Exploring spatio-temporal neural correlates of face learning

Ying Yang *

Center for the Neural Basis of Cognition
Machine Learning Department
Carnegie Mellon University,
Pittsburgh, PA, 15213, USA

May 25, 2015

Abstract

Background: Humans have great expertise in learning faces. Understanding the neural basis of such expertise is one of the fundamental goals of cognitive science. With functional neuroimaging, researchers have discovered a spatial network of face-sensitive regions in the brain, and some temporal signatures of face-processing. However, we still lack a joint spatio-temporal characterization, in the context of learning new faces.

Aim: In this work, we aim to explore the spatio-temporal neural correlates of face learning, by analyzing magnetoencephalography (MEG) recordings. To be more concrete, we aim to test whether the neural signals in the face-sensitive regions change during learning, and if yes, whether the temporal patterns of changes are different in the regions at various spatial locations.

Data: The MEG data were recorded when human subjects learned to distinguish two categories of computer-generated faces, with trial-by-trial feedback. In about 700 trials, the behavioral accuracy rose from chance to $\geq 70\%$. The face-sensitive brain regions were also defined for each subject according to an independent dataset.

Methods: To infer neural responses in the brain space based on MEG sensor data, we needed to solve the inverse of an underspecified linear problem (known as `source localization`). We ran the following two analyses on both the MEG sensor data and the source data in the face-sensitive regions, using source localization. (1) We regressed the MEG recordings against behavioral learning curves, to understand how correlated the trial-by-trial neural signals were with learning. For better inference in the brain space, we also developed a novel structured-sparsity-inducing regression model in the source localization framework. (2) With multivariate discriminant analysis, we examined whether the neural signals were able to discriminate the two face categories, and more importantly whether the discriminability changed between the early and late trials of learning.

Results: (1) In the regression analysis, we found that the MEG sensor data were significantly correlated with behavioral learning curves, and the effect was predominant in 150 to 300 ms after the stimulus onset. In further regression analysis in the face-sensitive regions, our new model produced smoother and thus more interpretable results than an alternative benchmark. In results by both methods, we observed that the correlation with behavioral learning was localized in the ventral face-sensitive regions, whereas regions in the temporal and frontal lobes did not show as strong effects. (2) With the discriminant analysis, we found that the neural signals recorded by MEG were able to differentiate the two face categories starting from about 140 ms and lasting up to 560 ms. Such discriminability was localized in most of the face sensitive regions. However, in a comparison between the early and late stages of learning, we did

*In collaboration with Yang Xu, Carol A. Jew, John A. Pyles, Michael J. Tarr and Robert E. Kass.

not find significant changes in discriminability that survived our permutation tests, although there appeared to be some trend in some of the ventral regions.

Conclusions: Our regression results revealed the spatio-temporal dynamics in the face-sensitive areas during face learning, on a finer-grained level than previous work. The temporal correlation near 150-300 ms may reflect changes in two waveforms in the literature: the M170 at around 170 ms, indexing recognition of individual faces, and the N250 at around 250 ms, indexing face familiarity. Spatially, the learning effects were localized in the ventral regions that are involved in processing visual properties of faces, but not as strongly in the non-ventral regions that are relevant to the social or semantic properties, suggesting the importance of changes in visual processing in face learning. Although we did not find significant changes of discriminability of face categories during learning, our discriminant analysis demonstrated that many of the face-sensitive regions encoded information to differentiate the face categories, and also provided a detailed spatio-temporal profile of such discriminability in these regions.

1 Introduction

Humans have great expertise in face recognition – we can detect and individuate faces accurately within half a second. Such proficiency is highly dependent on our ability to learn novel faces. Understanding the neural mechanisms of face-learning is a key step of understanding human visual cognition, which is fundamental in cognitive science.

Using high-spatial-resolution functional magnetic resonance imaging (fMRI), researchers have identified a network of regions in the brain that are more sensitive to faces than to other visual stimuli, and are structurally and functionally connected with each other [Ishai, 2008, Pyles et al., 2013]. (See Figure 2 for example locations of these regions) A group of such face-sensitive regions is on the bilateral ventral surfaces of the brain, including the "occipital face area" near the inferior occipital gyrus (IOG) [Pitcher et al., 2011], the "fusiform face area" in the middle fusiform gyrus (mFUS) [Kanwisher et al., 1997], and the anterior inferior temporal lobes (aIT) [Kriegeskorte et al., 2007, Nestor et al., 2008, Rajimehr et al., 2009]. These regions lie in the ventral visual pathway, which is the main pathway for object recognition [Mishkin et al., 1983] and is generally organized in a hierarchical manner: the very posterior areas are sensitive to low-level visual features such as edge orientations; more anterior areas are sensitive to shapes and contours, and ultimately the diagnostic features to recognize "what we see" [DiCarlo and Cox, 2007]. Therefore the ventral face-sensitive regions are also hypothesized to process the visual information of faces with such a hierarchy. The face-network also includes areas in the superior temporal sulcus (STS), hypothesized to process the social aspects of faces (e.g. expression and gaze), and areas in the inferior frontal gyrus (IFG) and orbitofrontal cortex (OFC), hypothesized to process the semantic or valance components of faces [Ishai, 2008].

However, due to the low temporal resolution, it is hard to show the temporal dynamics of face processing with fMRI. A second line of research used megnetoencephalography (MEG) and electroencephalogram(EEG) with millisecond resolution, and identified relevant temporal waveforms at: about 100 ms after the stimulus onset(M1, face detection), 170 ms (M170/N170, face detection and identification [Liu et al., 2000]) and 250 ms (N250, familiarity [Tanaka et al., 2006]), as well as some later waveforms after 300 ms that are associated with semantic processing [Barrett and Rugg, 1990]. Yet the spatial resolution of EEG/MEG highly depends on the challenging `source localization` procedure (see Section 2) to project the data into the brain space, therefore most MEG/EEG studies did not focus on spatial inference. Although some waveforms such as M170 were very roughly localized to the fusiform gyrus [Deffke et al., 2007], we still lack a detailed mapping

between the temporal signatures and the spatial face-sensitive regions.

The aforementioned work mainly focused on face processing per se, and only a few neuroimaging studies directly looked into the changes of neural signals in face learning. Two fMRI studies [DeGutis and D'Esposito, 2007, DeGutis and D'Esposito, 2009] demonstrated that the face-sensitive network was involved in face-learning, and two EEG experiments [Su et al., 2013, Itz et al., 2014] showed that waveforms at 170 ms (M170), 200 ms, and ≥250 ms (N250) changed after learning. However, these experiments only exploited either the spatial or temporal resolution of the imaging technique, and were not able to describe the learning process in spatio-temporal details. Additionally, most of these studies only recorded data at the beginning and the end of learning, ignoring the middle phase where the change with learning actually happened.

To understand the spatio-temporal changes of neural signals during face learning, especially in the face-sensitive regions, our lab conducted an MEG experiment, where human subjects learned to distinguish two categories of computer-generated novel faces, with trial-by-trial feedback. The task was completed in about 700 trials in one experimental session, and we were able to observe the entire learning process with high temporal resolution. Our lab also defined the face-sensitive regions for each subject, based on an independent dataset, so that we were able to achieve spatial inference in these regions. Since such inference is highly dependent on the `source localization` procedure, we briefly review the related background in Section 2.

2 Background on source localization in MEG

There are about 300 MEG sensors distributed on a helmet that surrounds the subject's head. These sensors continuously measure magnetic signals induced by electrical currents that are generated by a large number of local neurons [Hamalainen et al., 1993]. The MEG machine records the sensor read-out as a multi-dimensional time series. Given the positions of sensors and the subject's head, the sensor data can be approximated by a linear transform of the underlying neural currents in the brain space (termed as the `source space`), according to Maxwell's equations [Hamalainen and Ilmoniemi, 1994]. To make spatial inference about the neural signals, we need to solve the inverse of this linear problem, and this procedure is termed as `source localization`. In earlier MEG literature, it was assumed that the signals came from only a few current dipoles in the source space, and nonlinear regression was used to fit the positions and directions of the dipoles [Scherg and Von Cramon, 1985]. However, this assumption is unrealistic in the sense that most cognitive tasks involve multiple brain areas, and the few equivalent dipoles may not reflect these true locations. Alternatively, more recent methods define the `source space` as $\geq 10^4$ discrete `source points` covering the brain with certain spatial resolution (e.g. 5-7 mm). Since we are mostly interested in the cortical brain activity, and MEG sensors are mostly sensitive to groups of pyramidal cells, which align perpendicularly to the cortical surfaces, the `source points` are often defined on the cortical surfaces, and aligned to the local normal directions as well. In this case, we aim to make inference about all the source points simultaneously. However, because the number of sensors ($\sim 10^2$) is much smaller than the number of source points ($\geq 10^4$), the inverse problem is highly underspecified, and proper regularization is needed.

Many popular `source localization` methods use an L_2 penalty for regularization at each time point (minimum-norm estimate, MNE [Hamalainen and Ilmoniemi, 1994], dynamic statistical para-

metric mapping, dSPM, [Dale et al., 2000] and methods alike [Pascual-Marqui, 2002]). Recently, more sophisticated penalty structures have been applied. For example, noticing that L_2 penalty is equivalent to putting a simple Gaussian prior with zero mean and a diagonal covariance matrix on the source solutions, researchers have made improvements with more generic Bayesian regularization, and building in other prior knowledge such as results from fMRI [Mattout et al., 2006, Henson et al., 2011]. On the other hand, sparsity-inducing penalties such as L_{21} or "group lasso" penalties have been applied with the assumption that only a few `source points` in the source space are active, implemented in convex optimization algorithms[Gramfort et al., 2012] or greedy algorithms [Babadi et al., 2014].

Another issue in `source localization` is whether to incorporate the temporal structure of the time series. MNE and methods alike solve the inverse problem at each time point without considering the temporal smoothness, and thus may result in source solutions with a lot of high-frequency noise. [Galka et al., 2004, Lamus et al., 2012] have exploited linear state-space models such as Kalman filters and obtained relatively smooth estimates. However, such models assume the source time series to be stationary, which is often not satisfied. To model the non-stationary source time series, [Gramfort et al., 2013] represented them by sparse time-frequency components from short-time Fourier transform (STFT), and obtained solutions that were spatially sparse and temporally smooth.

Here, with our purpose of understanding how the neural signals change with behavioral learning, it is natural to do a regression of the source time series across trials against the behavioral accuracy, which increased from early trials to late trials. Since most of the existing methods were designed for single-trial data, we need a two-step procedure for such analysis: (1) obtain source time series in each trial; (2) run regression. But besides MNE and methods alike, it is often hard to use more sophisticated methods such as [Gramfort et al., 2013] directly in the two-step procedure, because the sparse structure may not be consistent across trials. Instead, a one-step model that embeds the regression in the source localization model is more desirable. Therefore, we adapted the model from [Gramfort et al., 2013] to a new one-step regression model for our analysis.

3 Problem definition

In this project, we were interested in how neural signals measured in MEG changed when subjects learned to distinguish two categories of novel faces, especially in the face-sensitive regions (regions of interests or ROIs) defined for the same subjects, based on an independent dataset. We focused on the recordings from -140 ms to about 560 ms (0 being the stimulus onset), where most of related temporal signals happened. The preprocessed data were a three dimensional tensor $\boldsymbol{M} \in \mathbb{R}_{n \times T \times q}$, from n MEG sensors, T time points and q trials. In each trial, the subjects were shown one exemplar face from one category, asked to respond which category it was, and given feedback after the response. In about 700 trials, their behavioral accuracy increased from chance to $\geq 70\%$. We recorded the behavioral responses as well as the face category labels in all trials. We used the following two types of analysis to examine the changes of neural signals in learning. In both approaches, we did the analysis on the sensor data first as a sanity check, and then applied source localization to do analysis in the ROIs.

