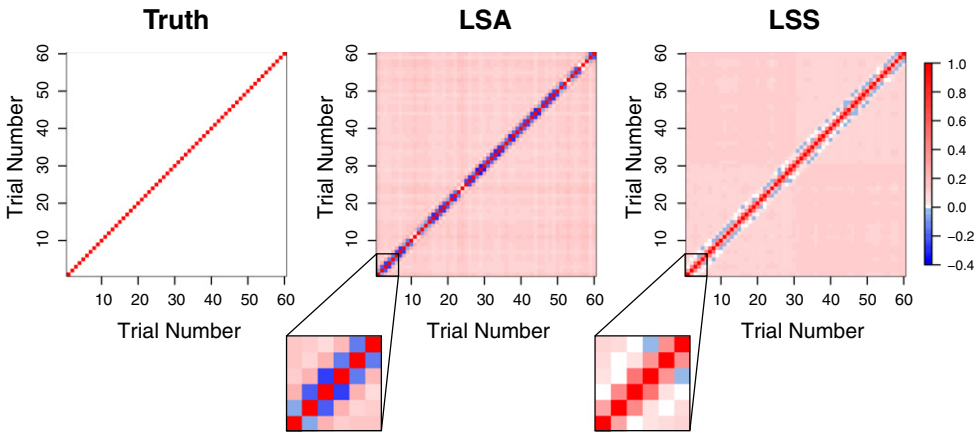

Fig. 2. Impact of collinearity in LSS and LSA models on similarity estimates. Each matrix displays true (left panel) or estimated (right two panels) similarities between all pairs of 60 trials, where the LSA and LSS estimates are based on Eqs. (7) and (12), respectively. There were two trial types presented in a blocked fashion: 30 exemplars of trial type 1, followed by 30 exemplars of trial type 2. The LSA and LSS plots include a blown up image of the lower 8 × 8 section of the matrix. Although the true off-diagonal similarities are all 0, collinearity in the LSA model causes a strong negative correlation between adjacent trials (blue off diagonal) and collinearity in the LSS model causes the large blocked pattern (2 faintly lighter blocks along the diagonal). LSS exhibits strong positive correlations between temporally close trials (red first off-diagonal).

large blocks follow the trial type groupings and imply that patterns of the same trial type are less similar than trials of different types. With the LSS model (Fig. 2), a weak collinearity occurs if the neighbors of a trial of interest are exemplars of the same category. With blocked trial order, almost all exemplars of type 1 have type 1 neighbors, causing a weak collinearity that results in negatively biased similarity estimates between each type 1 and all other type 1 exemplars (likewise for type 2 exemplars). This will also occur in the alternating trial type presentation, with an opposite effect, since each exemplar of trial type 1 will have type 2 neighbors, hence similarities between type 1 and type 2 exemplars will be negatively biased. Lastly, the strong positive similarities for early lags is a result of each trial's estimate coming from an independent model, thus each trial's estimate is not adjusted for the neighboring trials and neighboring trial estimates reflect a similar variability in the data. Notably, the impact of collinearity will diminish as ISI increases, but a weak negative correlation will develop between regressors. This weak effect will be shown to have a smaller, but opposite, impact on pattern similarities and is discussed below.

The top row of Fig. 3 shows distributions of within-subject similarity differences over sets of 30 simulated subjects, across all trial orderings for both LSS and LSA for the 2 ISI distributions. For the sake of brevity in table and figure labels, 3.03 and 7.03 s ISIs are rounded to 3 and 7,

