Searchlight-based trial-wise fMRI decoding in the presence of trial-by-trial correlations

Joram Soch[a,b,c,d,•]

[a] Berlin Center for Advanced Neuroimaging, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health
[b] Berlin Center for Computational Neuroscience, Berlin, Germany
[c] German Center for Neurodegenerative Diseases, Göttingen, Germany
[d] Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

- Corresponding author: joram.soch@bccn-berlin.de.
  BCCN Berlin, Philippstraße 13, Haus 6, 10115 Berlin, Germany.

Abstract

In multivariate pattern analysis (MVPA) for functional magnetic resonance imaging (fMRI) signals, trial-wise response amplitudes are sometimes estimated using a general linear model (GLM) with one onset regressor for each trial. When using rapid event-related designs with trials closely spaced in time, those estimates can be highly correlated due to the temporally smoothed shape of the hemodynamic response function. In previous work (Soch, J., Allefeld, C., & Haynes, J.-D. (2020). Inverse transformed encoding models – a solution to the problem of correlated trial-by-trial parameter estimates in fMRI decoding. *NeuroImage*, 209, 116449, 1-19. https://doi.org/10.1016/j.neuroimage.2019.116449), we have proposed inverse transformed encoding modelling (ITEM), a principled approach for trial-wise decoding from fMRI signals in the presence of trial-by-trial correlations. Here, we (i) perform simulation studies addressing its performance for multivariate signals and (ii) present searchlight-based ITEM analysis – which allows to predict a variable of interest from the vicinity of each voxel in the brain. We empirically validate the approach by confirming *a priori* plausible hypotheses about the well-understood visual system.

Keywords

fMRI decoding, multivariate pattern analysis, searchlight-based decoding, trial-wise parameter estimates, general linear model, multivariate GLM, model inversion

Contents

1 Introduction **1**

2 Methods **3**
2.1 Trial-wise decoding from fMRI 3
2.2 Statistical theory behind ITEM analysis 4
2.3 Practical steps of an ITEM analysis 6
2.4 Searchlight-based implementation 7

3 Simulation **9**
3.1 Methods 9
3.2 Results 10

4 Application **12**
4.1 Experiment 12
4.2 Analysis 12
4.3 Results 13

5 Discussion **15**
5.1 Assessment of simulation validation 15
5.2 Assessment of empirical validation 15
5.3 Assumptions and Limitations 16

6 Statements **18**
6.1 Ethics Statement 18
6.2 Data and Code 18
6.3 Author Contributions 18
6.4 Funding Statement 18
6.5 Competing Interests 18
6.6 Acknowledgements 18

7 Appendix **19**
A Detailed information for simulation study 19
B A note on parametric modulators in SPM 20

8 References **23**

1 Introduction

In functional magnetic resonance imaging (fMRI), data are frequently analyzed with univariate encoding models (Brodersen et al., 2011b) such as general linear models (GLMs) as well as multivariate decoding algorithms (Brodersen et al., 2011a) such as support vector machines (SVMs). Univariate encoding models construct a relationship between experimental variables and the measured signal in one voxel which allows to statistically test activation differences between experimental conditions (Smith, 2004; Monti, 2011). Multivariate decoding algorithms extract experimental variables from the measured signals in many voxels which allows to reliably decode experimental conditions from brain activation (Haxby et al., 2001; Haxby, 2012; Cox and Savoy, 2003; Norman et al., 2006; Haynes and Rees, 2006; Haynes, 2015). This is commonly called "multivariate pattern analysis" (MVPA) for neuroimaging data.

MVPA for fMRI can either be performed by decoding from measured signals in the entire brain ("whole-brain decoding") or by decoding from measured signals in a spatially well-circumscribed region of interest (ROI; "ROI-based decoding"). Another, third option is given by building a sphere of voxels around each voxel and then decoding from measured signals in each of these "searchlights" separately ("searchlight-based decoding"). Searchlight-based decoding was introduced in the early days of MVPA for fMRI to harness the statistical power of cross-validated prediction afforded by machine learning algorithms, but to also identify information at potentially unexpected locations in the brain (Kriegeskorte et al., 2006; Haynes et al., 2007).

Cross-validated prediction can either be performed over parameteter estimates calculated from fMRI recording sessions (so-called "run-wise betas") or it can be used to decode the identity of individual trials (decoding based on "trial-wise betas"). For trial-wise decoding, a common approach is to estimate trial-wise response amplitudes from the fMRI signal using a GLM with one onset regressor per trial (Rissman et al., 2004; Mennes et al., 2013), generated by convolution with a hemodynamic response function (HRF; Friston et al., 1998; Henson et al., 2001). When inter-trial-intervals are very short, those HRF regressors overlap in time due to the temporally extended shape of the canonical HRF, causing trial-wise estimates to be serially correlated and highly variable (Mumford et al., 2012; Turner et al., 2012) which can distort parameter estimates and invalidate statistical tests (Mumford et al., 2014).

One currently accepted approach for solving this problem is to estimate each trial's response via a GLM including a regressor for that trial and another regressor for all other trials (Mumford et al., 2012). This approach is called the "least squares, separate" method (LS-S) and was found to outperform the uncorrected "least squares, all" method (LS-A) as well as a range of other techniques by Mumford and colleagues (Mumford et al., 2012). The rationale behind LS-S is that the one-trial regressor is only weakly correlated to the all-other-trials regressor which effectively reduces the variance and auto-correlation of the trial-wise parameter estimates (Mumford et al., 2014). One disadvantage of LS-S is that each trial requires fitting a separate GLM, so that e.g. calculating activation patterns for 100 trials needs 100 GLMs.

In previous work, we have suggested "inverse transformed encoding models" (ITEM; Soch et al., 2020), a trial-wise modelling framework that builds on LS-A estimates, but accounts for the correlation between trials by incorporating their covariance matrix into a linear model operating at the trial level. This first contribution was somewhat limited by the fact that (i) ITEM as a technique for multivariate decoding was validated using an univariate simulation (adapted from Mumford et al., 2012); and (ii) we used ROI-based decoding for a visual stimulation dataset (acquired by Heinzle et al., 2011) which did not explore the full potential of ITEM for localization of information in the brain.

In this paper, we present and validate searchlight-based ITEM analysis (ITEM-SL), i.e. ITEM-style decoding from signals in spherical volumes of interest the center of which is moving through the brain. We perform a truly multivariate simulation in which we repeatedly generate data from synthetic searchlights and find that performance gains of ITEM-SL over LS-S are even higher than in our original simulation, pointing to the possibility that the advantage of the proposed over the state-of-the-art approach grows with increasing number of voxels that are decoded from. Additionally, we apply ITEM-SL to a visual fMRI data set (Soch et al., 2023) and are able to recover classically known principles of visual cortex organisation, e.g. contralateral processing of the visual hemifields and polar-coordinate representation of the visual field.

The structure of this paper is as follows. First, we will introduce ITEM-SL by recapitulating the theory behind ITEM-style analysis and describing the searchlight-based implementation (see Section 2 and Figure 2). Second, we will perform a simulation study on searchlight-based classification from synthetic fMRI data to demonstrate that ITEM is more powerful than combining the currently accepted approach with SVM-based classification (see Section 3 and Figure 3). Third, we will describe an empirical application in which ITEMs are used for searchlight-based reconstruction of massively parallel visual information in an extremely rapid event-related design and thereby recover well-known properties of early visual cortex (see Section 4 and Figure 4).

2 Methods

In this section, we briefly summarize the methodology on which searchlight-based ITEM analysis is based. We review the problem of trial-by-trial correlations in fMRI decoding (see Section 2.1) and recapitulate how inverse transformed encoding models solve this problem (see Section 2.2). Then, we describe how an ITEM analysis works in practice (see Section 2.3) and explain the searchlight-based implementation of this approach (see Section 2.4). For more theory behind the methodology, see Soch et al., 2020.

2.1 Trial-wise decoding from fMRI

In univariate fMRI analysis, the goal is usually to investigate whether experimental conditions have statistically significant (or, significantly different) effects on measured responses in single voxels. These data are often analyzed using the "standard" general linear model (standard GLM)

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathrm{N}(0, \sigma^2 V) \tag{1}$$

in which y is an $n \times 1$ vector of the measured BOLD signal in a single voxel (n = number of fMRI scans), X is an $n \times p$ design matrix containing predictor variables (p = number of predictor variables, or "regressors"; Smith, 2004), β is a $p \times 1$ vector of regression coefficients and ε is an $n \times 1$ vector of error terms which are controlled by the noise variance σ^2 and the $n \times n$ covariance matrix V.

In such analyses, X is typically known (because imposed by the experimental design) and V is also given (e.g. via estimated restricted maximum likelihood across voxels; see Friston et al., 2002b; Friston et al., 2002a), but β and ε are unknown and have to be estimated. Usually, X consists of "condition regressors", i.e. trial onsets and durations convolved with the hemodynamic response function (HRF), and other regressors given as scan-by-scan covariates. Estimation of and inference based on (1) is usually referred to as the "mass-univariate approach" (Monti, 2011).