3.1 Regression against behavioral learning curves

Since we recorded the trial-by-trial behavioral accuracy in binary values, we were able to estimate a smooth behavioral learning curve $\boldsymbol{x}(r)$, as a function of trial number r. To characterize whether the neural signals were correlated with the behavioral learning, it is natural to regress the trial by trial MEG data ($\boldsymbol{M}[:,:,r], r = 1, \cdots q$) on the behavioral learning curve $\boldsymbol{x}$. It is non-trivial to characterize how a time series correlates with a regressor. The magnitudes, latencies, or shapes of waveforms at various time windows can all change. For simplicity, we reduced the problem to regressing the values at each time point across trials against the behavioral learning curve, and then examining statistics of the regression coefficients. By doing so, we aimed to investigate (1) whether the neural signals in the face-sensitive ROIs were correlated with the behavioral learning curve; (2) if so, in what time window the correlation effect happened, and (3) whether different ROIs showed different temporal patterns of the correlation effect.

To obtain the correlation effects for the ROIs in the brain space, we adapted the model in [Gramfort et al., 2013] to produce a new short-time Fourier transform regression model (STFT-R) in the source localization framework. We represented the trial-by-trial time series for each source point with time-frequency STFT components and assumed that each component had a linear relationship with the regressors ($\boldsymbol{x}$ or multidimensional regressors in more general cases). To regularize the underspecified problem when solving the regression coefficients of STFT components, we designed a hierarchical group lasso (L_{21}) penalty [Jenatton et al., 2011] of three levels to induce structured sparsity (1) on the source points outside the ROIs; (2) over time and frequency on the STFT components; (3) finally for different regressors (if applicable) for each STFT component. We derived an efficient algorithm to solve STFT-R, and compare STFT-R with an alternative two-step procedure using MNE.

3.2 Discriminant analysis

We also examined whether the neural signals recorded in MEG could differentiate the two face categories as the subjects did, with multivariate discriminant analysis. Then we examined whether the discriminability in different ROIs changed from early trials to late trials, that is, whether the representation of the two categories of faces became more separable or less separable with learning.

4 Materials and methods

4.1 Data description

The MEG dataset was collected by Xu et al. in 2013, and details are documented in [Xu, 2013]. Here we briefly describe the experiment. Ten right-handed adults (6 females and 4 males) participated in the experiment with written consent. All procedures were approved by the Institutional Review Board of Carnegie Mellon University and University of Pittsburgh. In the experiment, the subjects were shown two categories of computer-generated faces. Each category had 364 exemplars, (728 in total for both categories). The faces had varied eye-sizes and mouth-widths, and the exemplars were sampled from a systematic grid in the two dimensional eye-size/mouth-width space, with a decision boundary (see Figure 1). Exemplars in Category A featured in larger eyes and smaller mouths than those in Category B. All the exemplars were novel faces to the subjects. In each trial, a category label (A or B) was provided by an audio cue, then one exemplar was presented

for 900 ms, and the subject had up to 900 ms to report whether the category of the shown face matched the audio cue by pressing "yes" or "no" buttons. Feedback ("corrected", "wrong", or "too slow") was provided after the response, such that the subject could learn the two categories in an online way. The audio cue was not associated with the true category label of the exemplar. Each of the 728 exemplars was presented once, and the position of the exemplar in the two-dimensional feature space was pseudo-randomized, and numbers of "A" and "B" were equal within every 20 trials. There were 728 trials together, and they were divided into 8 runs, where each run had 91 trials. More detailed timing of each trial is shown in Figure 1.

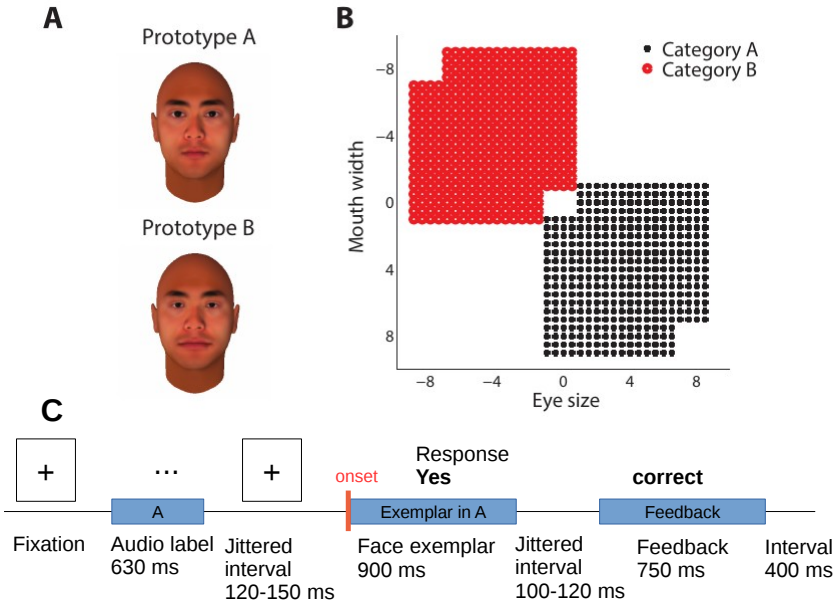


Figure 1: Stimulus design and trial structure [Xu, 2013]. A, prototypes of the two face categories. B, the two-dimensional (eye-size and mouth-width) feature space. Each point represents one exemplar. C, the trial structure.

During the entire experiment, the neural signals of the subjects were recorded in a 306-channel whole-head MEG system (Elekta Neuromag, Helsinki, Finland). The data were acquired at 1 kHz, high-pass filtered at 0.1 Hz and low-pass filtered at 330 Hz. Electrooculogram (EOG) and electrocardiogram (ECG) were recorded with three pairs of electrodes, such that we could estimate eye movements and heartbeats, and remove artifacts caused by them in the MEG recordings. Empty room MEG data was also recorded in the same experimental session, and used to estimate the covariance matrix of the sensor noise. For each subject, the behavioral response and reaction time for each trial were also recorded. In addition, a high-resolution T1-weighed anatomical MRI was obtained, and processed in an automatic cortical-surface-parcellation software (Freesurfer). This anatomical scan of the brain was used to construct the source space in source localization.

4.2 Data preprocessing

We preprocessed the MEG data with the `MNE` software and `MNE-python` package [Gramfort et al., 2014]. The MEG raw data was firstly futher filtered by a 1-110 Hz bandpass filter, and then a notch filter at 60 Hz to reduce the power-line interference. Secondly we applied temporal signal-

space separation (tSSS) [Taulu and Simola, 2006] to the MEG data, using the `MaxFilter` software provided by Elekta. This step further removed the noise that came from outside the MEG helmet. Thirdly, we also used the signal-space projection (SSP) method in `MNE`, where we constructed a low-dimensional linear space based on the empty room sensor noise, and removed the projection on this subspace from the experimental MEG recordings. Finally, we performed an independent component analysis (ICA) on the experimental recordings, and removed the components that were highly correlated with eye blinks and heartbeats detected by EOG and ECG recordings, using the standard script in `MNE-python` [1]. For each trial, we only selected the sensor data from 140 ms before the face stimulus onset (-140 ms) to 560 ms or 580 ms after the stimulus onset, and defined the $-140 \sim -40$ ms time window as the baseline. For each sensor, the mean of the baseline was subtracted for all time points, separately for each trial. We call these trial-by-trial sensor data `epochs`. We prepared different versions of epochs for different analyses.

- `Down-sampled epochs`: the epochs were down-sampled with 100 Hz sampling rate to reduce computational burdens; they were used in the regression analysis, intentionally not smoothed, to test if the results would be different between STFT-R and the alternative two-step MNE.
- `Smoothed epochs`: the epochs were first smoothed with a 50 ms Hanning window, to further reduce the high-frequency noise, then down-sampled with 100 Hz sampling rate; they were used in the discriminant analysis.

4.3 Estimating the behavioral learning curve

A logistic regression model was used to estimate the behavioral learning curves for individual subjects, based on the binary (correct or incorrect) behavioral responses for all 728 trials. We used Legendre polynomials of order 5 as a relatively non-parametric basis set in the logistic regression. Because subjects might have different learning rates for different face categories, we also included an interaction term of the basis with the face category of each trial in the regression. In such a way, we estimated two different learning curves for the two categories (See Figure A.1). Subject `s9` had almost flat behavioral learning curves for both categories, thus was excluded in the analysis.

4.4 Regions of interest (ROIs)

The definition of face-sensitive regions of interest (ROIs) was described in details in Chapter 4, Appendix I in [Xu, 2013]. In an independent experiment, the same subjects were presented color images of four conditions: faces, houses, everyday objects and scrambled objects. The stimuli were presented in blocks, where each block consisted of 16 images of one condition, and each stimulus was presented for 800 ms, with 200 ms inter-stimulus intervals. Xu et al. ran the following spatio-temporal excursion procedure to define the ROIs as spatially clustered source points that could differentiate faces and everyday objects. A 40 ms time window near 170 ms after the stimulus onset was first detected based on Hotelling's T-squared test of the sensor responses to faces versus everyday objects, then they obtained the MNE source solutions for each trial, and averaged the time series within the window for each source point, thirdly they ran a searchlight Hotelling's T-squared test, for each source points and its adjacent neighbors. They finally defined the contiguous source points where the p-values < 0.001. Six to eleven ROIs were found for individual subjects. Figure 2 illustrates the face-sensitive ROIs in one example subject (`s8`). Table 1 shows the abbreviations of ROIs and the number of source points in each ROI. (ventral visual regions:

[1] The ECG and EOG data for one subject (`s4`) were missing, so we skipped this step for this subject.

IOG, inferior occipital gyrus; mFUS, middle fusiform; aIT, anterior inferior temporal lobe. temporal regions: STS, superiortemporal sulcus. frontal regions: IFG, inferior frontal gyrus; OFC, orbitofrontal gyrus). The suffix "_L" indicates that it is in the left hemisphere, and "_R" in the right hemisphere. n_{part} in the last row indicates the number of subjects where the ROI was present.

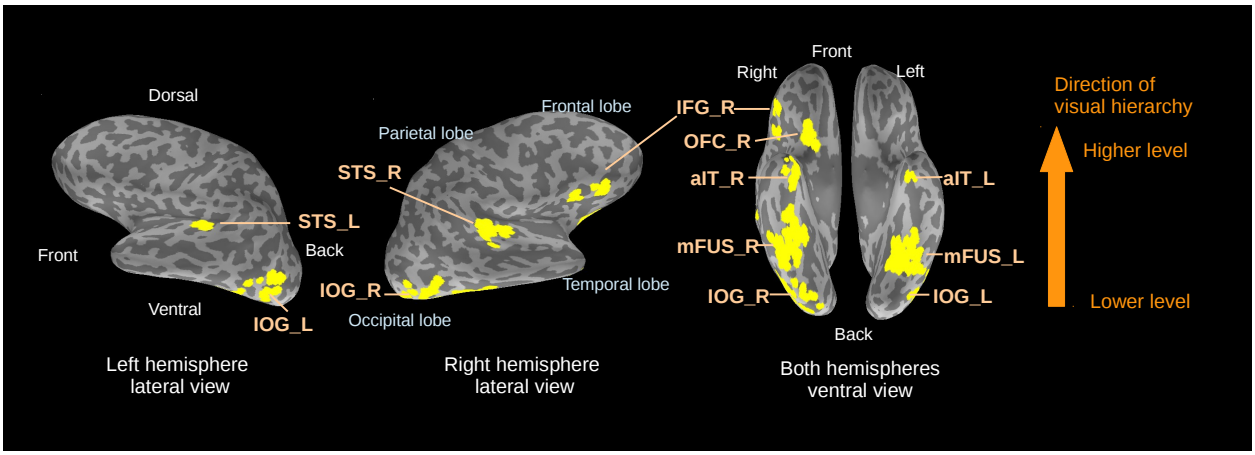


Figure 2: Illustration of face-sensitive ROIs in one example subject (`s8`).