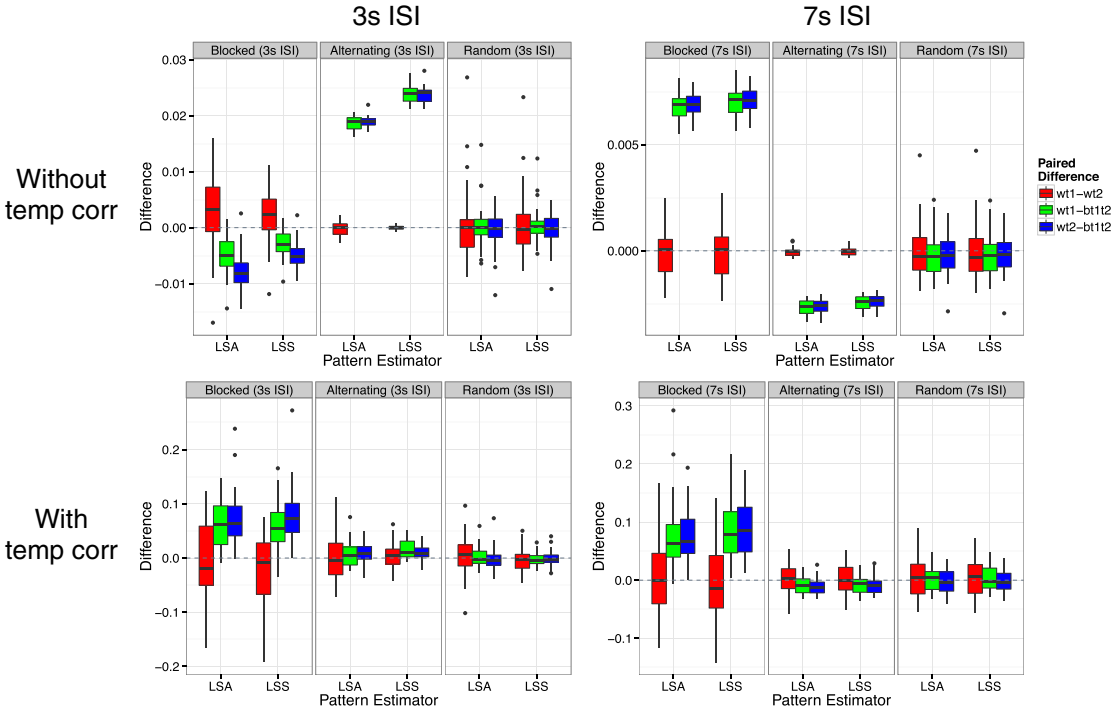

Fig. 3. Distributions of paired similarity differences across 30 simulated subjects for each simulation setting. Similarities are grouped according to exemplar pairs of the same type (*wt*1 and *wt*2) and opposite types (*bt*1*t*2), which are averaged, within-subject, and differenced. Results in the top row assume no temporal autocorrelation, whereas results in the bottom row assume temporally correlated data. The left and right columns use ISI distributions with means of 3.03 s a 7.03 s, respectively. Distributions that are centered about 0 indicate that the comparison is likely valid. See Table 1 for Type I error rates.

Table 1
Type 1 error rates across simulations. Pattern similarities are either between trials within the same run (WR) or between runs only (BR). Temporal correlation is assumed to be identity matrix (Ind) or follows a temporal covariance structure based on resting state data (Corr). Lastly, in the between-run cases, either the same trial order and ISIs were used (same) or were generated separately (different) for each run.

		Blocked			Alternating			Random		
		wt1–wt2	*wt1–bt1t2*	*wt2–bt1t2*	*wt1–wt2*	*wt1–bt1t2*	*wt2–bt1t2*	*wt1–wt2*	*wt1–bt1t2*	*wt2–bt1t2*
WR/Ind (3 s ISI)	LSA	0.050	1	1	0.051	1	1	0.048	0.049	0.049
	LSS	0.049	1	1	0.052	1	1	0.048	0.049	0.048
WR/Ind (7 s ISI)	LSA	0.049	1	1	0.051	1	1	0.048	0.047	0.052
	LSS	0.049	1	1	0.050	1	1	0.048	0.048	0.052
WR/Corr (3 s ISI)	LSA	0.047	1	1	0.055	0.489	0.497	0.053	0.054	0.052
	LSS	0.049	1	1	0.050	0.965	0.966	0.052	0.053	0.052
WR/Corr (7 s ISI)	LSA	0.049	1	1	0.051	0.755	0.743	0.049	0.055	0.057
	LSS	0.053	1	1	0.050	0.429	0.429	0.051	0.053	0.055
WR/Corr (15 s ISI)	LSA	0.048	0.862	0.865	0.050	0.244	0.241	0.053	0.059	0.060
	LSS	0.050	0.888	0.892	0.050	0.218	0.216	0.052	0.053	0.058
	Add6	0.051	0.498	0.503	0.052	0.079	0.079	0.049	0.057	0.058
BR/Corr (same) (3 s ISI)	LSA	0.048	0.047	0.052	0.047	0.048	0.049	0.056	0.048	0.052
	LSS	0.048	0.050	0.048	0.050	0.050	0.051	0.050	0.051	0.050
BR/Corr(different) (3 s ISI)	LSA	0.051	0.050	0.049	0.052	0.051	0.052	0.052	0.049	0.049
	LSS	0.045	0.045	0.046	0.050	0.051	0.050	0.051	0.049	0.048