In multivariate fMRI analysis, the goal is sometimes to perform trial-wise decoding, i.e. to provide predictions for individual trials rather than collapsing them into conditions. In this case, it is advantageous to estimate trial-wise response amplitudes using a "trial-wise" general linear model (trial-wise GLM)

$$y = X_t\gamma + \varepsilon_t, \; \varepsilon_t \sim \mathrm{N}(0, \sigma_t^2 V) \tag{2}$$

where X_t is an $n \times t$ design matrix with one HRF onset regressor for each trial (t = number of trials), instead of one such regressors for each condition, as in X (see Figure 1A); γ is a $t \times 1$ vector of trial-wise response amplitudes, sometimes also referred to as the "trial-wise betas" (cf. Rissman et al., 2004, p. 752); ε_t and σ_t^2 are the error terms and the noise variance of the trial-wise GLM, respectively.

Trial-wise response amplitudes can be estimated from (2) using e.g. weighted least squares which results in reponses commonly referred to as "LS-A estimates" (LS-A for "least squares, all"; Mumford et al., 2012)

$$\hat{\gamma} = (X_t^T V^{-1} X_t)^{-1} X_t^T V^{-1} y \; , \tag{3}$$

but this can become problematic: In rapid event-related designs, when inter-stimulus-intervals are short, the HRFs from adjacent trials can overlap in time, due to the comparably slow hemodynamic response with a peak at around 6 s and a post-stimulus undershoot until 20-30 s after stimulus onset (Friston et al., 1998). This induces serial correlations into the estimated trial-wise responses and makes those estimates more variable which reduces the statistical power of any analysis operating on them (Mumford et al., 2012; Turner et al., 2012).

For this reason, it has been proposed to estimate the response amplitude of each trial using a separate design matrix which leads to so-called "LS-S estimates" (LS-S for "least squares, separate"; Mumford et al., 2012)

$$\hat{\gamma}_i = \hat{\beta}_1^{(i)} \quad \text{where} \quad \hat{\beta}^{(i)} = (X_i^T V^{-1} X_i)^{-1} X_i^T V^{-1} y \quad \text{for} \quad i = 1, \ldots, t \quad , \tag{4}$$

where X_i is an $n \times 2$ design matrix with one HRF regressor for the i-th trial and another regressor for all other trials. This requires that a separate GLM is run for each trial-wise parameter estimate. The rationale for this approach is that, because the second regressor contains *all other* trials, correlation with the first regressor is reduced which makes estimated trial-wise responses more robust. LS-S has been validated in previous work (Mumford et al., 2012; Turner et al., 2012; Mumford et al., 2014; Weeda, 2018) and is currently the most widely used approach of extracting response estimates for trial-wise fMRI decoding.

2.2 Statistical theory behind ITEM analysis

Rather than artificially reducing the correlations between trial-wise parameter estimates which LS-S does, the ITEM approach attempts to naturally account for them by estimating and integrating their extent into the classification process. This starts by relating the design matrix of the standard GLM X to the design matrix of the trial-wise GLM X_t via a transformation matrix T

$$X = X_t T \tag{5}$$

where T is a $t \times p$ matrix mapping from trials to conditions. In the simplest case of a categorical design, T will simply be an indicator matrix where $t_{ij} = 1$ indicates that trial i belongs to condition j (see Figure 1A). However, T can also take a more complex form to emulate parametric modulators and nuisance regressors, as known from standard design matrices (see Soch et al., 2020, Fig. 1B).

Upon making the assumption given by (5), it can be shown (see Soch et al., 2020, App. A) that the trial-wise parameter estimates from (3) follow a new linear model operating on the trial-by-trial level in which the design matrix is given by the transformation matrix T and the covariance matrix is a function of X_t

$$\hat{\gamma} = T\beta + \eta, \; \eta \sim \mathrm{N}(0, \sigma^2 U) \tag{6}$$

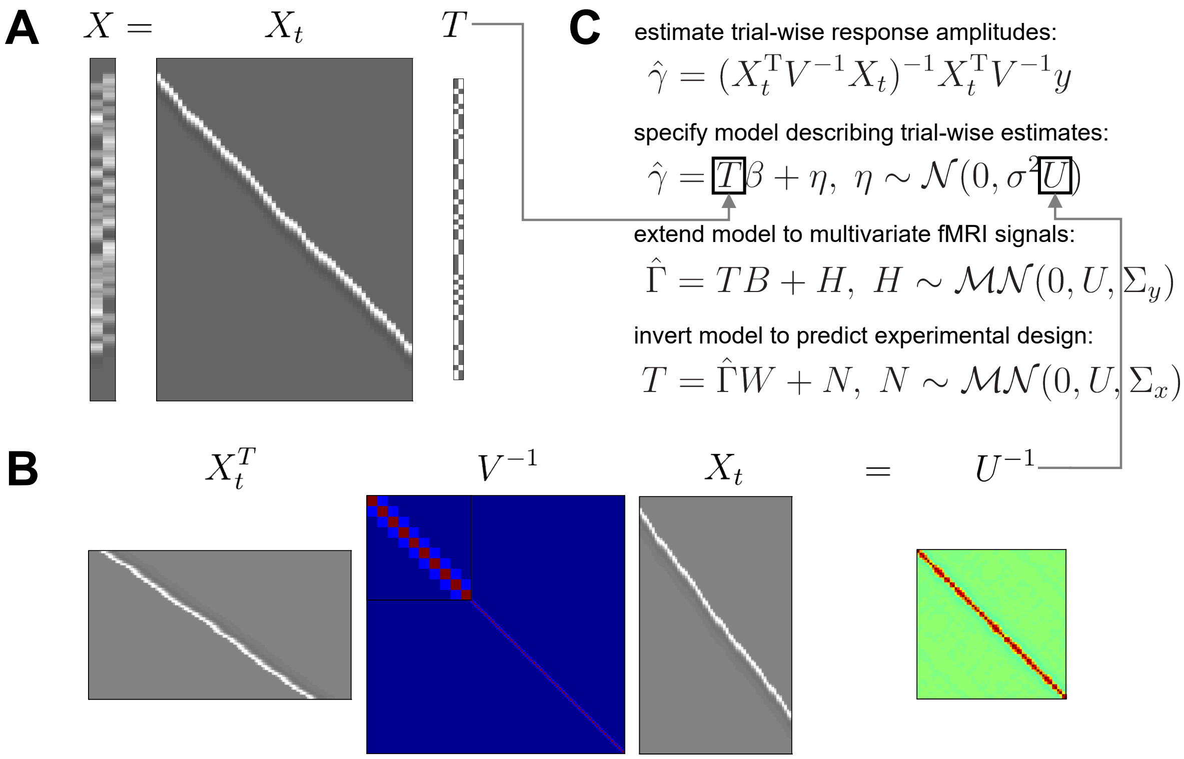
A
$X =$
X_t
T
B
X_t^T
V^{-1}
X_t
$=$
U^{-1}
C
estimate trial-wise response amplitudes:
$\hat{\gamma} = (X_t^{\mathrm{T}} V^{-1} X_t)^{-1} X_t^{\mathrm{T}} V^{-1} y$
specify model describing trial-wise estimates:
$\hat{\gamma} = \boxed{T}\beta + \eta, \; \eta \sim \mathcal{N}(0, \sigma^2 \boxed{U})$
extend model to multivariate fMRI signals:
$\hat{\Gamma} = TB + H, \; H \sim \mathcal{MN}(0, U, \Sigma_y)$
invert model to predict experimental design:
$T = \hat{\Gamma} W + N, \; N \sim \mathcal{MN}(0, U, \Sigma_x)$

Figure 1: *Mathematics of ITEM analysis.* **(A)** The trial-wise design matrix X_t (scans × trials) can be related to the standard design matrix X (scans × conditions) using a trial-level specification matrix T (trials × condition). In this simple case, T is just an indicator matrix specifying which trial belongs to which condition. **(B)** Under this relationship, the (inverse of the) trial-by-trial covariance matrix U (trials × trials) is equal to the trial-wise design matrix X_t, multiplied with itself and weighted by the scan-by-scan covariance matrix V (scans × scans). In the present case, overlapping HRFs induce serial correlation betweens temporally nearby trials. **(C)** Given that trial-wise response amplitudes γ are estimated from the voxel-wise fMRI signal y using the trial-wise design matrix X_t (first equation), it can be shown that they follow a linear model with the transformation matrix T as its design matrix and the uncorrelation matrix U as its temporal covariance (second equation). Combining the trial-wise parameter estimates from multiple voxels into a matrix ($\hat{\Gamma} = [\hat{\gamma}_1, \ldots, \hat{\gamma}_v]$) leads to a multivariate version of this model (third equation) and assuming a correspondence between forward and backward model ($BW = I_p$) leads to an inverted version of this model (fourth equation).

where $\hat{\gamma}$ is a $t \times 1$ vector given by (3), T is the $t \times p$ matrix defined by (5) and U is a $t \times t$ matrix which specifies the trial-by-trial covariance and can be directly calculated from the trial-wise design-matrix[1] (see Figure 1B):

$$U = (X_t^T V^{-1} X_t)^{-1} \; . \tag{7}$$

U is referred to as the “uncorrelation matrix”, because it allows to decorrelate trials and equation (6) is referred to as the “transformed encoding model”, because it operates on