Table 1: Number of source points in face-sensitive ROIs

	IOG_L	IOG_R	mFUS_L	mFUS_R	aIT_L	aIT_R	STS_L	STS_R	IFG_L	IFG_R	OFC_R
s1	20	18	13	68	5	17	0	10	7	7	3
s2	3	20	0	38	0	0	0	9	3	3	3
s3	20	21	44	50	6	0	5	3	19	0	0
s4	20	19	54	55	0	0	17	12	3	10	0
s5	18	14	25	43	7	0	3	0	16	0	0
s6	0	0	40	19	17	0	9	0	11	4	0
s7	0	20	46	28	0	9	0	14	0	6	0
s8	14	23	37	47	5	12	6	18	0	12	3
s10	19	14	40	19	0	0	0	13	7	19	0
n_{part}	7	8	8	9	5	3	5	7	7	7	3

4.5 Source localization

We constructed the source space as about 6000-7000 discrete source points almost evenly distributed on the bi-hemispheric cortical surfaces, with 7 mm separation on average. The direction of source points were constrained to align with the normal direction of the cortical surfaces. The linear operator that projected source signals to the sensor space (also known as the `forward matrix`) was computed using Maxwell equations by the `MNE` software, based on the position of the head and MEG sensors. We computed a forward solution for each of the 8 runs, to correct for run-to-run head movement, when applying both MNE (and dSPM) and STFT-R. But here, we use a constant forward matrix $\boldsymbol{G}$ for all the trials just for simplicity of description.

For the following text, we introduce some notations. We use bold letters to denote matrices and vectors. For a matrix $\boldsymbol{A}$, we denote the ith row by $\boldsymbol{A}[i,:]$, the jth column by $\boldsymbol{A}[:,j]$, and the entry in the ith row and jth column by $\boldsymbol{A}[i,j]$. We also use bold letters to denote a tensor (or a

three-dimensional matrix), for a tensor $\boldsymbol{M}_{m\times n\times q}$, the rth layer is denoted as $\boldsymbol{M}^{(r)}$ or $\boldsymbol{M}_r$.

Assume we have n sensors, m source points, T time points in each trial, and q trials together ($n = 306, m \approx 7000$). Then both the sensor data $\boldsymbol{M} \in \mathbb{R}^{n\times T\times q}$ and source time series $\boldsymbol{Y} \in \mathbb{R}^{m\times T\times q}$ are three dimensional tensors. Let the superscript $^{(r)}$ denote the data in the rth trial. With $\boldsymbol{G} \in \mathbb{R}^{n\times m}$ being the forward matrix, and $\boldsymbol{E}^{(r)} \in \mathbb{R}^{n\times T}$ being sensor noise of the rth trial, we have the following forward problem for each trial

$$\boldsymbol{M}^{(r)} = \boldsymbol{G}\boldsymbol{Y}^{(r)} + \boldsymbol{E}^{(r)} \tag{1}$$

We assume the noise to be independent for each trial, and temporally independent within each trial. In addition, in preprocessing, we estimated the covariance matrix of the sensors using empty room recordings, and pre-whitened the sensor data with the covariance matrix. So we assume that each column of $\boldsymbol{E}^{(r)}$ has an identity covariance matrix.

4.5.1 Minimum-norm estimate (MNE) and dynamic statistical parametric mapping (dSPM)

The minimum-norm estimate (MNE) is one of the most commonly used source localization methods [Hamalainen and Ilmoniemi, 1994]. It solves the linear regression in (1) with a simple L_2 penalty. At a time point t in the rth trial, with the penalty λ, we have

$$\hat{\boldsymbol{Y}}^{(r)}[:,t] = \arg\min(\frac{1}{2}\|\boldsymbol{M}^{(r)}[:,t] - \boldsymbol{G}\boldsymbol{Y}^{(r)}[:,t]\|^2 + \lambda\|\boldsymbol{Y}^{(r)}[:,t]\|_2^2)$$

The closed-form solution can be obtained by a simple linear operator on the sensor data in each trial,

$$\hat{\boldsymbol{Y}}^{(r)} = (\boldsymbol{G}^T\boldsymbol{G} + \lambda\boldsymbol{I})^{-1}\boldsymbol{G}^T\boldsymbol{M}^{(r)}$$

as long as λ is predefined, the linear operator $\texttt{Inv} = (\boldsymbol{G}^T\boldsymbol{G}+\lambda\boldsymbol{I})^{-1}\boldsymbol{G}^T$ can be computed only once, and applied to all trials.

For the trial-by-trial regression analysis in the source space, we used the following two-step model (termed as MNE-R): first, we obtained the source solution for each trial with MNE. Secondly, we regressed the solutions at each time point across trials against the regressors. To make this method comparable with our STFT-R model (defined in Section 4.5.2) in the time-frequency domain, in the actual implementation, we transformed the source solutions in the time domain to time-frequency domain with STFT. In such a case, because STFT did not depend on the regressors, regressions in the STFT domain produced equivalent coefficients as regressions in the time domain. We reconstructed the fitted sensor time series with the regression coefficients, and selected the best penalty parameter λ by minimizing two-fold cross-validated mean squared error of reconstructed sensor data across trials.

The dynamic statistical parametric mapping (dSPM) [Dale et al., 2000] is also a common variant of MNE. It normalizes the estimated source values to dimensionless statistical test variables. At each time point, assume that the variance of a source point is only a propagation of the sensor covariance matrix $\boldsymbol{I}$, then the covariance matrix of MNE source solution is $\texttt{Inv}^{\boldsymbol{T}}\texttt{Inv}$. dSPM normalizes the MNE solutions for each source point by dividing its expected standard deviation, obtained from the diagonal elements of $\texttt{Inv}^{\boldsymbol{T}}\texttt{Inv}$. Such normalization improves the MNE solution by reducing biases

to superficial source points. We obtained the dSPM solution from the `smoothed epochs` and used them in the discriminant analysis, where λ was set to 1. [2]

4.5.2 Short-time Fourier transform regression (STFT-R)

Short-time Fourier transform (STFT) In our model, we use the STFT implemented in [Gramfort et al., 2013]. Given a time series $\boldsymbol{U} = \{U(t), t = 1, \cdots, T\}$, a time step τ_0 and a window size T_0, we define the STFT as

$$\Phi(\{U(t)\}, \tau, \omega_h) = \sum_{t=1}^{T} U(t) K(t-\tau) e^{(-i\omega_h)} \tag{2}$$

for frequencies $\omega_h = 2\pi h/T_0, h = 0, 1, \cdots, T_0/2$ and time points $\tau = \tau_0, 2\tau_0, \cdots n_0\tau_0$, where $K(t-\tau)$ is a window function defined on T_0 time points and centered at τ, and $n_0 = T/\tau_0$. We concatenate STFT components at different time points and frequencies into a single vector in $\boldsymbol{V} \in \mathbb{C}^s$, where $s = (T_0/2+1) \times n_0$. Following notations in [Gramfort et al., 2013], we also call the $K(t-\tau)e^{(-i\omega_h)}$ terms STFT dictionary functions, and use a matrix's Hermitian transpose $\boldsymbol{\Phi^H}$ to denote them, i.e. $(\boldsymbol{U}^T)_{1\times T} = (\boldsymbol{V}^T)_{1\times s}(\boldsymbol{\Phi^H})_{s\times T}$.

Model Let $\boldsymbol{\Phi^H} \in \mathbb{C}^{s\times T}$ be s pre-defined STFT dictionary functions at different frequencies and time points. Suppose we have p regressors (e.g. a behavioral learning curve, or non-parametric spline basis functions), we write them into a design matrix $\boldsymbol{X} \in \mathbb{R}^{q\times p}$, which also includes an all-one column to represent the intercept. Besides the all-one column, all other columns have zero means. Let the scalar $X_k^{(r)} = \boldsymbol{X}[r,k]$ be the value of the kth regressor in the rth trial. When we represent the time series of the ith source point with STFT, we assume each complex STFT component is a linear function of the p regressors: the jth STFT component in the rth trial is $\sum_{k=1}^{p} Z_{ijk} X_k^{(r)}$, where the regression coefficients Z_{ijk}'s are to be solved. We use a complex tensor $\boldsymbol{Z} \in \mathbb{C}^{m\times s\times p}$ to denote the Z_{ijk}'s, and use $\boldsymbol{Z}_k \in \mathbb{C}^{m\times s}$ to denote the kth layer of $\boldsymbol{Z}$. Our STFT-R model reads

$$\boldsymbol{M}^{(r)} = \boldsymbol{G}\left(\sum_{k=1}^{p} X_k^{(r)} \boldsymbol{Z}_k\right)\boldsymbol{\Phi^H} + \boldsymbol{E}^{(r)} \quad \text{for} \quad r = 1, \cdots, q.$$

where the sensor error $\boldsymbol{E}^{(r)} \in \mathbb{R}^{n\times T}$ is an independently and identically distributed random matrix for each trial. To solve $\boldsymbol{Z}$, we minimize the sum of squared prediction error across q trials, with a hierarchical L_{21} penalty Ω on $\boldsymbol{Z}$:

$$\min_{\boldsymbol{Z}} \left(\frac{1}{2} \sum_{r=1}^{q} \|\boldsymbol{M}^{(r)} - \boldsymbol{G}(\sum_{k=1}^{p} X_k^{(r)} \boldsymbol{Z}_k)\boldsymbol{\Phi^H}\|_F^2 + \Omega(\boldsymbol{Z}, \alpha, \beta, \gamma, \boldsymbol{w}) \right) \tag{3}$$

where $\|\cdot\|_F$ is the Frobenius norm and

[2] We did not use dSPM in the two-step regression model, because given λ, the normalization is the same across all trials and time points, and using dSPM in the two-step regression model gives the same result as MNE up to a constant factor, which is essentially equivalent.

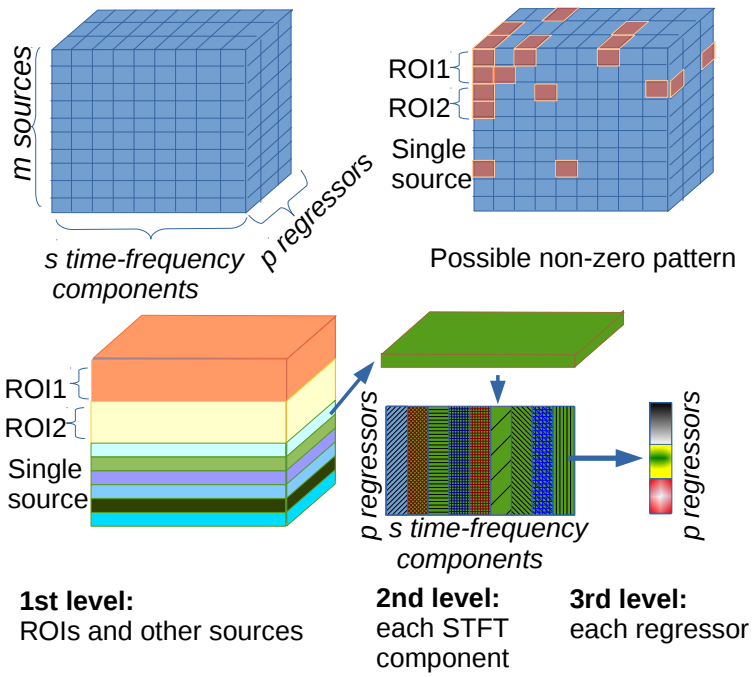


Figure 3: The hierarchical L_21 penalty

$$\Omega(\boldsymbol{Z}, \alpha, \beta, \gamma, \boldsymbol{w}) = \alpha \sum_l w_l \sqrt{\sum_{i \in \mathcal{A}_l} \sum_{j=1}^{s} \sum_{k=1}^{p} |Z_{ijk}|^2} \tag{4}$$

$$+ \beta \sum_{i=1}^{m} \sum_{j=1}^{s} \sqrt{\sum_{k=1}^{p} |Z_{ijk}|^2} \tag{5}$$

$$+ \gamma \sum_{i=1}^{m} \sum_{j=1}^{s} \sum_{k=1}^{p} |Z_{ijk}|. \tag{6}$$

The penalty Ω involves three terms corresponding to three levels of nested groups (Figure 3), and α, β and γ are tuning parameters. On the first level in (4), each group under the square root either consists of coefficients for all source points within one ROI, or coefficients for one single source point outside the ROIs. Therefore we have N_α groups, denoted by $\mathcal{A}_l, l = 1, \cdots, N_\alpha$, where N_α is the number of ROIs plus the number of source points outside the ROIs. Such a structure encourages the source signals outside the ROIs to be spatially sparse and thus reduces computational cost. With a good choice of weights for the N_α groups, $\boldsymbol{w} = (w_1, w_2, \ldots w_{N_\alpha})^T$, we can also make the penalty on coefficients for source points within the ROIs smaller than that on coefficients for source points outside the ROIs. On the second level, for each source point i, the term (5) groups the p regression coefficients for the jth STFT component under the square root, inducing sparsity over time points and frequencies. Finally, on the third level, (6) adds an L_1 penalty on each Z_{ijk} to encourage sparsity on the p covariates, for each STFT component of each source point.