respectively. Difference distributions with a mean of zero indicate a valid comparison and all corresponding Type I error rates are reported in Table 1. Although the pattern similarity matrices for LSA- and LSS-estimated patterns look very different (Fig. 2), the comparisons corresponding to our hypotheses of interest are similar. Regardless of ISI distribution and pattern estimator used, the blocked and alternating trial orders yield false positive differences when comparing similarities within a single trial type to similarities between trial types. The only valid comparisons for the blocked and alternating trial orders involve within-trial-type similarity comparisons (*wt1–wt2*). When trials are randomly ordered, employing a unique randomization for each subject, all three comparisons are valid. Notably, even when the ISI was increased to 15 s, the false positive rates remained similar to the 7 s and 3 s ISIs (Table 1).

The next set of simulations focuses on the added impact of temporal autocorrelation on false positive rates for within-run analyses. The bottom row of Fig. 3 shows the results of pairwise similarity comparisons when the temporal covariance structure is added. Although the addition of the temporal covariance seems to reduce some of the impact caused by trial order and ISI on the estimated similarities, the pattern of valid and invalid comparisons remains the same (Table 1). Only within-trial-type comparisons (*wt1–wt2*) are valid for blocked and alternating designs, while all pairwise comparisons are valid when the trial type and ISIs are randomly generated for each individual subject.

Table 1 shows LSS and LSA for a long ISI of 15 s as well as the Add6 model (WR/Corr (15 s ISI)). Although the Add6 model shows some improvement over LSS and LSA, the same comparisons remain invalid even with this long ISI.

Although the results so far support the use of randomly ordered trials, those simulations assumed that a different trial order and a set of ISIs were generated for each subject. Often it is the case that random trial orders are only random across runs, within-subject, but the same trial orderings are used for all subjects. For example, in designs for univariate voxel-wise analysis, investigators will often generate a single or small set of optimized sequences that are used in every subject. In this case, for each set of 30 subjects in an iteration of the simulation, the same trial order and ISIs were used. If, for example, the average amount of fixation around trials of type 1 is slightly larger, the similarities between two trials of type 1 (*wt1*) would be expected to be larger than either *wt2* or *bt1t2*, due to noisier pattern estimates for trials of type 2, leading to a false positive difference. Table 2 shows that for both estimators and all comparisons, if the same trial order is used for all subjects, the false positive rate is at least 0.29.

Pattern similarity: between-run

The impact on pattern similarities in the above simulations is driven either by collinearities in the pattern estimator model or temporal covariance. Using pattern similarities derived from patterns in different runs should alleviate these problems. Indeed, as the bottom two rows of Table 1 illustrate, Type I error is preserved across all trial ordering and both pattern estimators when patterns are only compared between-run.

Classification analysis

The last set of simulations was designed to generalize our conclusions about the impact of collinearities due to trial order and ISI on pattern similarity to MVPA analysis strategies employing pattern classifiers. In all cases, a 2-fold, within-subject, cross validation with a SVM was used, where the data were either split within-run (WR) or between-run (BR). In all cases the patterns were random, so the true classification accuracy is chance, 0.5. In the BR case the trial order and ISIs were either the same for both runs or different (BR (same)/BR (diff)). Fig. 4 illustrates the classification accuracy distributions over 1000 simulated, single subject data sets. Since pooled classification accuracies can often be misleading, the within-trial-type classification accuracies are shown in each case. The top row uses an average ISI of 3.03 s and bottom row 7.03 s. The LSA-related plots are in red and magenta and LSS are in blue and cyan. Starting with the blocked trial design, the within-run CV leads to biased classification accuracies for both ISIs. However, between-run CV yielded classification accuracies that are not different from chance for both ISIs. When trials alternated, there is a clear issue with the WR CV with a short ISI, but the impact lessens when the ISI increases. The results using randomly ordered trials appear to be stable across all ISIs and CV settings. Only between-run cross validations are valid for all trial orderings and both ISI settings.