[1] See: https://statproofbook.github.io/P/tglm-dist.

a transformed version of the measured data y, namely $\hat{\gamma}$. Realizing that the response estimates from several voxels can be collected together, it extends into the "multivariate transformed encoding model" (see Figure 1C)

$$\hat{\Gamma} = TB + H, \; H \sim \mathrm{MN}(0, U, \Sigma_y) \tag{8}$$

where $\hat{\Gamma} = [\hat{\gamma}_1, \ldots, \hat{\gamma}_v]$ is a $t \times v$ matrix of trial-wise parameter estimates (v = number of voxels) and Σ_y is a $v \times v$ matrix describing the spatial covariance, i.e. correlations in the activity of nearby voxels. Finally, this model can be turned into an "inverse transformed encoding model" (see Soch et al., 2020, App. C)

$$T = \hat{\Gamma} W + N, \; N \sim \mathrm{MN}(0, U, \Sigma_x) \tag{9}$$

where W is a weight matrix mapping from trial-wise response amplitudes $\hat{\Gamma}$ to experimental design variables T, defined as the inverse of the activation pattern B via $BW = I_p$; which implies that $\Sigma_x = W^T \Sigma_y W$ is a $p \times p$ matrix describing the ensuing condition-by-condition covariance of the ITEM[2].

In an ITEM analysis, the goal to estimate (9) in a cross-validated fashion in order to come up with a prediction for T. More precisely, the weight matrix W is estimated from the trial-wise responses $\hat{\Gamma}$ and the trial-by-trial correlations U of all but one fMRI recoding session (e.g. sessions 2-4) and then used to predict the trial-level specification matrix T in the left-out session (e.g. session 1).

2.3 Practical steps of an ITEM analysis

In practice, an ITEM analysis will proceed as follows:

1. *custom fMRI preprocessing*: Before any statistical analysis, fMRI data are preprocessed. Preprocessing is completely independent from ITEM-style analysis and can therefore be performed according to the preferences of the individual researcher.
2. *standard GLM analysis*: Then, a standard GLM analysis is performed. This is performed in the exact same way as for univariate fMRI analysis, i.e. using a condition-based, not trial-wise GLM, and can be done via standard packages, e.g. Statistical Parametric Mapping, Version 12 (SPM12; Ashburner et al., 2021).
3. *trial-wise GLM estimation*: Then, based on this design information, a trial-wise design matrix is specified and trial-wise response amplitudes are estimated via (3). The ITEM toolbox for SPM (see Section 6.2) allows to distinguish conditions that are broken up into trials (e.g. target stimuli) vs. not broken up into trials (e.g. cue stimuli).
4. *trial-wise fMRI decoding*: Next, the actual predictive analysis is perfmored. In this step, trial-wise parameter estimates from a number of voxels (e.g. from a region of interest (ROI); or within searchlights (SL), see Section 2.4) are loaded and the desired decoding operation is performed using ITEM-style inversion of the model given by (9). The ITEM toolbox for SPM allows to select between
   a. classification vs. regression; as well as
   b. ROI-based decoding vs. searchlight-based decoding.

[2] See: https://statproofbook.github.io/P/iglm-dist.

5. *group-level analysis*: If applicable, decoding results can be generalized to the population by integrating single performance values from ROI-based decoding or voxel-wise performance maps from SL-based decoding across subjects.

In the later empirical validation (see Section 4), we describe for each of these five steps, how our exemplary ITEM analysis was conducted.

2.4 Searchlight-based implementation

While trial-wise parameter estimation works on a voxel-wise level[3], trial-wise fMRI decoding was previously only implemented as ROI-based analysis[4]. With this work, we provide searchlight-based implementations of ITEM-style analyses[5] and also perform a comprehensive application of searchlight decoding (see Section 4).

Searchlight-based ITEM analysis consists of the following steps: First, the desired searchlight radius is used to generate searchlights by taking each in-mask voxel as the center voxel, placing a sphere with the given radius around it and including all in-mask voxels inside the sphere (see Figure 2, top-right). Second, the trial- and voxel-wise responses $Y = \hat{\Gamma}$ are extracted from each searchlight. For example, if the number of trials is 120 and the number of voxels per searchlight is 50, then Y will be a 120×50 signal matrix. Third, the transformation matrix T and uncorrelation matrix U are gathered to specify the transformed encoding model for each fMRI recoding session (see Figure 2, center). Fourth, this model is inverted and variables of interest T are predicted via cross-validated estimation of the inverted model (see Figure 2, bottom-left).

Finally, if this has been repeated for all searchlights, a measure of decoding performance is calculated at each voxel by comparing the actual against the predicted values of the experimental design variables, across sessions:

- For *classification*, the column in the estimated matrix $\hat{T}$ $(t \times p)$ with the highest value is selected as the predicted condition in each trial, i.e. for each row. Then, decoding performance is calculated as decoding accuracy, i.e. the number of correct classifications, divided by the total number of trials.
- For *regression*, a column in the estimated matrix $\hat{T}$ is taken as the set of predicted target values for this regressor. Then, decoding performance is calculated as the correlation between the predicted and actual values for this regressor.

Decoding performances are stored in a single map for (a) each contrast to classify or (b) each regressor to predict and can later be used for voxel-wise group-level analysis.

[3] See function `ITEM_est_1st_level` in the repository https://github.com/JoramSoch/ITEM.

[4] See functions `ITEM_dec_class` and `ITEM_dec_recon` of the ITEM toolbox.

[5] See functions `ITEM_dec_class_SL` and `ITEM_dec_recon_SL` of the ITEM toolbox.

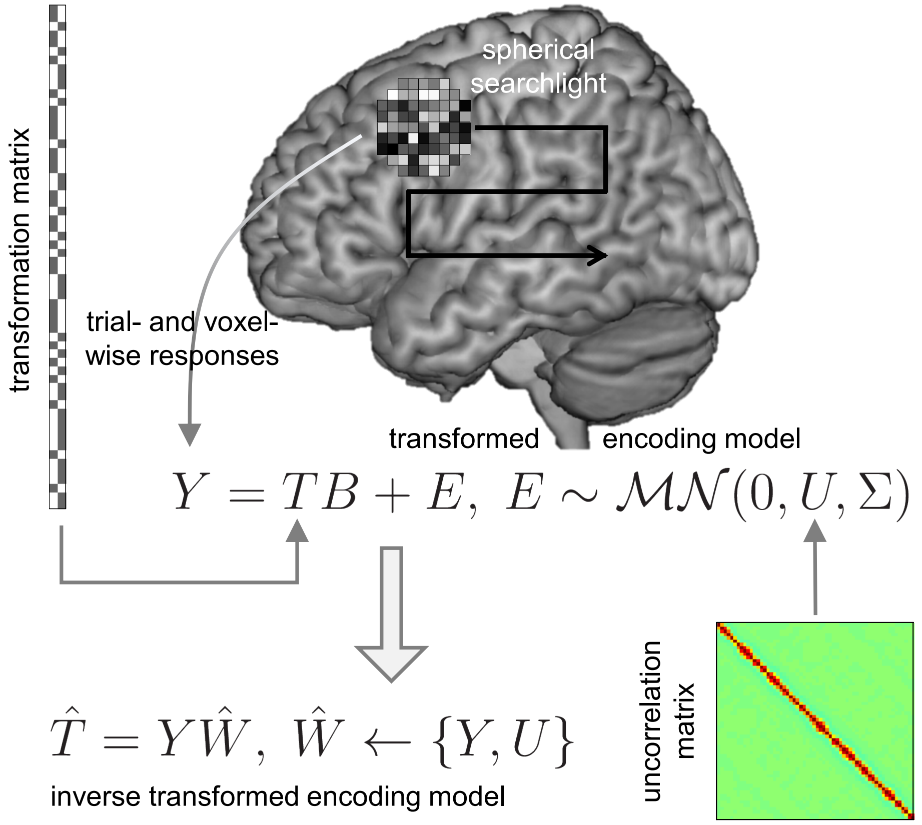
transformation matrix
spherical searchlight
trial- and voxel-wise responses
transformed
encoding model
$Y = TB + E,\ E \sim \mathcal{MN}(0, U, \Sigma)$
uncorrelation matrix
$\hat{T} = Y\hat{W},\ \hat{W} \leftarrow \{Y, U\}$
inverse transformed encoding model

Figure 2: *Searchlight-based ITEM analysis.* The trial-wise parameter estimates Y (trials × voxels) are extracted from all voxels belonging to a spherical searchlight which is successively centered on each voxel in the brain (black arrow). The estimated responses Y follow a multivariate linear model which uses the transformation matrix T (see Figure 1A) as design matrix and the uncorrelation matrix U (see Figure 1B) as temporal covariance (gray arrows; for simplicity, $\hat{\Gamma}$ from Figure 1C is here replaced by Y and Σ_y from Figure 1C is here replaced by Σ). This is called a "transformed encoding model" (center equation). Such a model can be inverted by predicting experimental design variables T from the estimated responses Y and their covariance U via an estimated weight matrix $\hat{W}$. This is called "inverse transformed encoding modelling" (bottom-left equation).