The FISTA algorithm We use the fast iterative shrinkage-thresholding algorithm (FISTA [Beck and Teboulle, 2009]) to solve (3), with a constant step size, following [Gramfort et al., 2013]. Let

$\boldsymbol{z}$ be a vector that is concatenated by all entries in $\boldsymbol{Z}$, and let $\boldsymbol{y}$ be a vector of the same size. In each FISTA step, we need the proximal operator associated with the hierarchical penalty Ω:

$$\arg\min_{\boldsymbol{z}}(\frac{1}{2}\|\boldsymbol{z}-\boldsymbol{y}\|^2+\Omega(\boldsymbol{z},\alpha,\beta,\gamma,\boldsymbol{w}))=\arg\min_{\boldsymbol{z}}(\frac{1}{2}\|\boldsymbol{z}-\boldsymbol{y}\|^2+\sum_{h=1}^{N}\lambda_h\|\boldsymbol{z}|_{g_h}\|_2) \tag{7}$$

where we concatenate all of the nested groups on the three levels in Ω into an ordered list $\{g_1, g_2, \cdots, g_N\}$ and denote the penalty on group g_h by λ_h. For example, $\lambda_h = \alpha w_l$ if g_h is the lth group on the first level, $\lambda_h = \beta$ if g_h is on the second level, and $\lambda_h = \gamma$ if g_h is on in the third level. $\{g_1, g_2, \cdots, g_N\}$ is obtained by listing all the third level groups, then the second level and finally the first level, such that if h_1 is before h_2, then $g_{h_1} \subset g_{h_2}$ or $g_{h_1} \cap g_{h_2} = \emptyset$. Let $\boldsymbol{z}|_{g_h}$ be the elements of $\boldsymbol{z}$ in group g_h. As proved in [Jenatton et al., 2011], (7) is solved by composing the proximal operators for the L_{21} penalty on each g_h, following the order in the list; that is, initialize $\boldsymbol{z} \leftarrow \boldsymbol{y}$, for $h = 1, \cdots N$ in the ordered list,

$$\boldsymbol{z}|_{g_h} \leftarrow \begin{cases} \boldsymbol{z}|_{g_h}(1-\lambda_h/\|\boldsymbol{z}|_{g_h}\|_2) & \text{if } \|\boldsymbol{z}|_{g_h}\|_2 > \lambda_h \\ 0 & \text{otherwise} \end{cases}$$

Details of FISTA are shown in Algorithm 1, where $\boldsymbol{y}$ and $\boldsymbol{z_0}$ are auxiliary variables of the same

Algorithm 1: FISTA algorithm given the Lipschitz constant L

Data: $L, f(\boldsymbol{z}) = \frac{1}{2}\sum_{r=1}^{q}\|\boldsymbol{M}^{(r)} - \boldsymbol{G}\left(\sum_{k=1}^{p}X_k^{(r)}\boldsymbol{Z_k}\right)\boldsymbol{\Phi^H}\|_F^2, \Omega(\boldsymbol{z}) = \Omega(\boldsymbol{Z},\alpha,\beta,\gamma,\boldsymbol{w})$
Result: the optimal solution $\boldsymbol{z}$
initialization: $\boldsymbol{z_0}, \zeta = 1, \zeta_0 = 1, \boldsymbol{y} \leftarrow \boldsymbol{z_0}, \boldsymbol{z} \leftarrow \boldsymbol{z_0}$;
while *change of* $\boldsymbol{z}$ *in two iterations is not small enough* **do**
  $\boldsymbol{z_0} \leftarrow \boldsymbol{z}$; Compute $\nabla f(\boldsymbol{y})$;
  Apply the proximal operator $\boldsymbol{z} = \arg_{\boldsymbol{x}}\min(\frac{1}{2}\|\boldsymbol{x} - (\boldsymbol{y} - \frac{1}{L}\nabla f(\boldsymbol{y}))\|^2 + \frac{1}{L}\Omega(\boldsymbol{x}))$;
  $\zeta_0 \leftarrow \zeta$; $\zeta \leftarrow \frac{1+\sqrt{4\zeta_0^2+1}}{2}$; $\boldsymbol{y} \leftarrow \boldsymbol{z} + \frac{\zeta_0-1}{\zeta}(\boldsymbol{z} - \boldsymbol{z_0})$;
end

shape as $\boldsymbol{z}$, and ζ, ζ_0 are constants used to accelerate convergence. The gradient of $f(\boldsymbol{z})$ is computed in the following way: $\frac{\partial f}{\partial \boldsymbol{Z}_k} = -\boldsymbol{G}^T\sum_{r=1}^{q}X_k^{(r)}\boldsymbol{M}^{(r)}\boldsymbol{\Phi} + \boldsymbol{G}^T\boldsymbol{G}(\sum_{r=1}^{q}X_k^{(r)}\sum_{k'=1}^{p}\boldsymbol{Z}_{k'}X_{k'}(r))\boldsymbol{\Phi}^H\boldsymbol{\Phi}$. We use the power iteration method in [Gramfort et al., 2013] to compute the Lipschitz constant of the gradient.

The active-set strategy In practice, it is expensive to solve the original problem in (3). Thus we derive an active-set strategy (Algorithm 2), according to Chapter 6 in [Bach et al., 2011]: starting with a union of some groups on the first level ($J = \cup_{l\in\mathcal{B}}\mathcal{A}_l, \mathcal{B} \subset \{1, \cdots, N_\alpha\}$), we compute the solution to the problem constrained on J, then examine whether it is optimal for the original problem by checking whether the Karush-Kuhn-Tucker(KKT) conditions are met, if yes, we accept it, otherwise, we greedily add more groups to J and repeat the procedure.

Let $\boldsymbol{z}$ denote the concatenated $\boldsymbol{Z}$ again, and let diagonal matrix $\boldsymbol{D}_h$ be a filter to select the elements of $\boldsymbol{z}$ in group g_h (i.e. entries of $\boldsymbol{D}_h\boldsymbol{z}$ in group g_h are equal to $\boldsymbol{z}|_{g_h}$, and entries outside g_h are 0). Given a solution $\boldsymbol{z}_0$, the KKT conditions are

$$\nabla f(\boldsymbol{z})_{\boldsymbol{z}=\boldsymbol{z}_0} + \sum_h \boldsymbol{D}_h\boldsymbol{\xi}_h = 0, \text{and} \begin{cases} \boldsymbol{\xi}_h = \lambda_h\frac{\boldsymbol{D}_h\boldsymbol{z}_0}{\|\boldsymbol{D}_h\boldsymbol{z}_0\|_2} & \text{if } \|\boldsymbol{D}_h\boldsymbol{z}_0\|_2 > 0, \\ \|\boldsymbol{\xi}_h\|_2 \le \lambda_h & \text{if } \|\boldsymbol{D}_h\boldsymbol{z}_0\|_2 = 0 \end{cases}$$

where $\boldsymbol{\xi}_h, h = 1, \cdots, N$ are Lagrange multipliers of the same shape as $\boldsymbol{z}$. We defer the derivations to Appendix. We minimize the following problem

$$\min_{\boldsymbol{\xi}_h, \forall h} \frac{1}{2} \|\nabla f(\boldsymbol{z})_{\boldsymbol{z}=\boldsymbol{z}_0} + \sum_h \boldsymbol{D}_h \boldsymbol{\xi}_h\|_2^2,$$

$$\text{subject to } \begin{cases} \boldsymbol{\xi}_h = \lambda_h \frac{\boldsymbol{D}_h \boldsymbol{z}_0}{\|\boldsymbol{D}_h \boldsymbol{z}_0\|_2} & \text{if } \|\boldsymbol{D}_h \boldsymbol{z}_0\|_2 > 0, \\ \|\boldsymbol{\xi}_h\|_2 \leq \lambda_h & \text{if } \|\boldsymbol{D}_h \boldsymbol{z}_0\|_2 = 0 \end{cases}$$

and use $\frac{1}{2}\|\nabla f(\boldsymbol{z})_{\boldsymbol{z}=\boldsymbol{z}_0} + \sum_h \boldsymbol{D}_h \boldsymbol{\xi}_h\|_2^2$ at the optimum to measure the violation of KKT conditions. Additionally, we use $\frac{1}{2}\| \left(\nabla f(\boldsymbol{z})_{\boldsymbol{z}=\boldsymbol{z}_0} + \sum_h \boldsymbol{D}_h \boldsymbol{\xi}_h\right) |_{\mathcal{A}_l} \|_2^2$, constrained on each non-active first-level group $\mathcal{A}_l \not\subset J$, as a measurement of violation for the group.

Algorithm 2: Active-set strategy

initialization: choose initial J and initial solution Z; compute the KKT violation for each $\mathcal{A}_l \not\subset J$;
while *the total KKT violation is not small enough* **do**
  add 50 non-active groups that have the largest KKT violations to J;
  compute a solution to the problem constrained on J using FISTA ;
  compute the KKT violation for each $\mathcal{A}_l \not\subset J$;
end

***L_2* regularization and bootstrapping** The hierarchical L_{21} penalty may give biased results [Gramfort et al., 2013]. To reduce bias, we computed an L_2 solution constrained on the non-zero entries of the hierarchical L_{21} solution. Tuning parameters in the L_{21} and L_2 models were selected to minimize cross-validated prediction error.

To obtain the standard deviations of the regression coefficients in $\boldsymbol{Z}$, we performed a data-splitting bootstrapping procedure. The data was split to two halves (odd and even trials). On the first half, we obtained the hierarchical L_{21} solution, and on the second half, we computed an L_2 solution constrained on the non-zero entries of the hierarchical L_{21} solution. Then we plugged in this L_2 solution $\boldsymbol{Z}$ to obtain residual sensor time series of each trial on the second half of the data ($\boldsymbol{R}^{(r)} = \boldsymbol{M}^{(r)} - \boldsymbol{G}(\sum_{k=1}^p X_k^{(r)} \boldsymbol{Z}_k)\boldsymbol{\Phi}^{\boldsymbol{H}}$). We rescaled the residuals according to the design matrix $\boldsymbol{X}$ [Stine, 1985]. Let $X_r = \boldsymbol{X}[r,:]^T = (X_1^{(r)}, X_2^{(r)}, \cdots, X_p^{(r)})^T$, and $h_r = X_r^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} X_r$. The residual in the rth trial was rescaled by $1/(1-h_r)^{0.5}$. The re-sampled residuals $\boldsymbol{R}^{(r)*}$s were random samples with replacement from $\{\boldsymbol{R}^{(r)}/(1-h_r)^{0.5}, r = 1, \cdots, q\}$ and the bootstrapped sensor data for each trial were

$$\boldsymbol{M}^{(r)*} = \boldsymbol{G}(\sum_{k=1}^p X_k^{(r)} \boldsymbol{Z}_k)\boldsymbol{\Phi}^{\boldsymbol{H}} + \boldsymbol{R}^{(r)*}$$

After B re-sampling procedures, for each bootstrapped sample, we re-estimated the solution to the L_2 problem constrained on the non-zero entries again, and the best L_2 parameter was determined by a 2-fold cross-validation.

Applying STFT on the data The X here included the behavioral learning curve column and an all one-column to fit the intercept. We set the weights on α within the face-sensitive ROIs to zero, and the weights for all the other source points to an equal positive value that summed to 1. We did a grid search for α, β and γ, and selected the optimal parameters via two-fold cross-validation as mentioned above. We used T_0 =160 ms time windows, and τ_0 = 40 ms step for the STFT, which

resulted in 0-50 Hz spaced by 6.25 Hz, according to our sampling rate (100 Hz). B was 20 in the bootstrapping.