Just as it was important that trials were randomly ordered in the pattern similarity analyses, the within-run CV behaves poorly if the

Table 2
False positive rates when trial randomization is fixed across all subjects within a study. Note, a different randomization was used for each set of 30 subjects in the simulation.

		Random (same for all subjects)		
		wt1–wt2	*wt1–bt1t2*	*wt2–bt1t2*
WR/Corr (3 s ISI)	LSA	0.606	0.580	0.593
	LSS	0.308	0.293	0.291

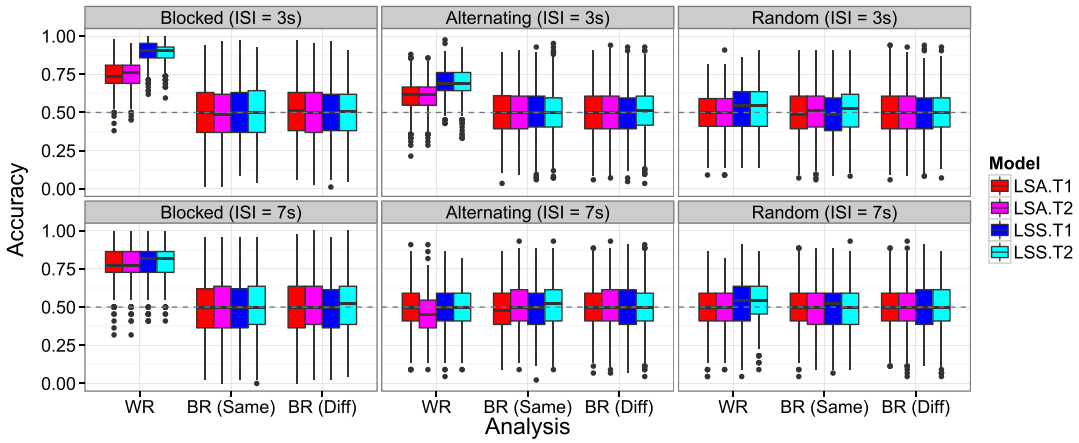

Fig. 4. Classification accuracy distributions across 1000 simulated data sets. Accuracy distributions are separated according to trial type and pattern estimator. The distributions should be centered about 0.5. The three cross validations considered were all 2-fold CV where it is either: within-run where trials were randomly grouped into 2 sets (WR), between-run where the same trial presentation was used for both runs (BR(Same)), or between-run where different trial presentations were used for each run (BR(Diff)).

same randomization is used for all subjects in the sample. Fig. 5 illustrates the within-trial-type accuracy distributions for a single, representative set of 30 simulated subjects. Although the pooled classification accuracy is near chance, the individual trial accuracies are either over- or under-estimated.

Discussion

Overview: pattern similarity

This work highlights the importance of stimulus order and ISI within the context of pattern classification and similarity analyses when trying to preserve the Type I error rate. A common tradeoff in study design for pattern analyses is to increase the ISI, which often decreases the number of stimuli and/or lengthens the scanning time. The motivation behind increasing the ISI is that the trial-specific activation patterns will have more power and will be more independent from each other, due to a reduction in the impact of temporal autocorrelation and collinearity in the pattern estimator model. This work shows that, under the null condition, even large ISIs do not guarantee independence between the pattern estimates and this can drive false positive differences when comparing distributions of pattern similarities for different pairings of stimuli from the same functional run. The serious issues with the Type I error rate at ISIs of 3 second long persist with ISIs as long as 15 s. If there is any structure in the trial presentation order, for example blocked, alternating or randomly ordered trials where the same random order is used across all subjects, false positive differences between pattern similarity distributions can result. Notably, this problem occurs with all three pattern estimators considered here: LSA, LSS and Add6. In terms of preserving the Type I error rate when running a within-run pattern similarity analysis, there is no need to shorten the ISI in a pattern similarity analysis, but the most important factor is to randomly order the trials with a different randomization for each subject, as this is this only way to preserve Type I error (Fig. 3 and Table 1). Importantly, the trial order must be truly random and simulations can be used to verify that the design does not bias the results. If there are restrictions in the study design that do not allow for trials to be randomly ordered, for example if the trials are grouped according to subject-specific criteria, the Type I error can be controlled if multiple runs of data are collected and pattern similarities are only computed between pairs of patterns from different runs.