3 Simulation

To validate searchlight-based ITEM using synthetic data, we adapt a simulation from Soch et al., 2020 which was itself adapted from Mumford et al., 2012. The main change was replacing the previously univariate generative model by multivariate signals generated in the present simulation. All simulation code is available from GitHub (see Section 6.2).

3.1 Methods

In our simulation, we compare the three approaches of trial-wise decoding from fMRI signals introduced earlier: the naïve approach operating on uncorrected trial-wise parameter estimates (Mumford: "least squares, all", LS-A), the state-of-the-art approach based on parameter estimation using separate models (Mumford: "least squares, separate", LS-S) and the new approach proposed here, i.e. searchlight-based inverse transformed encoding modelling (ITEM-SL). LS-A entails decoding without accounting for correlation and taking trial-wise parameter estimates $\hat{\gamma}$ from equation (3) "as is". LS-S is based on trial-wise parameter estimates using a separate design matrix X_i for each trial $i = 1, \ldots, t$, including one regressor for this trial and one regressor for all other trials. ITEM uses the same estimates as LS-A, but accounts for their correlation by incorporating the trial-by-trial covariance matrix U as in equation (9) (see Figure 2).

In the simulation, data were generated as follows: First, trials were randomly sampled from two experimental conditions, A and B. Second, voxel-wise average responses $\mu_{A,j}$ and $\mu_{B,j}$ (j indexes voxel) were sampled from the standard normal distribution $\mathcal{N}(0,1)$. Third, trial-wise response amplitudes $\gamma_{i,j}$ (i indexes trials) were sampled from normal distributions $\mathcal{N}(\mu_{A,j}, \sigma_\gamma^2)$ and $\mathcal{N}(\mu_{B,j}, \sigma_\gamma^2)$ where $\sigma_\gamma = 0.5$.

Fourth, inter-stimulus-intervals t_i were sampled from uniform distributions $\mathcal{U}(0,4)$ or $\mathcal{U}(2,6)$ or $\mathcal{U}(4,8)$. Fifth, the design matrix X_t was generated based on the t_i's and convolution with the canonical HRF using stimulus duration $t_{\text{dur}} = 2$ s and repetition time $\text{TR} = 2$ s. An exemplary design matrix for the case $t_i \sim \mathcal{U}(0,4)$ is given in the middle of Fig. 1A. Finally, a multivariate signal was generated by multiplying the trial-wise design matrix X_t with trial-wise response amplitudes Γ and adding zero-mean Gaussian noise E with variance σ^2 where $\sigma^2 \in \{0.8, 1.6, 3.2\}$.

This was repeated for $N = 1{,}000$ simulations with $S = 2$ sessions and $t = 100$ trials per session (50 per condition). Each simulation can be seen as an individual searchlight and the number of voxels per simulation/searchlight was $v = 33$ which corresponds to a spherical searchlight with a radius of 2 voxels[6]. A proportion of voxels with information r was specified as $r = 20\%$, such that for $1 - r = 80\%$ of the voxels, $\mu_{B,j}$ was set to $\mu_{A,j}$, implying no difference between A and B in those voxels. A more detailed description of the simulation is given in Appendix A.

After data generation, decoding was performed as follows: For LS-A and ITEM, trial-wise parameter estimates $\hat{\Gamma}$ were obtained by least-squares estimation using design matrix X_t. For ITEM, $\hat{\Gamma}$ was subjected to an additional restricted maximum likelihood (ReML) analysis (see Soch et al., 2020, App. B), in order to separate the natural trial-to-trial variability (coming from σ_A and σ_B) from the induced trial-by-trial correlations (coming from X_t). For LS-S, $\hat{\Gamma}$ was obtained as described above using trial-specific design matrices

[6] See the third entry at https://oeis.org/A000605.

X_i where $i = 1, \ldots, t$. Afterwards, parameter estimates were subjected to support vector classification (SVC) with training on one session and testing on the other session to assess cross-validated decoding accuracy.

3.2 Results

Given that there are differences between them, the two experimental conditions can be decoded from the generated data using a meaching-learning classification algorithm. For this purpose, we here chose support vector machines for classification (SVC). For LS-A and LS-S, condition labels for A and B are coded as 1 and 2 and the corresponding support vector machine is calibrated based on training data. Then, condition labels are predicted from trial-wise response estimates in the left-out session.

For ITEM, as trial-by-trial correlations cannot be easily accounted for by SVC, the linear decoding procedure outlined above (see Section 2.2) was employed for cross-validated classification of trial types. For all approaches, decoding accuracy (DA), i.e. the percentage of trials correctly assigned across both sessions, was used as the measure of decoding performance. Each procedure leads to one DA value per simulation/searchlight, the distributions of which are visualized as box plots.

We found that, when setting the proportion of activated voxels to $r = 0\%$, such that no difference between the conditions exists, all approaches considered have an average decoding accuracy of around 50% (results not shown), for all levels of trial collinearity (t_i) and signal-to-noise ratio (σ^2). Thus, there is no evidence for systematic above-chance classification in the absence of a real effect.

Furthermore, when setting r to its value specified above, such that there is a real effect, ITEM outperforms LS-S in each simulation scenario (see Figure 3) by between 0.0% ($\sigma^2 = 0.8$, $t_i \sim \mathcal{U}(4, 8)$) and 14.0% ($\sigma^2 = 0.8$, $t_i \sim \mathcal{U}(0, 4)$) in terms of median decoding accuracy. This is particularly the case when inter-stimulus-intervals are short (all σ^2 for $t_i \sim \mathcal{U}(0, 4)$). Note that even LS-A outperforms LS-S for low-variance situations (all t_i for $\sigma^2 = 0.8$). In conclusion, the ITEM approach outperforms the previously best known approach in terms of sensitivity, for the simulation scenarios investigated here.