4.6 Statistical analysis

Regression analysis For regression of the sensor data in the time domain, we ran a separate regression against the behavioral learning curve for trials in each face category, at each time point. With only one regressor in this case, we mainly focused on the slope coefficient. We ran T-tests on the slope coefficients and obtained two-sided p-values against the null hypothesis that the coefficients were 0.

For regression in the face-sensitive ROIs in the source space (STFT-R and MNE-R), the original regression was essentially run on the STFT components of each source point. To visualize the coefficients ($\boldsymbol{Z}_k$) for the kth regressor in the time domain, we applied inverse STFT to transform $\boldsymbol{Z}_k$ to time series of coefficients ($\boldsymbol{Z}_k\boldsymbol{\Phi}^{\boldsymbol{H}}$) [3]. We used permutation tests to get p-values of whether the slope coefficients for the behavioral learning curve were significantly non-zero, by randomly permuting the rows of the regressors ($\boldsymbol{X}$). The permutation was only done in the second half of the split data, and for STFT-R, in each permutation, we solved L_2 solutions constrained on the non-zero entries of $\boldsymbol{Z}$ obtained from the first half of the data. It is also worth noting that in the source space, signs of the signals of a source point indicated the orientation of the electrical current. Due to the folding of sulci and gyri in the brain, two source points that were close in the euclidean space can have opposite orientations. Since the forward matrix only had limited spatial resolution in the euclidean space, the estimated source solutions may have an opposite sign to the truth. Therefore, to summarize the coefficients across the source points in an ROI, we took the the square of the coefficients averaged across source points, and compared them with the permuted counterparts.

Discriminant analysis To test whether multivariate neural signals from multiple sensors or source points were able to distinguish the two face categories, we ran Hotelling's two-sample T-squared tests on the multivariate signals at each time point, for the smoothed epochs. Let $\boldsymbol{y}_r \in \mathbb{R}^n$ be the response of n sensors or source points at a certain time point, in the rth trial. Let A and B denote the set of trials with face Category A and B, and q_A, q_B be the number of trials in each category. We computed the Hotelling's T-square in the following way: we took the sample mean for each category,

$$\bar{\boldsymbol{y}}_A = \frac{1}{q_A}\sum_{r\in A}\boldsymbol{y}_r \quad \bar{\boldsymbol{y}}_B = \frac{1}{q_B}\sum_{r\in B}\boldsymbol{y}_r$$

and then estimated a common covariance matrix for both categories,

$$\boldsymbol{W} = \frac{\sum_{r\in A}(\boldsymbol{y}_r - \bar{\boldsymbol{y}}_A)(\boldsymbol{y}_r - \bar{\boldsymbol{y}}_A)^T + \sum_{i\in B}(\boldsymbol{y}_r - \bar{\boldsymbol{y}}_B)(\boldsymbol{y}_r - \bar{\boldsymbol{y}}_B)^T}{q_A + q_B - 2}$$

[3] Since for the rth trial, $\boldsymbol{M}^{(r)} = \boldsymbol{G}\left(\sum_{k=1}^p X_k^{(r)}\boldsymbol{Z}_k\right)\boldsymbol{\Phi}^{\boldsymbol{H}} + \boldsymbol{E}^{(r)}$ and the entry in the regressor $X_k^{(r)}$ is a scalar, it is equivalent to $\boldsymbol{M}^{(r)} = \boldsymbol{G}\left(\sum_{k=1}^p X_k^{(r)}(\boldsymbol{Z}_k\boldsymbol{\Phi}^{\boldsymbol{H}})\right) + \boldsymbol{E}^{(r)}$, where $(\boldsymbol{Z}_k\boldsymbol{\Phi}^{\boldsymbol{H}})$ is the regression coefficients in the time domain.

Then the testing statistics, T-square was defined as

$$t^2 = \frac{(\bar{\boldsymbol{y}}_A - \bar{\boldsymbol{y}}_B)^T \boldsymbol{W}^{-1} (\bar{\boldsymbol{y}}_A - \bar{\boldsymbol{y}}_B)}{1/q_A + 1/q_B}$$

Under the null hypothesis that the means of two categories are the same, t^2 is related to an F-distribution

$$\frac{q_A + q_b - n - 1}{q_A + q_B - 2} t^2 \sim \mathcal{F}(n, q_A + q_B - 1 - n)$$

with which we obtained the p-values.

In addition, we might not have enough trials to estimate the covariance matrix for a large number of sensors or source points, therefore we implemented two different types of dimension reduction before applying Hotelling's T-squared tests. In the first approach, for source points within an ROI, which were often highly correlated, we did principle component analysis discarding the category labels, and then used the projections to the first several principle components preserving 99% of the variance in the Hotelling's T-squared tests. We labelled this approach as `PCA-Hotelling`. We found that the PCA-Hotelling procedure did not perform well on the 306-dimensional sensor data, possibly because the most discriminant sensors did not explained a big part of the variance. Therefore, for sensor data, we used a second approach, where we split the trials randomly into two parts, and on the first half trials, we ran a univariate two-sample T-tests on each sensor, and selected the top 20 sensors with the lowest p-values, and then applied Hotelling's T-squared tests only on these sensors in the second half trials. We labelled this approach as `split-Hotelling`. The split was done multiple times independently, and we averaged $-\log_{10}$ of the p-values across the splits.

Visualization and tests of statistics
(1) `Pivotal confidence intervals` After obtaining the statistics of tests for individual subjects, which were usually time series, we averaged them across subjects. To visualize the uncertainty of the average, we computed pivotal confidence intervals [Wasserman, 2010] by bootstrapping the time series from multiple subjects. In such a way, we preserved the correlation between adjacent time points, but it is worth noting that these confidence intervals were not corrected for multiple comparisons at different time points.

(2) `Permutation-excursion tests` When testing if in some time window(s), a time series was significantly different from the null hypothesis, we needed to correct for multiple comparisons at different time points. With permutation-excursion tests [Maris and Oostenveld, 2007, Xu et al., 2011], we were able to control the family-wise false discovery and obtain a global p-value for a window. In such tests, we looked for clusters of continues time points where the testing statistics were greater than a threshold, then summed the statistics within each cluster. We permuted the condition labels to obtain the statistics under the null hypothesis, and applied the same procedure to the permuted statistics. Finally the global p-value was defined as the proportion of permutations where the largest summed statistics were greater than the observed one. Particularly, in the regression and discriminant analysis in the sensor space, we tested whether the averaged $-\log_{10}(\texttt{p-value})$ time series across subjects were significantly greater than chance. The baseline window ($-140 \sim -40ms$) was de-meaned during the preprocessing and thus was a good representation of chance. We subtracted the averaged $-\log_{10}(\texttt{p-value})$ in the baseline window off individually for each subject, and tested if the group mean of this difference was not zero. In this case, the testing statistics were the

T-statistics across subjects at each time point, and each permutation was implemented by assigning a random sign to the time series of each subject. This test was implemented in `MNE-python`, and we termed it as `permutation-excursion T-test`. The number of permutations of this test was set to 1024, and the threshold of the T-statistics were equivalent to uncorrected $p \leq 0.05$.

(3) `Fisher's method` We also used Fisher's method to combine individual p-values $p_i, i = 1 \ldots, K$ of independent tests. This method tests against that the null hypothesis is true for every individual test. Under the null, $-2\sum_{i=1}^{K} \log p_i$ has a χ^2_{2K} distribution with $2K$ degrees of freedom, and we can obtain a combined p-value based on the χ^2_{2K} distribution. This test is sensitive when used to combine p-values for individual subjects, especially in our case where the number of subjects were moderate for many face-sensitive ROIs (See the last row of Table 1).

5 Results

5.1 Regression against the behavioral learning curve

In this section, we first show the results of the regression analysis against behavioral learning curves of the sensor time series, and then we show the results by STFT-R and MNE-R, in the source space and focused on the face-sensitive ROIs. The main goal was to examine in which ROIs at what time window during a trial the signals were correlated with increasing behavioral accuracy. Note that although the stimulus in each trial was a distinct exemplar, the sampling of the exemplar in the "eye-size/mouth-width" space was balanced as much as possible during the learning procedure. Therefore in our regression analysis, we were mainly looking at the change of neural response to the face category as a integrated group, and within-category difference between exemplars was not considered.

5.1.1 Regression of sensor data

Figure 4a shows the estimated the behavioral learning curves for both categories for one example subject `s1`. We used these curves as the regressors and regressed the data at each time point across trials for each sensor against them. Observing that for some subjects, the learning curves seemed different between the two face categories, we ran the regression for the two categories separately, for example, only trials in Category A were used in the regression against the learning curve for Category A. Only the first 250 trials for each category, (500 trials in total), where the learning curves were most steep, were used in our analysis. In Figure 4, we demonstrated the result from one example sensor near the right occipital side of the brain (Figure 4b) for Category A. The signal-to-noise ratio of single trials was usually low, therefore, to visually examine the trend from early trials to middle trials, we averaged the time series every 20 trials, to help clearly demonstrate the waveforms. In Figure 4c, we plotted the first six averaged time series (indexed by 0 to 5) covering trials 1-120, of Category A. There appeared to be some systematic change in 150-300 ms, as we move from the averages of the first 20 trials to the sixth 20 trials. We further plotted the intercepts (Figure 4d) and slopes (Figure 4e) for this sensor at each time point. The intercepts show the averaged time series of all 250 trials, and the slope time series was negative in 150-300 ms, which corresponded the changes in Figure 4c.

Next we tested whether there was a significant non-zero correlation with behavioral learning at a group level, by looking at the p-values of the regressions across all 9 subjects (Figure 5). To

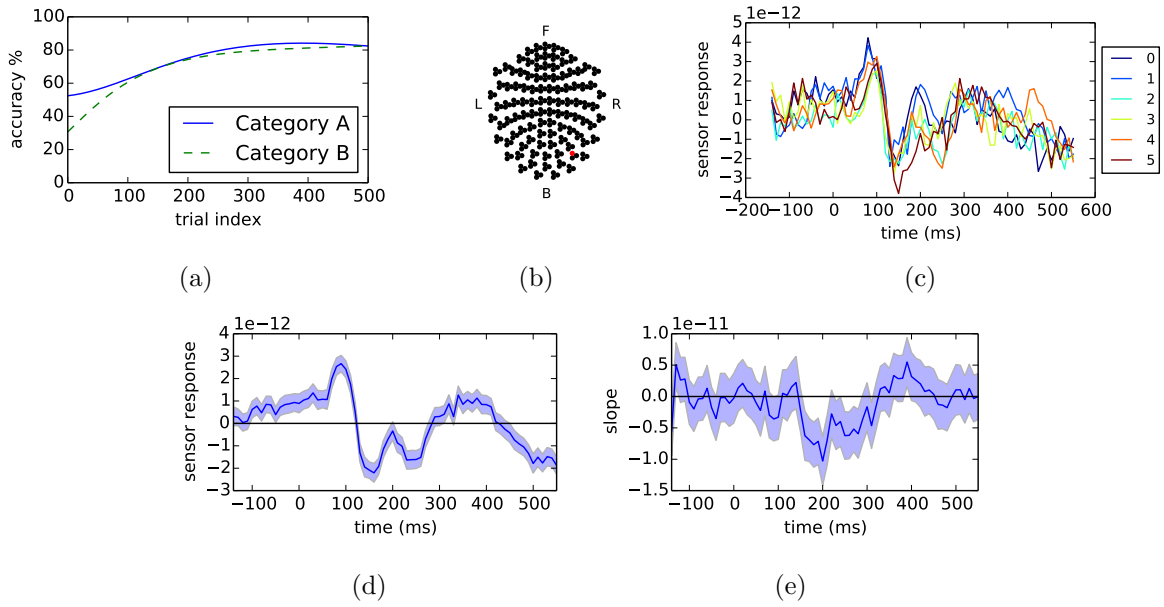
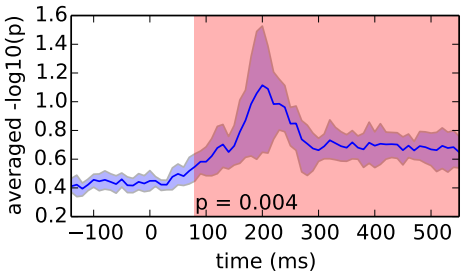


Figure 4: Regression of sensor data against the behavioral learning curves: one example subject (`s1`). (a) Behavioral learning curves of the example subject (`s1`), estimated with a logistic regression. (b-e) Demonstration of one example MEG sensor for subject `s1`. (b) The topological map of the sensors on a MEG helmet, viewed from top. Left(L), right(R), front(F) and back(B) of the head are labelled. The red dot is the example sensor. (c) Averaged responses of the example sensor in a moving and non-overlapping 20-trial window, for face Category A. Numbers in the legend show the indices of the trial windows, (0, the first window, 1 the second, etc.). (d-e) Estimated intercept time series (d) and slope time series (e) for the example sensor, for face Category A. The blue bands show 95% confidence intervals assuming asymptotic Gaussian distributions of the slopes and intercepts.