Overview: pattern classification

Intuitively one would think that a within-run CV would be susceptible to a peeking bias, since test and training data sets may not be independent (Kriegeskorte et al., 2009) and this is reflected in the inflated classification accuracies that were found in most of our simulations. This is especially the case for blocked and alternating trials when the ISI was only 3 second long, although the bias diminished as ISI was

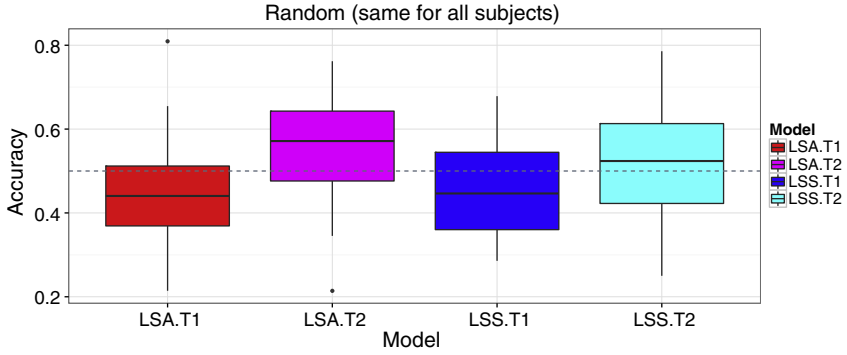

Fig. 5. Classification accuracy distributions across 30 simulated data sets, all using the same trial randomization and ISI. Accuracy distributions are separated according to trial type and pattern estimator. Although the pooled accuracies, across trial types, would come out at chance, variability differences in the data, due to the trial order and ISI, cause one trial type to be over classified.

increased and when trials were randomly ordered for each subject (Fig. 4). On the other hand, between-run CV is stable regardless of trial order and is the recommended approach. Note if the trials are not randomly ordered for each subject (Fig. 5), within-trial-type classification accuracies are not centered about 0.5, likely due to variability differences in the estimated patterns driven by ISI differences. Combined with Mumford et al. (2012), the present results suggest that, in terms of pattern classification, LSS is overall more advantageous as a shorter ISI can be used without any detriment to the Type I or II error rates, as long as a between-run CV is used.

Although we have found that appropriate randomization can be used to maintain an acceptable Type I error rate in within-run similarity and CV, this is only a limited set of simulations and it is likely that for different trial durations, or ISI distributions, the random trial order could have issues on within-run comparison. In particular, if random sequences are selected via commonly employed optimization algorithms designed for univariate studies (Dale, 1999; Liu, 2004), then true randomization is unlikely to hold as typically only a few optimal sequences are chosen and repeated across subjects due to the time intensive nature of optimization. Further, such optimized sequences often contain "mini blocks" where there are several trials in a row of the same type, which may result in a peeking effect and other problems that we observed related to blocking. Beyond biases due to a priori optimization, studies selecting trials for classification or similarity analysis based on behavior instead of experimentally pre-defined conditions (e.g., subsequent memory; Wagner et al. (1998)) may also never reach true randomization as many behavioral measures are intrinsically autocorrelated (Gilden, 2001). Thus in many common analysis contexts, between-run comparisons may be the only option for truly non-biased comparisons. If one were to use within-run comparisons based on our findings, it is recommended that simulations similar to the ones run here be used to validate the approach on the specific nuances of their design.