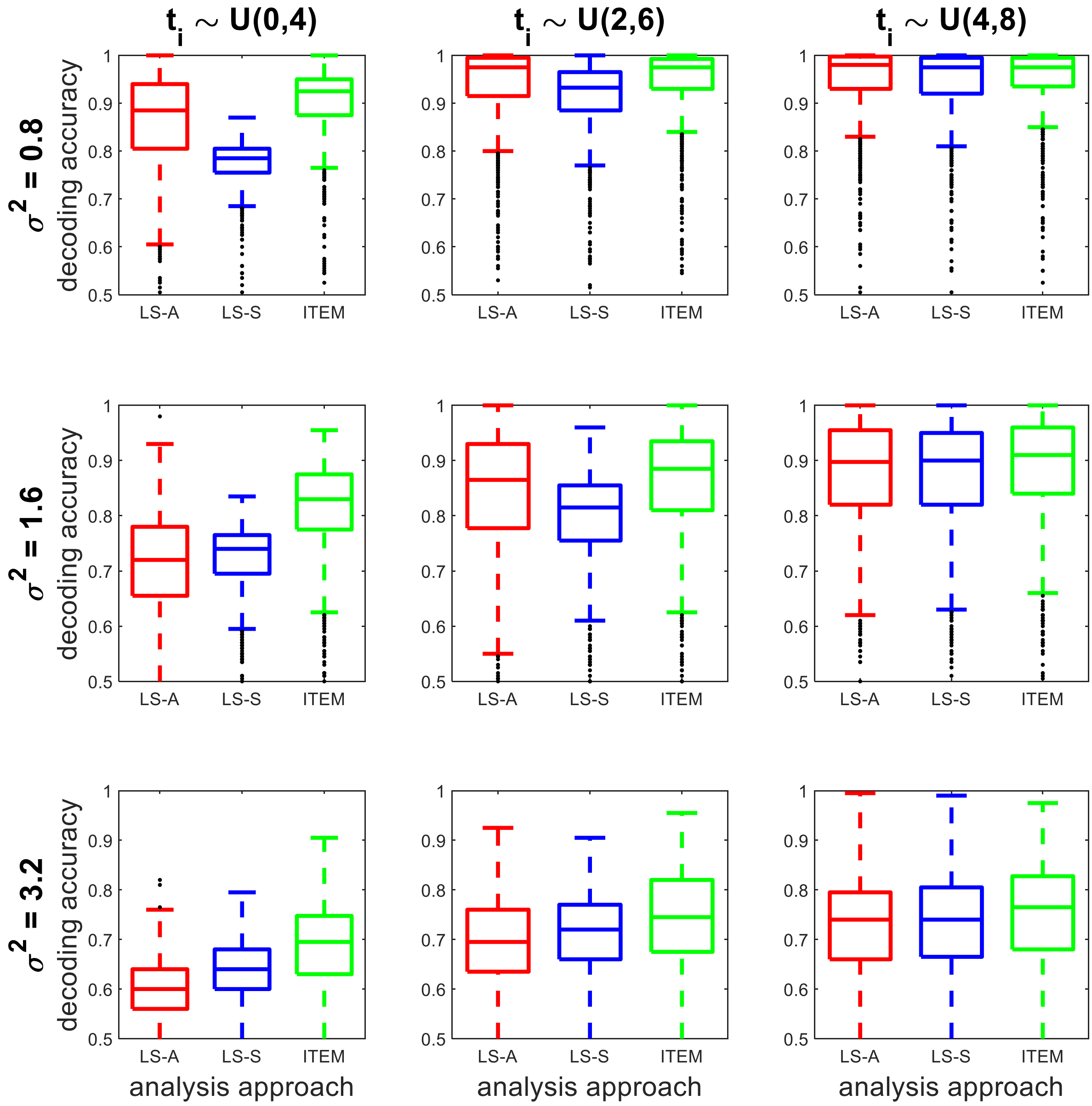

Figure 3: *Simulation validation of multi-voxel ITEM analysis.* For each combination of inter-stimulus-intervals (t_i) and noise variance (σ^2), decoding accuracies for classifying two experimental conditions are given for the naïve approach (LS-A, red), the standard approach (LS-S, blue) and the proposed approach (ITEM, green). For long t_i and low σ^2, average decoding accuracies of all algorithms are close to 1. When the noise variance is high (bottom row) or inter-stimulus-intervals are short (left column), the ITEM approach outperforms the state-of-the-art approach. In each boxplot, the central mark is the median; the box edges are the 25th and 75th percentiles; whiskers correspond to the most extreme data points within 1.5 × the interquartile range from the box edges; and black dots represent outliers. For details of the simulation study, see Appendix A.

4 Application

To validate searchlight-based ITEM using empirical data, we analyze fMRI data that were acquired by Heinzle et al., 2011 and are more closely described in Soch et al., 2023. This data set was originally acquired to investigate relationships between sensory-visual and cortico-cortical receptive fields and is here used to recover the spatial organization of early visual cortex. The entire data set is available from OpenNeuro (see Section 6.2).

4.1 Experiment

Because the descriptor of this data set is available open access (Soch et al., 2023)[7], the experimental design is reported rather shortly in this section.

Four right-handed, healthy subjects (24-28 years, 3 male, 1 female) participated in a visual stimulation experiment in which they viewed a dartboard-shaped stimulus that consisted of 48 "sectors" organized into 4 "rings" and 12 "segments" (see Figure 4A). Each sector was a flickering checkerboard randomly changing its local visual contrast every 3 s across 8 recording sessions with 100 trials per session.

Intensity levels were logarithmically spaced between 0.1 and 1 and used for analysis as linearly spaced between 0 and 1 in steps of 1/3. Importantly, there was no inter-stimulus-interval, implying the maximum possible overlap between HRFs at this stimulus duration and constituting a perfect application case of the presented approach of trial-wise decoding in the presence of trial-by-trial correlations.

To maintain fixation at the center of the visual display, subjects were engaged in a cognitive control task. Landolt's C was presented in the middle of the screen and subjects had to indicate whether it opened to left or to the right side.

Functional magnetic resonance imaging (MRI) data were collected on a 3-T Siemens Trio with a 12-channel head coil. In each session of the visual stimulation experiment, 220 T2*-weighted, gradient-echo EPIs were acquired at a repetition time TR = 1500 ms, echo time TE = 30 ms, flip-$\alpha = 90°$ in 25 slices (slice thickness: 2 mm (+1 mm gap); matrix size: 64 × 64) resulting in a voxel size of 3 × 3 × 3 mm.

4.2 Analysis

The five steps of ITEM analysis (see Section 2.3) for these data were as follows:

1. *custom fMRI preprocessing*: Data were converted to the BIDS format (Gorgolewski et al., 2016), reoriented to the axis from commissura anterior (AC) to commissura posterior (PC), corrected for acquisition time (slice timing) and head motion (spatial realignment) using SPM12.
2. *standard GLM analysis*: The first-level design matrix included 1 "condition" regressor for continuous visual stimulation; 48 parametric modulators describing the intensity levels in all sectors; 2 regressors of no interest for the control fixation task; further nuisance regressors for movement parameters and temporal filter; and a constant regressor modelling the implicit baseline.
3. *trial-wise GLM estimation*: The ITEM toolbox function `ITEM_est_1st_lvl` was used to estimate trial-wise response amplitudes where only the continuous visual stimulation

[7] Also see: https://twitter.com/JoramSoch/status/1631568277800378368.

events (100 trials), but not the control fixation task events (Landolt's C) was broken up into trial-wise structure, as our major interest was on predicting local visual contrast from concurrent early visual cortex activity.

4. *trial-wise fMRI decoding*: The ITEM toolbox function `ITEM_dec_recon_SL` was used to perform searchlight-based regression of intensity levels in all sectors from searchlights all over the brain using a searchlight radius of 6 `mm` to yield a correlation coefficient (CC) map for each sector and subject.
5. *group-level analysis*: Finally, CC maps were normalized into the common MNI space and subjected to a repeated-measures ANOVA with visual field radius (4 rings = 4 levels) and visual field angle (12 segments = 12 levels) as within-subject factors. Using suitable contrasts, we were looking for voxels in which the average CC for a subset of sectors was significantly larger than zero (see Figure 4B).

The complete empirical data analysis can be reproduced using MATLAB code available from GitHub (https://github.com/JoramSoch/ITEM-SL-paper).

4.3 Results

Results obtained from the repeated-measures ANOVA matched well-known properties of early visual cortex (see Figure 4B): (i) contralateral processing of visual hemifield, i.e. left visual field activates right visual cortex and vice versa; (ii) representation of visual field half, i.e. top visual stimulation activates medial parts, bottom visual stimulation activates lateral parts of visual cortex; (iii) representation of eccentricity along a posterior-anterior axis, i.e. more outer parts activate more anterior regions; (iv) representation of angular direction along a dorsal-ventral axis, i.e. more bottom parts activate more dorsal regions; and (v) taking (iii) and (iv) together, polar-coordinate representation of the visual field in primary visual cortex (Zeidman et al., 2018).

All the results were significant at $\alpha = 0.05$, whole-brain corrected for family-wise error; except for result (iii) for which uncorrected inference with a significance threshold $\alpha = 0.001$ and an extent threshold $k = 10$ was applied (see Figure 4B).

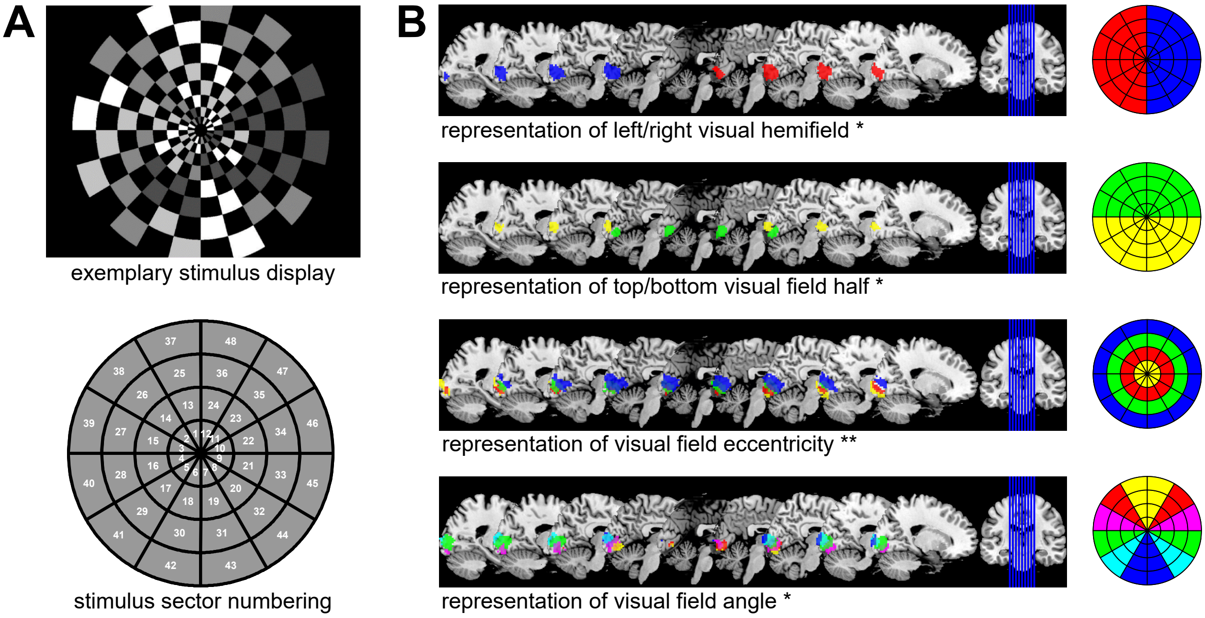
A
exemplary stimulus display
37
48
38
25
36
47
26
35
13
24
39
27
14
23
34
46
15
22
40
28
16
21
33
45
17
20
29
18
19
32
41
30
31
44
42
43
stimulus sector numbering
B
representation of left/right visual hemifield *
representation of top/bottom visual field half *
representation of visual field eccentricity **
representation of visual field angle *

Figure 4: *Empirical validation of searchlight-based ITEM analysis.* **(A)** During fMRI scanning, subjects were stimulated with flickering checkerboard patterns (top) whose illumination intensity changed from trial to trial. The visual field was partitioned into 48 sectors (bottom) organized into 4 rings and 12 segments. Trial-wise intensity levels in all 48 sectors were reconstructed using ITEM-based searchlight decoding from searchlights centered on each voxel (SL radius = 6 mm) and predictive correlations between actual and reconstructed intensities were calculated for each searchlight. **(B)** Then, predictive correlation maps were normalized to standard space and submitted to a repeated-measures ANOVA with eccentricity (4 levels) and angular direction (12 levels) as within-subject factors. Colored voxels indicate searchlights from which the visual contrast in highlighted sectors could be decoded with average predictive correlation significantly greater than zero (* FWE, $p < 0.05$, $k = 0$; ** unc., $p < 0.001$, $k = 10$).