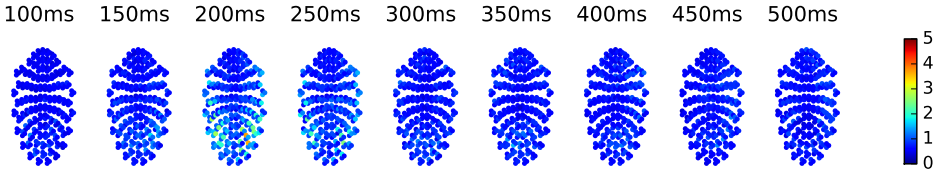
visualize the overall significance across sensors, we took $-\log_{10}$ of the regression p-values for each sensor and category, and then averaged them across categories and sensors for each subject, resulting in 9 time series of $-\log_{10}(\texttt{p-value})$. High $-\log_{10}(\texttt{p-value})$s indicate strong correlation with behavioral learning. Figure 5a shows the average of $-\log_{10}(\texttt{p-value})$s across subjects, with 95% pivotal confidence intervals. Based on an permutation-excursion T-test against the baseline, we found a significant time window ($p < 0.01$) in ∼90 ms to 560 ms, and the effect of correlation with behavioral learning was predominant at about 150 ms to about 300 ms. To visualize which sensors mainly contributed to the effect, we plotted the averaged $-\log_{10}(\texttt{p-value})$ across categories and subjects, on a topology map of sensors (Figure 5b), further averaged over 60 ms window flanking the labelled time points. Again, we observed high $-\log_{10}(\texttt{p-value})$ near 150 to about 300 ms, and this effect was larger at the posterior sensors and left and right temporal sensors (i.e. bottom and left/right lateral side of the topology maps), which were close to the visual cortices in the occipital and temporal lobes.

5.1.2 Regression in the face-sensitive ROIs

For the regression analysis in the face-sensitive ROIs in the source space, we applied our method STFT-R and the alternative MNE-R. To reduce computational cost, we only analyzed the first 250 trials for each category, and because of the data-split paradigm, the effective number of trials we analyzed in each category was 125. To visualize the coefficients in the time domain, we inversely transformed the regression coefficients of the STFT components back to the time domain,

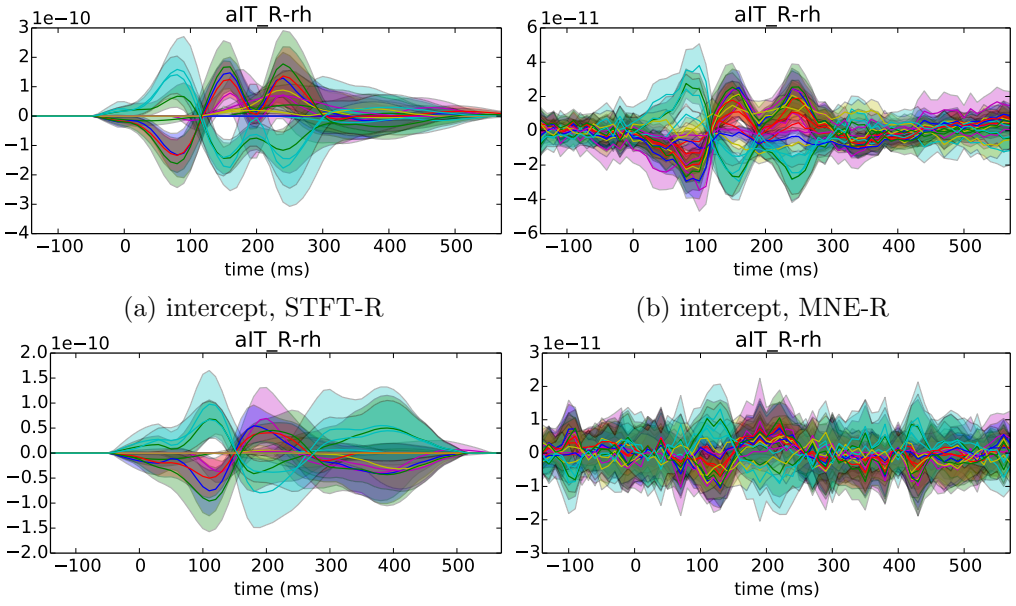


(a)



(b)

Figure 5: Regression of sensor data against the behavioral learning curves: results averaged across subjects. (a) Averaged $-\log_{10}$(p-value)s of all 306 sensors and two face categories across subjects. The shaded red area indicates a window where the $-\log_{10}$(p-value)s were significantly larger than the baseline ($-140 \sim -40$ ms), (right sided `permutation-excursion T-test`, 9 subjects). The blue shaded bands show 95% pivotal confidence intervals across subjects. (b) Averaged $-\log_{10}$(p-value)s across two categories and subjects, further averaged in 60 ms windows flanking the labelled time, on a sensor topology map.



(a) intercept, STFT-R

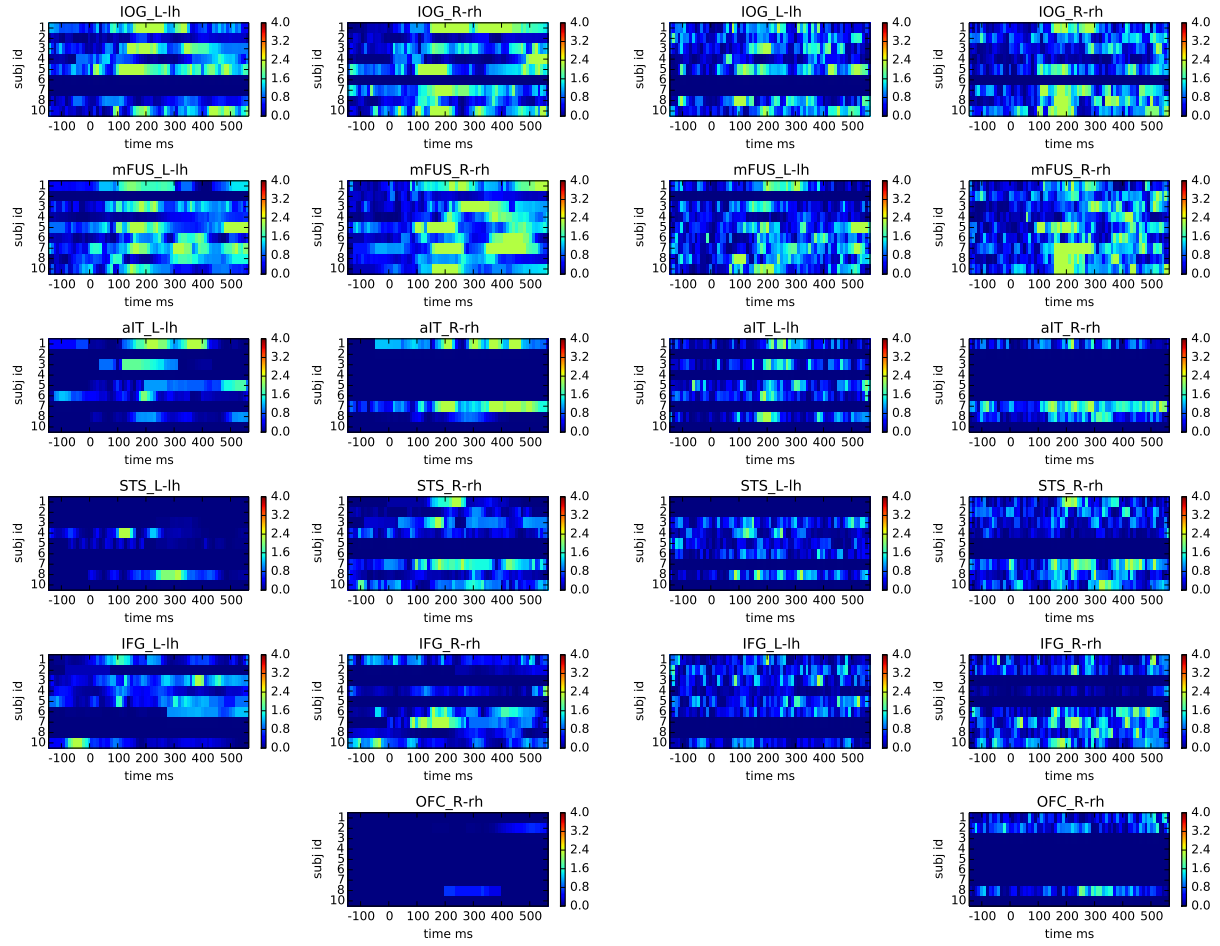
(b) intercept, MNE-R

(c) slope, STFT-R

(d) slope, MNE-R

Figure 6: Regression against the behavioral learning curve in the right aIT for Category A, sample subject `s1`. The intercept and slope time series are shown for STFT-R and MNE-R. Each curve with one color represents one source point, and the shaded bands show 95% confidence intervals, assuming asymptotic Gaussian distributions.

and called them `slope time series` and `intercept time series`. Figure 6 show the estimated intercept and slope time series of one example subject `s1` in the right aIT (aIT_R) for Category A. Each colored curve represents one source point in the ROI. STFT-R produced smoother time series than MNE-R, especially for slopes, because the penalty in STFT-R selected sparse time-frequency components that mostly explained the variance of the sensor data, which were often relatively lower frequency components (see Figure A.2 to visualize the slope coefficients in the original time-frequency domain).