Although we found no benefit of one pattern estimator over the others in terms of Type I error rate for either the similarity or classification analyses, this does not mean that all pattern estimators are equally useful for MVPA. Another critical study design aspect that we did not directly address here is how the statistical power of pattern similarity tests varies as a function of pattern estimator. Given that LSS outperforms both LSA and Add6 in classification analyses (Mumford et al., 2012), it is likely to have added power in pattern similarity analyses as well. Further work will be necessary to evaluate power differences between the models within the pattern similarity framework.

No observed benefit with increased ISI

One of the more surprising results is even with long ISIs there is an impact of trial order on pattern similarity estimates. Under the assumption that data are not temporally correlated (top left panel, Fig. 3), there was a bias in the comparisons of *wt*1–*bt1t*2 and *wt*2–*bt1t*2 for blocked and alternating trials. This effect persists after temporal covariance is added with a slight improvement in Type I error for the alternating trial ordering. Effects driven by temporal autocorrelation occur because time points that are closer tend to be more highly correlated than temporally distant time points, hence pattern similarities for trials that are close to each other will be higher than trials that are further apart. In the case of the blocked trials, most between-trial-type similarities are very far apart and so their similarities will be smaller, whereas the within-trial-type similarities are at smaller lags and temporal correlation will cause the pattern similarities to be larger. This is why the *wt1–bt1t2* and *wt2–bt1t2* distributions tend to be significantly larger than 0 (bottom row, Fig. 3). Although it appears that the within- versus between-trial-type comparisons are smaller when temporal autocorrelation is added, this is not the case, but the range on the y-axis is much larger in the bottom row of plots.

Generally, at small ISIs the different pattern estimators suffer from collinearity, driven by positively correlated regressors in the models. When the ISI is increased this alleviates collinearity, yet there will always be a slight negative correlation between regressors since when one regressor increases, the other is flat. This is a very weak effect, but is enough to drive biases in the pattern similarity estimates. This can be seen when the ISI is increased from 3 to 7 s when temporal covariance was ignored (top row, Fig. 3). For example, with a small ISI, adjacent trials are positively correlated, causing similarities between trials to be biased negatively. Hence, in the blocked trials case, within-trial-type similarities are smaller than they should be (causing *wt*1–*bt1t*2 and *wt*2–*bt1t*2 to be negative) and in the alternating trial setting, between-trial-type similarities are smaller than they should be, having the opposite effect. When the ISI is increased there is a weak negative correlation, which is slightly stronger at early lags and this reverses the impact. For example, in the LSA design matrix with an ISI of 15 s the average correlation between adjacent regressors is −0.051 while at lags of 2 or higher the correlation is around −0.014. Although the bias is very small, it still yields false positive comparisons as shown in Table 1.

Benefits of between-run similarity analyses

The inflation of Type I error rates that arose in the within-run similarity analyses was driven by correlations, either between covariates in the model used to estimate the patterns, or the temporal covariance. The results of the between-run-based similarities are intuitive, given that independent models and time series are being compared. One assumption of this analysis is that the time series from two different runs are completely independent from each other. In fact, the simulations used temporal covariance estimates from different subjects in place of temporal covariance estimates from two runs of the same subject. Since temporal covariance for small differences in time seems to have the largest impact, as evidenced by the false positive differences in the blocked design, it seems that two runs, which would typically have a couple of minutes between them, would not be problematic. Beyond removing potential biases induced by temporal correlations, there may be additional benefits for between-run comparisons that derive from increasing the generalization performance of MVPA methods (Coutanche et al., 2012).

Why and when within-run CV fails

Within-run cross validations are especially problematic for the blocked trial order and somewhat for alternating trial orders while, surprisingly, randomly ordered trials seem to perform fine within the within-run CV setting, when trials are randomly ordered for each subject. The reason the results vary according to study design is due to different levels of peeking bias. In order for a peeking bias to occur the correlation, caused by collinearity in the model, must be informative about the trial types. For example, when trials are blocked, highly collinear trial pairs tend to be of the same class, causing a relationship between trial pairs and the strength of the collinearity. On the other hand, in the alternating case, the trials of the same class are always separated by at least 1 other trial, so the relationship is weaker. This still induces a peeking bias, but the effect is not as strong, which is what was found in Fig. 4. With randomly ordered trials, this is not as likely and the within-run CV is not shown to be impacted by peeking bias. Importantly, again, the trials must be randomly ordered for each subject.