5 Discussion

We have extended inverse transformed encoding models (ITEM), a previously proposed method for dealing with trial-by-trial correlations in fMRI decoding, from region-of-interest (ROI) to searchlight-based (SL) analysis. Whereas earlier contributions to trial-level prediction from fMRI responses have suggested ad-hoc solutions, e.g. estimating each trial's reseponse amplitude using a different model in order to reduce trial-by-trial correlations (Mumford et al., 2012, 2014), the present technique offers a principled approach by accounting for the actual distribution of trial-wise parameter estimates. By going beyond individual ROIs to decoding from the vicinity of each voxel in the brain (see Figure 2), we have demonstrated that SL-based ITEMs can be successfully used for information-based mapping (Kriegeskorte et al., 2006; Haynes et al., 2007) of cognitive functions, e.g. visual perception (see Figure 4).

5.1 Assessment of simulation validation

In our earlier study, results on the relative advantage of ITEM over LS-S, the currently accepted approach, were somewhat inconclusive: Whereas simulation studies found that gains in statistical power and classification accuracy were rather marginal (mostly between 0 and +2%, minimally −0.83%, maximally +8.33% in favor of ITEM; see Soch et al., 2020, Fig. 5), the empirical application resulted in significantly higher predictive correlations for the ITEM approach ($p < 0.001$; ITEM: mean $r = 0.31$; LS-S: mean $r = 0.25$; see Soch et al., 2020, Fig. 7).

We believe that this was due to the simulation study using only univariate signals, therefore producing only mild differences between methodologies (cf. LS-A in Soch et al., 2020, Fig. 5). Here, we have closed this gap and developed a truly multivariate ("searchlight") simulation in which the proposed approach outperformed the state-of-the-art approach in each simulation scenario considered (always $\geq 0\%$, maximally +14% in favor of ITEM; see Figure 3 and Section 3.2).

We hypothesize that the improvement of ITEM over LS-S with increasing number of voxels is due to the fact that the removal of temporal correlation is beneficial for each individual voxel. Thus, the more voxels the multivariate signal consists of, the higher the overall benefit in terms of decoding accuracy will be.

Also note that, like in the previous simulation study (see Soch et al., 2020, Fig. 5), we observed that with low error variance, LS-A outperforms LS-S (see Figure 3, top-left), suggesting that the contamination of LS-A estimates with trial-by-trial correlations is not too harmful when the overall noise level is low, and LS-S (but not ITEM) might actually reduce statistical power compared to the naïve approach.

5.2 Assessment of empirical validation

In our earlier study, we applied ITEM in an ROI-based manner which required a feature selection step in which only voxels responsive to the task as such – no matter what aspect they are responsive to – were filtered out using a Bayesian model selection strategy (see Soch et al., 2020, p. 10). As a consequence of this, contrast levels in some sectors of the visual field could not be reconstructed with satisfying precision – simply because the most task-responsive voxels exclusively represented visual receptive fields near the horizontal midline of the visual vield (see Soch et al., 2020, Fig. 7D).

We believe that this was an unnecessary simplification and here replaced this ROI-based procedure by a searchlight-based approach (ITEM-SL) in which ITEM-style reconstruction was performed for multi-voxel trial-wise response amplitudes extracted from a spherical searchlight with radius $r = 6\ \mathtt{mm}$ around each voxel inside the analysis mask. Following this, the fully parametric model accounting for covariation between all levels of angular direction and field radius was specified and estimated.

ITEM-SL showed very good sensitivity, as it reliably recovered the polar-coordinate organization of receptive field representation in primary visual cortex (see Figure 4B): When correcting for multiple comparisons at the whole-brain level, there were significant differences in reconstruction performance for visual hemifield, field half and field angle. For uncovering the representation of visual field eccentricity, uncorrected inference had to be applied (see Figure 4B, 1st/2nd/4th vs. 3rd row).

ITEM-SL also exhibited high specificity, as no searchlights outside the occipital lobe could be used to decode visual contrast in low-level visual receptive fields and thus, no differences in reconstruction performance between visual field sectors were observed outside the visual cortex. Taken together, ITEM-SL can therefore be a powerful tool in the multivariate localization of cognitive functions in the human brain.

5.3 Assumptions and Limitations

When applying a statistical technique to empirical data, it is important to keep in mind the assumptions made by this method. For ITEM-SL, the two most important assumptions are (i) linearity of the effects and (ii) normality of the errors[8]:

- The multivariate general linear model (MGLM) for fMRI assumes that the effects of the predictor variables (i.e. experimental conditions and modulator variables) on the measured variables (i.e. the measured BOLD signals) are linear in every voxel. For discrete experimental conditions, this means that average responses between conditions can differ. For continuous modulator variables, this means that average responses parametrically follow the levels of the parametric modulator. If there is no effect, this corresponds to a linear effect with a weight of zero.
- The MGLM further assumes that errors, i.e. additive parts of the measurements that cannot be explained by the predictor variables, are matrix-normally distributed with a fixed temporal covariance between trials that is derived from the trial-wise design matrix and an unknown spatial covariance between voxels which is unconstrained and fully estimated (see U and Σ on Figure 2 or in Equation 8). Inside the searchlight, the voxel-by-voxel covariance is assumed to be constant over trials and the trial-by-trial covariance is assumed to be constant over voxels.

If any of the above assumptions is not met, the proposed technique should not be applied – or be applied with caution. Further research is necessary to assess how robust ITEM-SL is relative to violations of these assumptions.

For fMRI, there is good evidence that functional responses are linear relative to stimulation, especially in V1 (Boynton et al., 1996), and when the non-linear character of the hemodynamic response is accounted for (Friston et al., 1998). Furthermore, errors

[8] See Allefeld and Haynes, 2014, pp. 352-354 for a comprehensive discussion of these aspects.

are usually assumed to follow normal distributions based on the central limit theorem and the rationale that every voxel's signal represents a sum of a large number of physiological sources (Allefeld and Haynes, 2014). The most critical dependency of ITEM-SL is therefore whether the assumed trial-to-trial covariance structure holds. Our empirical validation provides grounds to believe this.

Something which ITEM-SL is not capable of is to capture (i) non-linearity in multivariate patterns, e.g. non-linear boundaries between experimental conditions or saturating effects of modulator variables, and (ii) non-stationarity in multivariate patterns, i.e. changes of neural responses or noise structure over time. In such a case, a statistical model explicitly accounting for such possibilities should be employed. For example, support vector machines allow for curved class boundaries using the kernel trick (Boser et al., 1992), but they are, without further extension, also limited to constant responses over time. Moreover, we want to emphasize that any other method, while possibly capturing non-linearity or non-stationarity, will likely be limited in its ability to capture between-trial correlations – which is the critical part of the present contribution.

6 Statements

6.1 Ethics Statement

When acquiring the data set used in this study, written informed consent was obtained from all subjects before participating in the experiments (Heinzle et al., 2011; Soch et al., 2023). The study was approved by the ethics committee of the University of Leipzig, Germany and conducted according to the Declaration of Helsinki.

6.2 Data and Code

SPM12-compatible MATLAB code for searchlight-based ITEM classification and regression has been added to the ITEM toolbox[9]. All code underlying the analyses in this paper is also available from GitHub[10].

The data set used for empirical validation in Section 4 has been BIDS-formatted and uploaded to OpenNeuro[11]. Further instructions on data processing can be found in the readme file of the accompanying repository[12].

6.3 Author Contributions

JS: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.

6.4 Funding Statement

This work was supported by the Bernstein Computational Neuroscience Program of the German Federal Ministry of Education and Research (BMBF grant 01GQ1001C).

6.5 Competing Interests

The authors have no conflict of interest, financial or otherwise, to declare.

6.6 Acknowledgements

The authors would like to thank Carsten Allefeld and John-Dylan Haynes for helpful discussions of the manuscript's contents as well as Jakob Heinzle for acquiring the fMRI data set used for empirical validation.

[9] URL: https://github.com/JoramSoch/ITEM.

[10] URL: https://github.com/JoramSoch/ITEM-SL-paper.

[11] URL: https://openneuro.org/datasets/ds002013.

[12] See: https://github.com/JoramSoch/ITEM-SL-paper/blob/main/README.md.