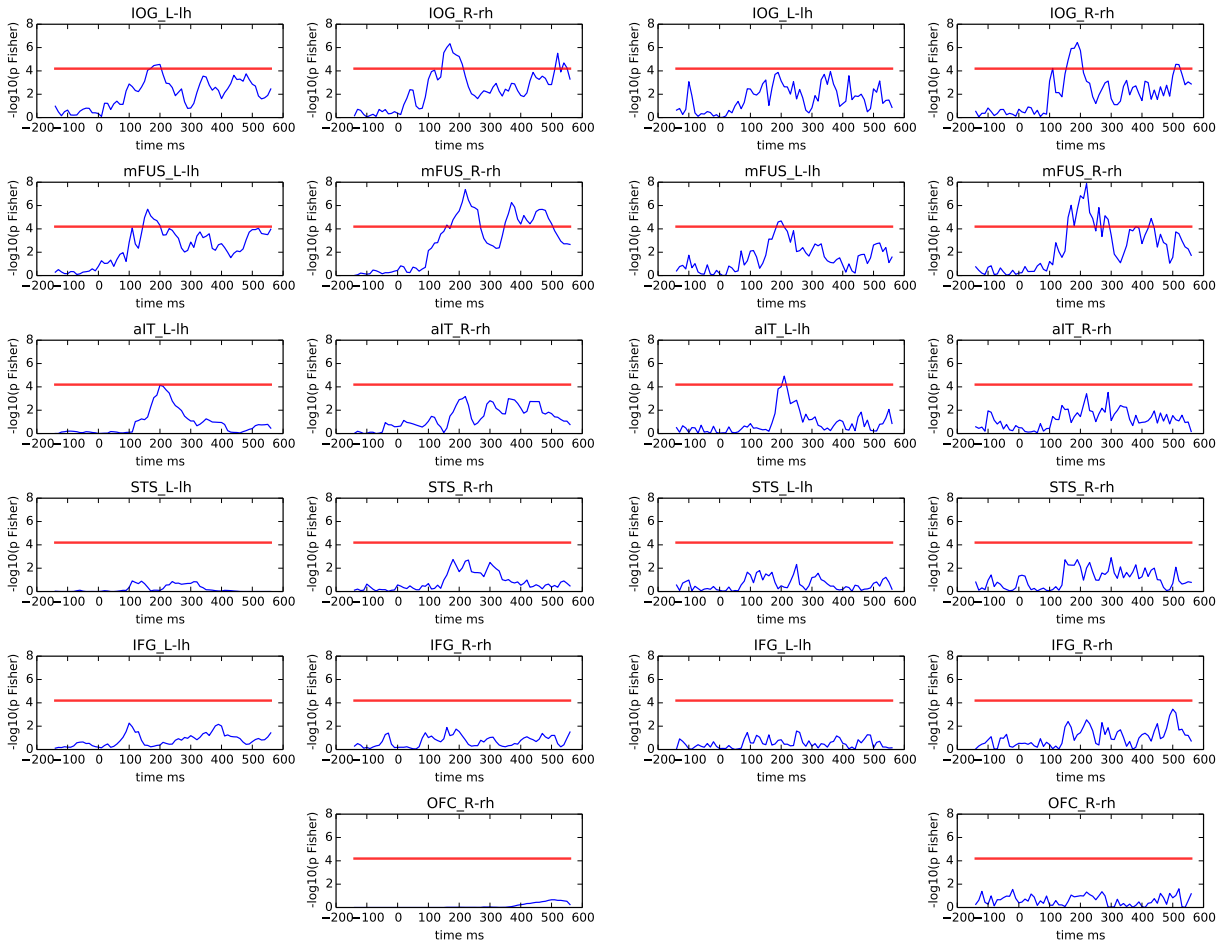


(a) STFT-R

(b) MNE-R

Figure 7: Regression of the source data in the ROIs: individual $-\log_{10}$(p-value)s of the permutation tests combined by Fisher's method across the two face categories from STFT-R(a) and MNE-R(b). Dark blue rows indicate that the ROI was absent for this subject.

We used permutation tests to examine whether the aggregated slope coefficients in each ROI were non-zero, by averaging the squares of slope coefficient time series across source points in each ROI, and comparing them with permuted counterparts (see Section 4.6). 40 permutations of the trial indices of the regressor (behavioral learning curves) were run for each subject and each face category, independently. We used Fisher's method to first combine the permutation p-values between the two categories in each ROI of each subject. Figure 7 shows $-\log_{10}$ of these combined p-values, which indicate whether for at least one category, the slope coefficients were above chance. Note



(a) STFT-R (b) MNE-R

Figure 8: Regression of the source data in the ROIs: $-\log_{10}$ of the combined p-values across subjects for each ROI, using Fisher's method, from STFT-R(a) and MNE-R(b). The red lines indicate the threshold of significance with Bonferroni correction at a level of 0.05.

that STFT-R produced smoother patterns of the individual $-\log$(p-value)s, which appeared to be easier to interpret, than the patterns by MNE-R.

Then we used Fisher's method to further combine these p-values across subjects for each ROI, to test the null hypothesis that the ROI signals in none of the subjects showed significant correlation with behavioral learning. The $-\log_{10}$ of the combined p-values are shown in Figure 8. The red lines indicate a significant threshold at level 0.05, with Bonferroni correction for multiple comparisons (71 times points and 11 ROIs). Both STFT-R and MNE-R produced similar patterns, where most of the ventral visual ROIs, (the bilateral IOGs and mFUS, and the left aIT), showed large correlation effects with behavioral learning near 150 to 250 ms, where as the non-ventral visual ROIs (STS and the two frontal regions, IFG and OFC) did not show as strong effects, although it is worth noticing that some ROIs were only present in a few subjects (e.g. right aIT, left STS and right OFC), which resulted in less power of the tests than other ROIs. Interestingly, some subjects also showed correlation effects after 300 ms (see in Figure 7, which is mostly predominant in the right mFUS.

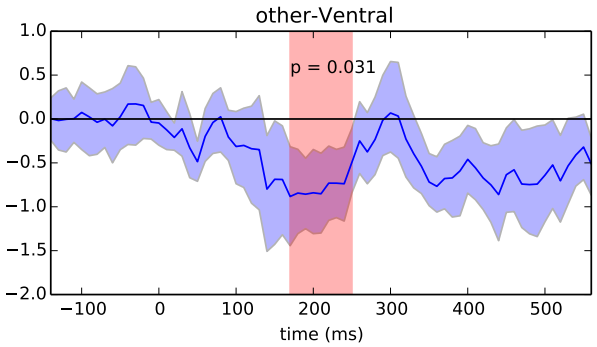


Figure 9: Mean difference of $-\log_{10}$(p-value) of the correlation with behavioral learning between the non-ventral ROIs and the ventral ROIs. The blue shaded bands shows 95% pivotal confidence intervals across subjects, and the red areas show significant time window from two-sided `permutatation-excursion T-test` on 9 subjects, with p-values labelled.

To further test whether the ventral ROIs showed larger correlation with behavioral learning than the non-ventral ROIs, we merged the ventral ROIs (bilateral IOG,mFUS and aIT) into one group, and the non-ventral ROIs (bilateral STS, IFG and the right OFC) into another group, then ran the similar permutation tests for each group as above, and directly examined the difference of $-\log$(p-values) between the two groups, with the STFT-R solutions. After obtaining the difference time series of $-\log$(p-values) of each subject, we used the `permutatation-excursion T-test` and found a significant time window near 170-250 ms (Figure 9), where the correlation effects indicated by $-\log$(p-values) were significantly smaller in the non-ventral ROIs than the ventral ROIs.

5.2 Discriminant analysis

In this subsection, we examined whether the MEG sensor data, and the data projected to the face-sensitive ROIs using dSPM could differentiate the two face categories, and more over, whether the discriminability changed from the early to the late stage of learning.

5.2.1 Sensor space discriminant analysis

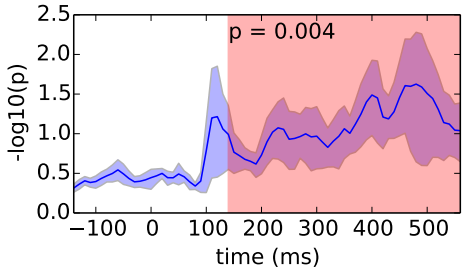


Figure 10: Discriminant analysis of the sensor data: averaged $-\log_{10}$(p-value)s across subjects in `split-Hotelling` tests over 728 trials. The shaded blue bands show 95% pivotal confidence intervals across 9 subjects. The red shaded area show a significant time window by the `permutation-excursion T-test`.

We ran the `split-Hotelling` test described in Section 4.6 on the 306-dimensional sensor data at each time point. Figure 10 shows the averaged $-\log_{10}$(p-value)s across subjects, using all 728 trials.

Again, we tested whether $-\log_{10}$(p-value)s were greater above chance by comparing with the baseline (-140∼-40 ms), using the `permutation-excursion T-test`. We found significant ($p < 0.01$) discriminablity starting from about 140 ms and lasting up to 560 ms, which indicated that the MEG signals carried information that distinguished the two categories after 150 ms of the stimulus onset.

5.2.2 Discriminant analysis in the face-sensitive ROIs

We further looked at the discriminability between face categories of the dSPM source solutions in the ROIs, by applying the `PCA-Hotelling` procedure for all 728 trials. We computed the averaged $-\log_{10}$(p-value)s of the `PCA-Hotelling` tests across subjects, for each face-sensitive ROI, and then used the `permutation-excursion tests` to search for significantly above-chance $-\log_{10}$(p-value)s. We randomly permuted the face-category labels of all trials for 500 times, and then compared the averaged $-\log_{10}$(p-value)s across subjects with the permuted counterparts. (The permutation sequences were the same across subjects). Figure 11 shows the results, where the green bands show 95% marginal intervals of the permutations, and the red shaded areas show significant time windows, at a level of 0.05 (Bonferroni correction for 11 ROIs). All the 11 ROIs except the right OFC showed significant discriminability. More interestingly, the non-ventral regions (IFGs and STSs) and the higher-level ventral regions (aITs) tended to show later discriminability after 300 ms. The lower- and mid-level ROIs in the ventral pathway (IOGs and mFUSs) showed additional earlier discriminability before 300 ms.

Finally, we directly compared the discriminability between the early stage of learning (the first 200 trials) and the late stage of learning (the last 200 trials), with two-sided excursion-permutation tests. We used the absolute value of the averaged difference of $-\log_{10}$(p-value)s in the late and early stages across subjects as our testing statistics for each ROI, and permuted the assignment of "early" and "late" of trials simultaneously for all subjects for 500 times. When looking at clusters where the statistics were above the threshold in the permuted cases, we took the largest sum from all ROIs, therefore, the resulting p-value was with regard to all ROIs, and no further correction for multiple comparison was needed. Note that the category labels were balanced in every 20 trials, therefore we had almost equal trials of Category A and Category B. Figure 12 show the difference of discriminability ($-\log_{10}$(p-value)s) between the late (the last 200 trials) and the early (the first 200 trials) stages of learning, averaged across subjects for each ROI. Green bands show marginal 95% intervals of the permutations. Although there appeared to be a decrease of determinability, in the left aIT near 200 to 220 ms ($p = 0.91$) and in the right aIT near 260 to 380 ms ($p = 0.07$), we did not find a significant time window at the level of 0.05.

6 Discussion

6.1 Discussion on the scientific findings

Regression analysis In this project, we discovered significant correlation between the trial-by-trial MEG data and subjects' behavioral learning curves. The correlation was mostly near 150 to about 300 ms, peaked near 200 ms, and was mostly focused in the MEG sensors that were close to the occipital and temporal lobes. We speculate that this time window can be related to both the M170/N170 and N250 components, at near 170 ms and 250 ms after the stimulus onset in