Add6 model

Intuitively, with the Add6 approach, one might think that there will not be a model-based effect, since a model is not necessary to extract the patterns. However, these models can be expressed in terms of the GLM (Mumford et al., 2012) and, accordingly, the results of the Add6 model are very similar to LSA at a long ISI of 15 s (Table 1). That is because the problem with LSA at high ISIs is not due to overlapping regressors,

a positive correlation between adjacent regressors, but a small, weak, negative correlation between regressors, which was described above. Although the Type I errors were reduced with Add6, they were still not valid due to these weak negative correlations.

Impact on future study designs

Given the benefits of using between-run similarity estimates and between-run cross validation for MVPA analyses, it seems that using multiple, shorter runs would be more advantageous than using longer runs (Coutanche et al., 2012). An example would be if one was interested in studying pattern similarities for stimuli before and after learning. Since within-run patterns cannot be used, one would want to split up runs so that each run captures a different level of learning. This would likely require both shorter runs and tasks where learning occurs slow enough that ceiling is not immediately reached. Although it might be tempting to take a long run of data and divide it into shorter "runs" after data collection, this will not provide results that mimic the between-run similarity simulations done here, since there will be slight correlations between the end of a run and the beginning of the proceeding run. Whether or not this can be alleviated by trimming trials off the boundaries is a question for a future data analysis.

Conclusion

This work highlights important design considerations for pattern similarity or classification analysis with fMRI. It establishes that the LSS pattern estimator can be used without negatively impacting Type I or II errors with designs using ISI's as short as 3 s on average. For pattern classification, multiple runs of data are required to guarantee that there is not a peeking bias impacting classification accuracy estimates. In the case of pattern similarity analyses, studies must be designed more carefully if patterns within the same run are to be compared, in which case only designs with truly randomly ordered trials can be used without inflating Type I error. This trial randomization must be unique for each subject and, to verify complete randomization, simulations, such as the ones run here, can be used to verify that there is no impact of the design on Type I error. When there is structure to the trial order, even with a large ISI of 15 s, the patterns suffer from false positive correlations that bias the similarity analysis. To avoid this issue, correlations should only be computed between patterns from independent runs.

References

Carlson, T.A., Schrater, P., He, S., 2003. Patterns of activity in the categorical representations of objects. J. Cogn. Neurosci. 15, 704–717.

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.

Dale, A.M., 1999. Optimal experimental design for event-related fMRI. Hum. Brain Mapp. 8 (2–3), 109–114.

Davis, T., Poldrack, R.A., 2013. Measuring neural representations with fMRI: practices and pitfalls. Ann. N. Y. Acad. Sci. 1296 (1), 108–134.

Gilden, D.L., 2001. Cognitive emissions of 1/f noise. Psychol. Rev. 108 (1), 33.

Greve, D.N., Fischl, B., 2009. Accurate and robust brain image alignment using boundary-based registration. NeuroImage 48 (1), 63–72.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430.

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

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5 (2), 143–156.

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17 (2), 825–841.

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

Kriegeskorte, N., Kievit, R.A., 2013. Representational geometry: integrating cognition, computation, and the brain. Trends Cogn. Sci. 17 (8), 401–412.

Kriegeskorte, N., Mur, M., Bandettini, P., 2008a. Representational similarity analysis – connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2, 4.

Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., Bandettini, P.A., 2008b. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60 (6), 1126–1141.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S., Baker, C.I., 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci. 12 (5), 535–540.

Liu, T.T., 2004. Efficiency, power, and entropy in event-related fMRI with multiple trial types. NeuroImage 21 (1), 401–413.

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.

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

Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. NeuroImage 45, 199–209.

Rissman, J., Gazzaley, A., D'Esposito, M., 2004. Measuring functional connectivity during distinct stages of a cognitive task. NeuroImage 23 (2), 752–763.

Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17 (3), 143–155.

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.

Wagner, A.D., Schacter, D.L., Rotte, M., Koutstaal, W., Maril, A., Dale, A.M., Rosen, B.R., Buckner, R.L., 1998. Building memories: remembering and forgetting of verbal experiences as predicted by brain activity. Science 281 (5380), 1188–1191.