7 Appendix

A Detailed information for simulation study

The generative model underlying our simulation study can be described as follows. First, in each session from each simulation run, $t = 100$ trials are evenly distributed into 2 experimental conditions or trial types (tt)

$$\mathrm{tt}_i = 1 \quad \text{or} \quad \mathrm{tt}_i = 2 \tag{A.1}$$

where $\mathrm{tt}_i = 1$ and $\mathrm{tt}_i = 2$ for 50 trials each.
Second, voxel-wise averages (i.e. average activity per condition and voxel, not trial-wise responses) are independently sampled from a standard normal distribution

$$\mu_{kj} \sim \mathcal{N}(0, 1),\ k \in \{1, 2\}\,,\ j = 1, \ldots, v \tag{A.2}$$

where k indexes trial type and j indexes voxel.
For a percentage of $1 - r = 80\%$ voxels, activities were equalized between trial types

$$\mu_{2j} = \begin{cases} \mu_{2j}\ , & \text{with probability } r \\ \mu_{1j}\ , & \text{with probability } 1 - r \end{cases} \tag{A.3}$$

where the first line indicates that the activities sampled in (A.2) were kept. As the conditions are thus different with probability r, this value may be seen as the proportion of voxels with information about the conditions.
Third, trial-wise responses are independently sampled from a normal distribution

$$\gamma_{ij} \sim \mathcal{N}(\mu_{\mathrm{tt}_i,j}, \sigma_\gamma^2),\ i = 1, \ldots, t \tag{A.4}$$

where i indexes trial and σ_γ is set to 0.5.
Fourth, inter-stimulus-intervals are independently sampled from a uniform distribution

$$t_i \sim \mathcal{U}(t_{\mathrm{min}}, t_{\mathrm{max}}),\ i = 1, \ldots, t - 1 \tag{A.5}$$

where $t_{\mathrm{min}} \in \{0, 2, 4\}$ and $t_{\mathrm{max}} = t_{\mathrm{min}} + 4$.
Based on the sampled inter-stimulus-intervals (ISIs), the canonical hemodynamic response function (cHRF) as well as stimulus duration t_{dur} and repetition time TR, an $n \times t$ trial-wise design matrix X_t is generated which instantiates the sampled ISIs (see Figure 1A as an example for $t_{\mathrm{isi}} \sim \mathcal{U}(0, 4)$).
Moreover, an $n \times n$ temporal correlation matrix V is generated with entries

$$v_{ij} = \rho^{|i-j|},\ i, j = 1, \ldots, n \tag{A.6}$$

and a $v \times v$ spatial correlation matrix Σ is generated with entries

$$\sigma_{ij} = \nu^{|i-j|},\ i, j = 1, \ldots, v \tag{A.7}$$

where the time constant is $\rho = 0.12$ and the space constant is $\nu = 0.48$. This induced mild temporal correlations between subsequent scans $i = 1, \ldots, n$ and stronger spatial correlations between adjacent voxels $j = 1, \ldots, v$.

Finally, an $n \times v$ noise matrix with standard deviation $\sigma \in \{0.8, 1.6, 3.2\}$ is sampled from the matrix-normal distribution with covariance matrices V and Σ

$$E_t \sim \mathcal{MN}(0_{nv}, \sigma^2 V, \Sigma) \tag{A.8}$$

and the simulated fMRI signals are generated according to the trial-wise GLM as

$$Y = X_t \Gamma + E_t \tag{A.9}$$

where the $t \times v$ matrix of trial-wise response amplitudes is given by

$$\Gamma = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1v} \\ \vdots & \ddots & \vdots \\ \gamma_{t1} & \cdots & \gamma_{tv} \end{bmatrix} . \tag{A.10}$$

The combination of the 3 different options for $t_{\mathrm{min}}/t_{\mathrm{max}}$ and the 3 different options for σ^2 leads to 9 different simulation scenarios (see Figures 3). In each scenario, $N = 1{,}000$ simulations with $S = 2$ sessions per simulation were performed.

After data generation, trial-wise activations are estimated and trial types are decoded. In the “least squares, all” (LS-A) approach, $\hat{\Gamma}$ was obtained via equation (3) and a support vector classification was trained on the estimates from one session to predict trial types in the other session and vice versa. Decoding accuracy was quantified as the proportion of trials correctly assigned to trial types 1 and 2 in the test session. In the “least squares, separate” (LS-S) approach, $\hat{\Gamma}$ was obtained via equation (4) and the same support vector classification was applied. For inverse transformed encoding modelling (ITEM), $\hat{\Gamma}$ was obtained via equation (3) and trial types were decoded via ITEM-style inversion, with decoding accuracy being assessed as described in Section 2.4.

The present simulation differs from the one in Soch et al., 2020 in the following respects:

- The number of trials t was changed from 60 to 100 in order to increase the statistical power for all approaches compared to let t be sufficiently large in comparison with the number of voxels $v = 33$.
- The number of voxels was changed from 1 (univariate signal) to 33 (multivariate signals). This required to specify the spatial covariance for which we assumed that voxels come from a one-dimensional array and voxel noise is more correlated the closer the voxels are to each other (cf. eq. A.7).
- We replaced logistic regression (LR) by support vector classification (SVC) as the decoding algorithm, since LR becomes unstable for larger number of features and since SVC is more widely applied in neuroimaging data analysis.

B A note on parametric modulators in SPM

The fMRI data set analyzed here represents a very unusual experiment in the sense that it can be understood as a single experimental condition (continuous visual stimulation) that is parametrically modulated with 48 variables (sector intensity levels) – which poses challenges for fMRI modelling.

When we started to analyze the data set with SPM, we noted that (i) parametric regressors were correlated to the onset regressor and (ii) parametric regressors are more correlated to each other than modulator variables from which they were generated (see Figure 5A). To us, both seemed to be unintended consequences, since onset and parametric regressors are typically aimed to model orthogonal components of the fMRI signal (condition mean vs. modulator effect) and parametric regressors should usually preserve the correlation of modulator variables.

We found out that this happens, because SPM does not transform modulator variables before HRF convolution, e.g. via mean-centering. Instead of mean-centering, SPM offers to orthgonalize regressors. Upon choosing this option, we noted that (i) parametric regressors were not correlated anymore, neither with the onset regressor nor among each other, but (ii) the closer one gets to the end of the design matrix (e.g. 48th vs. 1st parametric regressor), the lower is the correlation of parametric regressors with the original modulator variable (see Figure 5B).

We found out that this happens, because SPM uses sequential orthogonalization, i.e. the first parametric regressor is orthogonalized with respect to the onset regressors, the second parametric regressor is orthogonalized with respect to the first one, etc. (Ashburner et al., 2021). This has the consequence that, for a large number of parametric modulators, later parametric regressors are less and less veridical and interpretable. Specifically, only contrasts addressing all parametric regressors together are permitted, but not contrasts only looking at a subset of modulator variables.

We then went on to write our own function for creating an HRF-convolved design matrix from onsets, durations and parametric modulators[13]. Instead of sequential orthogonalization, we used mean-centering of each modulator variable to achieve orthogonality with the onset regressor. We found that our approach avoids both problems, i.e. (i) the correlation structure of the parametric regressors is close to that of the underlying modulator variables and (ii) interpretability of parametric regressors relative to modulator variables is preserved (see Figure 5D). The same was observed when mean-centering modulator variables before entering them into SPM and switching of SPM's orthogonalization procedure which was revalidating our approach (see Figure 5C).

In sum, we therefore suggest – especially in designs with multiple modulator variables for an experimental condition – to turn off orthogonalization in SPM and to subtract the mean (or a neutral value, see below) from modulator variables before SPM model specification. This comes at the cost that parametric regressors may still be correlated, but it is worth paying the price, because this yields better interpretability of the parameter estimates for the parametric regressors (Mumford et al., 2015). Also, partial collinearity of parametric regressors – if it is due to the design – is valuable information that can be handled by linear model estimation (Vanhove, 2020).

In the case that the means of modulator variables are different across subjects (or between modulators), another option is to subtract a neutral value rather than the variable mean. For example, if stimulus ratings are collected from subjects during or after the experiment, those ratings will rarely be distributed uniformly for each subject (and likely be distributed differently between subjects). Then, it makes sense to subtract the value corresponding to the neutral rating from the modulator variable. In this case, it is pos-

[13] See function `ITEM_get_des_mat` in the repository https://github.com/JoramSoch/ITEM.

sible that parametric regressors are correlated to their onset regressor, but in exchange, inter-subject interpretability is preserved. In the past, we have successfully applied this strategy to an fMRI episodic memory task (Soch et al., 2021b; Soch et al., 2021a) in which stimulus presentations were parametrically modulated with subsequent memory response ranging between 1 ("the stimulus is new") and 5 ("the stimulus is old"), such that 3 ("I don't know") corresponded to the subtracted neutral response.

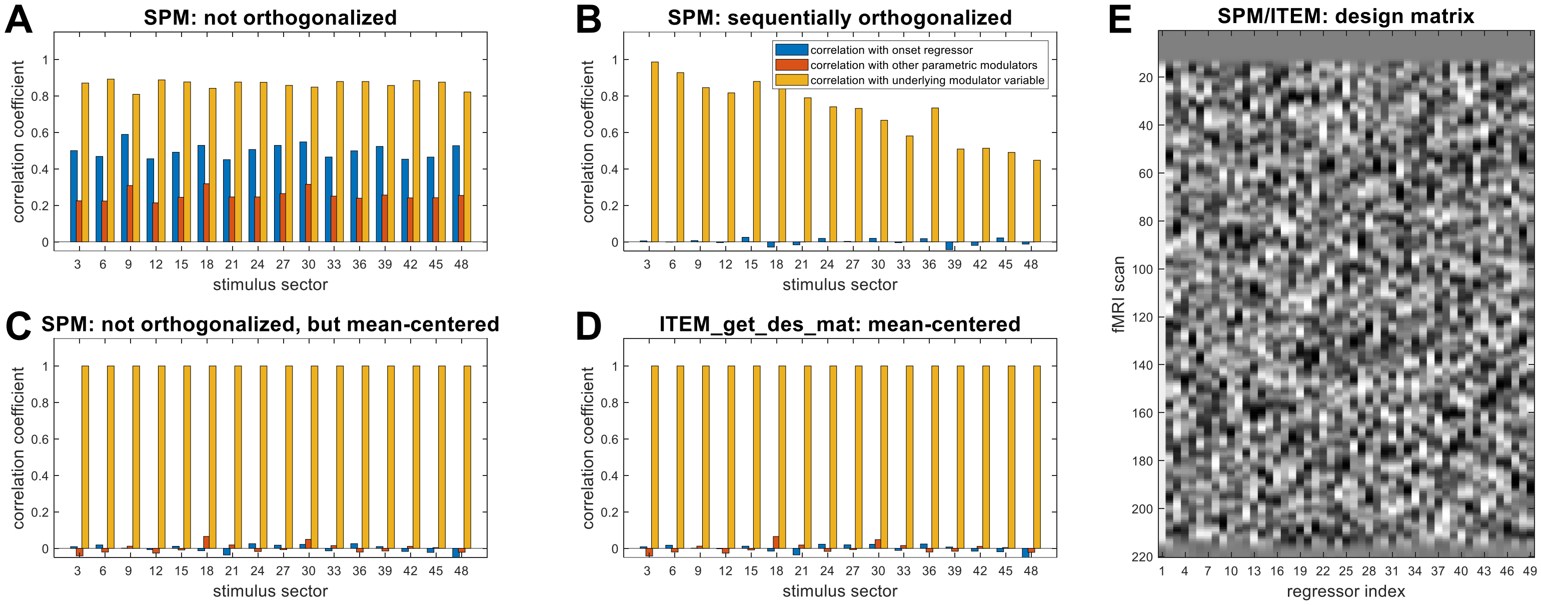

Figure 5: *Effects of orthogonalization on parametric modulators.* Pearson correlation coefficients of all 48 parametric modulator regressors (i) with the onset regressor (blue), (ii) with the other parametric regressors (orange) and (iii) with the original modulator variables (yellow) **(A)** when using default SPM with orthogonalization switched off, **(B)** when using SPM's sequential orthogonalization, **(C)** when using SPM without orthogonalization, but mean-centering beforehand and **(D)** when using the ITEM toolbox script for design matrix creation (which applies mean-centering, but not orthogonalization). For better readability of the figure, only every third parametric modulator is plotted. Note that not orthogonalizing or mean-centering induces correlations with the onset regressor (A, blue) and between the regressors (A, orange), but sequentially orthogonalizing makes parametric regressors become less and less faithful to the underlying modulator variables (B, yellow). **(E)** The design matrix that results from the procedure in C or D, without nuisance variables. Note that there are 49 regressors, the onset regressor for continuous stimulation (labeled as first regressor) and the parametric modulators for sector intensities (labeled with regressor index, if plotted in A-D).

8 References

Allefeld, C. and Haynes, J.-D. (2014). Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. *NeuroImage*, 89:345–357.

Ashburner, J., Friston, K., Penny, W., Stephan, K. E., et al. (2021). SPM12 Manual.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, Pittsburgh Pennsylvania USA. ACM.

Boynton, G. M., Engel, S. A., Glover, G. H., and Heeger, D. J. (1996). Linear Systems Analysis of Functional Magnetic Resonance Imaging in Human V1. *The Journal of Neuroscience*, 16(13):4207–4221.

Brodersen, K. H., Haiss, F., Ong, C. S., Jung, F., Tittgemeyer, M., Buhmann, J. M., Weber, B., and Stephan, K. E. (2011a). Model-based feature construction for multivariate decoding. *NeuroImage*, 56(2):601–615.

Brodersen, K. H., Schofield, T. M., Leff, A. P., Ong, C. S., Lomakina, E. I., Buhmann, J. M., and Stephan, K. E. (2011b). Generative Embedding for Model-Based Classification of fMRI Data. *PLOS Computational Biology*, 7(6):e1002079.

Cox, D. D. and Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) “brain reading”: detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, 19(2):261–270.

Friston, K., Glaser, D., Henson, R., Kiebel, S., Phillips, C., and Ashburner, J. (2002a). Classical and Bayesian Inference in Neuroimaging: Applications. *NeuroImage*, 16(2):484–512.

Friston, K., Penny, W., Phillips, C., Kiebel, S., Hinton, G., and Ashburner, J. (2002b). Classical and Bayesian Inference in Neuroimaging: Theory. *NeuroImage*, 16(2):465–483.

Friston, K. J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M. D., and Turner, R. (1998). Event-Related fMRI: Characterizing Differential Responses. *NeuroImage*, 7(1):30–40.

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., Poline, J.-B., Rokem, A., Schaefer, G., Sochat, V., Triplett, W., Turner, J. A., Varoquaux, G., and Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, 3:160044.

Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: The early beginnings. *NeuroImage*, 62(2):852–855.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, 293(5539):2425–2430.

Haynes, J.-D. (2015). A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. *Neuron*, 87(2):257–270.

Haynes, J.-D. and Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7):523–534.

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., and Passingham, R. E. (2007). Reading Hidden Intentions in the Human Brain. *Current Biology*, 17(4):323–328.

Heinzle, J., Kahnt, T., and Haynes, J.-D. (2011). Topographically specific functional connectivity between visual field maps in the human brain. *NeuroImage*, 56(3):1426–1436.

Henson, R., Rugg, M. D., and Friston, K. J. (2001). The choice of basis functions in event-related fMRI. *NeuroImage*, 13(6):149.

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

Mennes, M., Kelly, C., Colcombe, S., Castellanos, F. X., and Milham, M. P. (2013). The Extrinsic and Intrinsic Functional Architectures of the Human Brain Are Not Equivalent. *Cerebral Cortex*, 23(1):223–229.

Monti, M. (2011). Statistical Analysis of fMRI Time-Series: A Critical Review of the GLM Approach. *Frontiers in Human Neuroscience*, 5.

Mumford, J. A., Davis, T., and 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., Poline, J.-B., and Poldrack, R. A. (2015). Orthogonalization of Regressors in fMRI Models. *PLOS ONE*, 10(4):e0126255.

Mumford, J. A., Turner, B. O., Ashby, F. G., and 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., and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9):424–430.

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

Smith, S. M. (2004). Overview of fMRI analysis. *The British Journal of Radiology*, 77(suppl_2):S167–S175.

Soch, J., Allefeld, C., and Haynes, J.-D. (2020). Inverse transformed encoding models – a solution to the problem of correlated trial-by-trial parameter estimates in fMRI decoding. *NeuroImage*, 209:116449.

Soch, J., Görgen, K., Heinzle, J., and Haynes, J.-D. (2023). A tightly controlled fMRI dataset for receptive field mapping in human visual cortex. *Data in Brief*, 47:109018.

Soch, J., Richter, A., Schütze, H., Kizilirmak, J. M., Assmann, A., Behnisch, G., Feldhoff, H., Fischer, L., Heil, J., Knopf, L., Merkel, C., Raschick, M., Schietke, C., Schult, A., Seidenbecher, C. I., Yakupov, R., Ziegler, G., Wiltfang, J., Düzel, E., and Schott, B. H. (2021a). A comprehensive score reflecting memory-related fMRI activations and deactivations as potential biomarker for neurocognitive aging. *Human Brain Mapping*, 42(14):4478–4496.

Soch, J., Richter, A., Schütze, H., Kizilirmak, J. M., Assmann, A., Knopf, L., Raschick, M., Schult, A., Maass, A., Ziegler, G., Richardson-Klavehn, A., Düzel, E., and Schott, B. H. (2021b). Bayesian model selection favors parametric over categorical fMRI subsequent memory models in young and older adults. *NeuroImage*, 230:117820.

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

Vanhove, J. (2020). Collinearity isn't a disease that needs curing. preprint, PsyArXiv.

Weeda, W. (2018). Estimating Single-trial BOLD Amplitude and Latency in Task-based fMRI Data with an Unknown HRF. In *Organization for Human Brain Mapping*, volume 2018, page Poster #2532, Singapore. OHBM.

Zeidman, P., Silson, E. H., Schwarzkopf, D. S., Baker, C. I., and Penny, W. (2018). Bayesian population receptive field modelling. *NeuroImage*, 180:173–187.