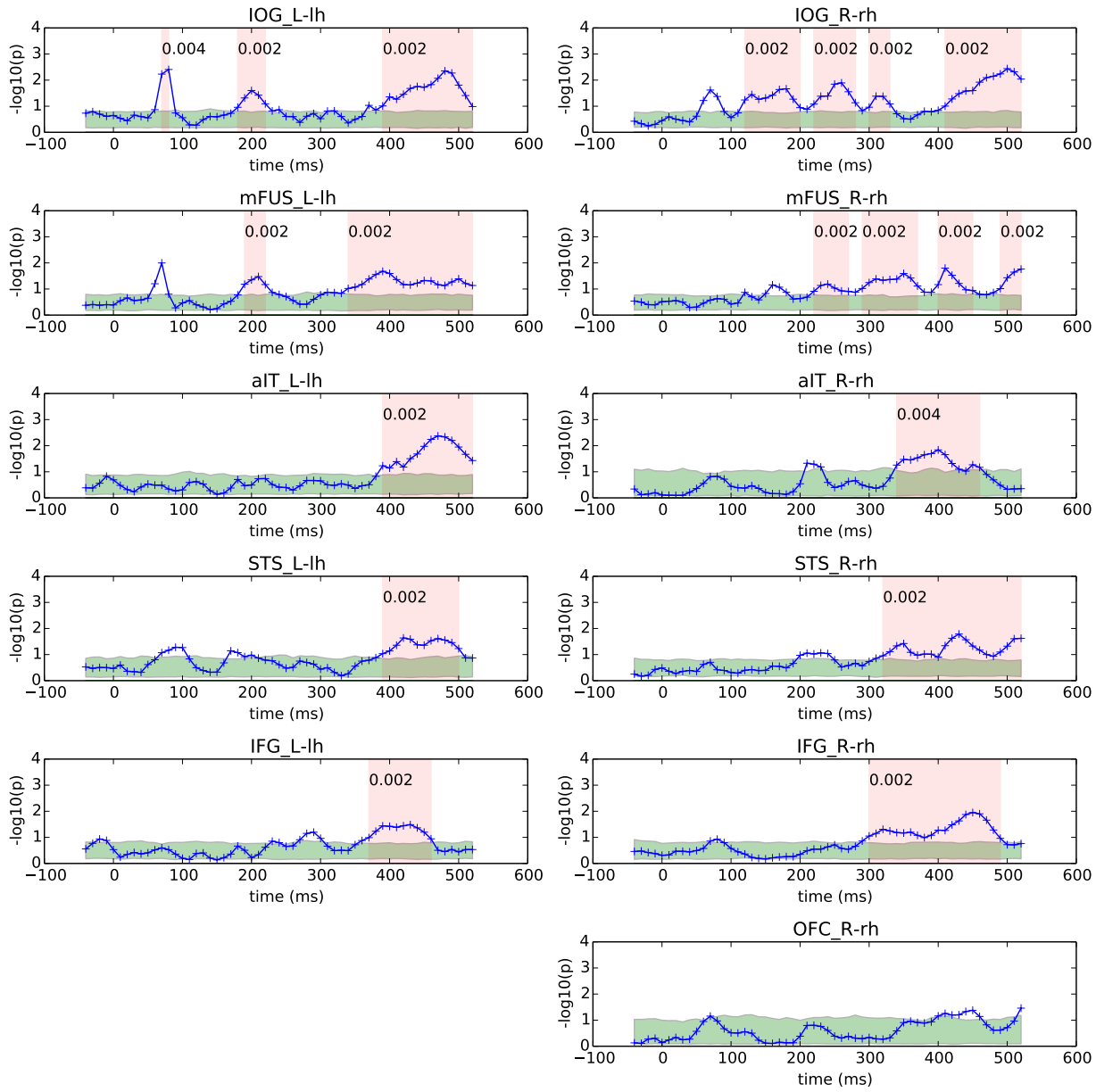


Figure 11: Averaged $-\log_{10}$(p-value)s across subjects for each face-sensitive ROI, from `PCA-Hotelling` tests, on 728 trials. Green bands show 95% marginal intervals of 500 permutations. The red shaded areas show significant windows where the averaged $-\log_{10}$(p-values) were significantly above chance, at a level of 0.05 (one-sided tests), with Bonferroni correction of multiple comparisons in 11 ROIs. Uncorrected p-values are labelled at each window. The threshold of the `excursion-permutation test` were 0.9 for all ROIs.

the MEG/EEG literature. Changes in the M/N170 components were reported in EEG face gender discrimination learning [Su et al., 2013], and short-term face memory tasks [Itier and Taylor, 2004]. The N250 was reported to be more negative when faces or objects are more familiar to the subject, and hence is hypothesized to index perceptual memory and familiarity [Schweinberger et al., 2004, Tanaka et al., 2006, Pierce et al., 2011]. In addition, [Itz et al., 2014] also found that enhancements of face features affected the N250 component and also a waveform near 200 ms. These findings are consistent with our results.

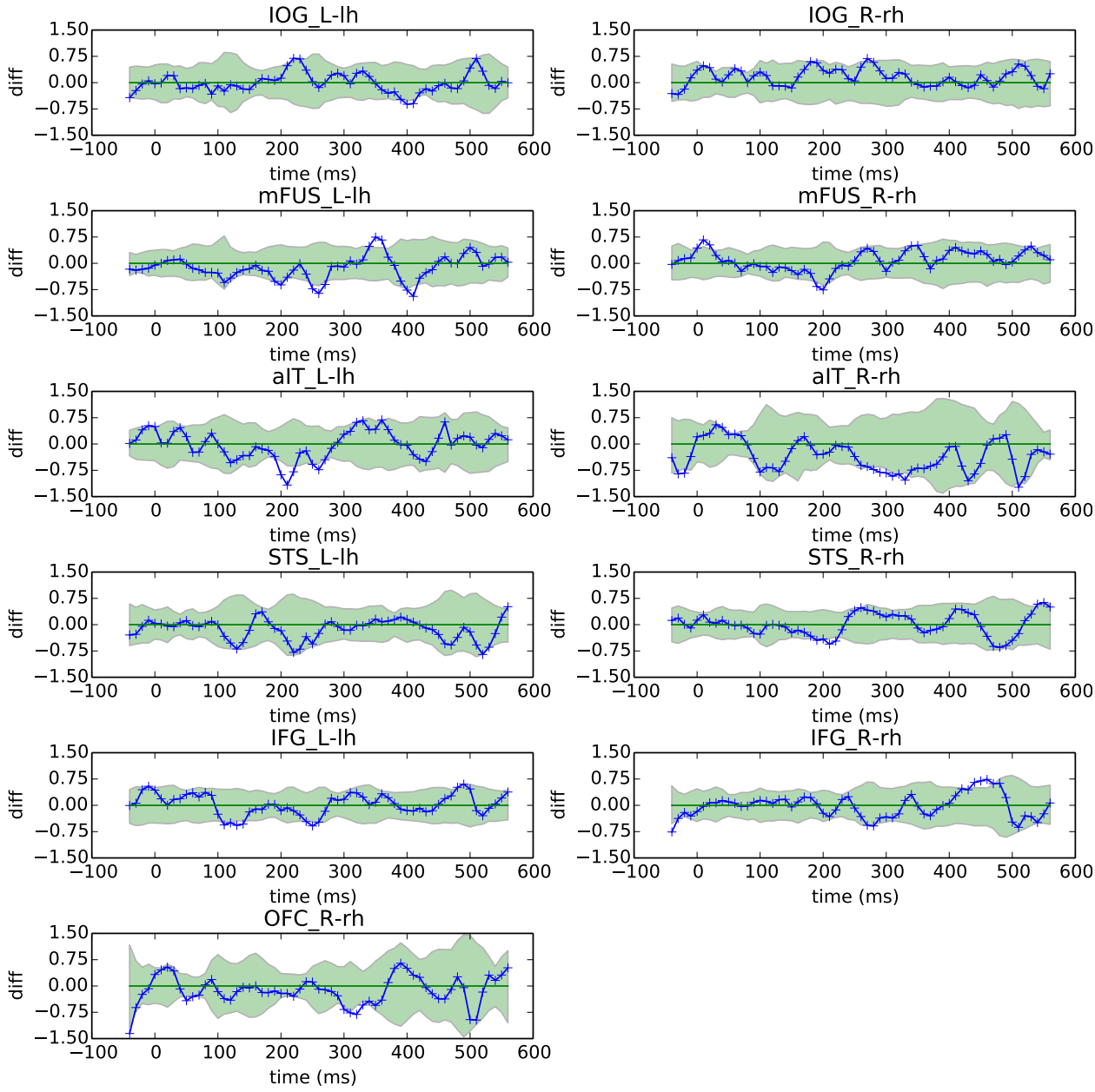


Figure 12: Difference of discriminability between late (the last 200 trials) and early stages (the first 200 trials) of learning averaged across subjects for each ROI (late - early). Green bands show marginal 95% intervals of 500 permutations. The threshold of `excursion-permutation test` were 0.5 for all ROIs (two-sided tests).

Yet, unlike many traditional EEG studies, we did not constrain our analysis only on the N170 or N250 time windows, because we aimed for an assumption-free exploration. One future analysis would be to describe the changes of the neural signals in a more parametric way, such as how the latency or amplitude changed, to better relate to the EEG literature.

Few previous MEG/EEG studies examined the neural correlates of face learning in the source space. Although the M170/N170 and N250 were roughly localized in the fusiform gyrus, [Deffke et al., 2007, Schweinberger et al., 2004], our work novelly described the correlation between the neural

signals in the face-sensitive ROIs and the behavioral learning curves. We found that the learning effect was predominant in the ventral visual ROIs (especially in the IOG and mFUS), whereas the non-ventral regions including the STS in the superior temporal lobe and the frontal regions (IFG and OFC) did not show as strong effects. This could reflect that the ventral visual ROIs and the non-ventral ROIs play different roles, and that the learning in this task was mainly related to visual processing. Interestingly, the STS was grouped into a "core" face-network, which also included the IOG and mFUS, based on their spatial proximity [Ishai, 2008]. However, our results suggest that STS may be functionally different in terms of face learning, which is consistent with the finding that structurally STS was not as much connected to other face areas[Pyles et al., 2013].

It is worth noting that head movement between runs might be a confound factor in our regression analysis, because the subject might move their heads when they gradually became tired of the task. The movement might also be a monotonic function towards the most comfortable position. In the sensor space analysis, we were not able to control this factor. But for source space analysis, we fit different forward solutions based on the head positions for each run to control the effect, and based on the results, it is unlikely that movement mainly contributed to the correlation with behavioral learning.

Discriminant analysis In the discriminant analysis with all 728 trials, we found that the MEG sensor data significantly discriminated the two face-categories, starting as early as about 140 ms. The window included the M170/N170 MEG/EEG component [Liu et al., 2000] which is hypothesized to index face identification, and also the 200 to 500 ms time window reported in direct electrode recordings in the fusiform gyrus [Ghuman et al., 2014]. With the dSPM source solutions, we found that the averaged discriminability across subjects was above chance in the bilateral ventral visual ROIs, STSs and IFGs, especially in the later time window after 300 ms. More interestingly, IOG, the lower-level ventral ROI, appeared to show earlier discriminability, than mFUS and aIT, and mFUS earlier than aIT too. This may be consistent with the hypothesis of IOG→mFUS→aIT organization of the ventrual visual ROIs, where information flows from the lower to higher level regions. However, more direct hypothesis testing of differences between ROIs (such as repeated measures ANOVA) is needed to make this claim, and with the moderate number of subjects, especially for some ROIs, we might not have enough power for such tests.

We did not find significant changes of discriminability between the early and late stages of learning. Although the right aIT had a marginally significant decrease near 260 to 380 ms, the fact that only three subjects had this region weakened the result. It may seem confusing when contrasting the null results here with the significant correlation of neural signals with behavioral learning, however, one possibility could be that the representations of the two face categories changed more or less similarly during learning, whereas the difference between the representations did not change largely enough to be detected, with the dSPM source localization and our current testing methods. In future analysis of this dataset, we may need to use more sensitive methods, and also explore other areas of the brain beyond the current ROIs.

6.2 Discussion on the methods

In this work, we proposed a novel one-step regression model (STFT-R) in the source localization framework, with short-time Fourier transform (STFT) representation of the source time series and sparsity-inducing penalty. Due to the sparsity in the STFT space, we obtained relatively smooth

intercept and slope time series of the regression. Therefore the results by STFT-R were easier to interpret than those by the alternative two-step method with the standard MNE source localization. In addition, because of the sparse non-zero set, it was also computationally faster to obtain bootstrapped solutions or permutation test results using STFT-R than using MNE-R.

As a first step, here we only regressed the MEG signals on one-dimensional behavioral learning curves. However, we can also apply higher dimensional regressors in the same STFT model. For example, we can regress against non-parametric basis functions to more flexibly model how neural signals changed with increasing trial number, so that we can investigate whether the changes lead or lagged the behavioral learning curves. Another future application of STFT-R is data-driven search of high-dimensional features that are important for human cognition. For example, for images, modern computer vision algorithms can produce a large number of visual features, and the sparsity-inducing property of STFT-R would be able to select the important features that explain the brain activities at various spatial and temporal sites.

Another observation worth noting is that the sparse STFT representation favored relatively low frequency signals, which usually explain the bulk part of the MEG data variance, and are often the main signals that MEG/EEG researchers are interested in. One future direction would be to use simpler basis functions than the STFT (e.g spline basis) that can also capture lower frequency waveforms, and possibly reduce the complexity of the model. On the other hand, another line of analyses focuses on higher frequency oscillations such as gamma band (30-100Hz), which is involved in face processing[Gao et al., 2013]. Unfortunately, the power at higher frequencies is often lower, and the phases might not align well across trials, therefore STFT-R is not suitable for picking up such signals. Instead, we can use two-step methods with MNE or dSPM, or develop suitable one-step models for this purpose.

In the discriminant analysis, we only compared the difference between the early and late stages of learning, but did not model the learning process as a continuous function, which might reduce our power if the actual changes of discriminability happened rapidly in the very beginning. An additional future direction is to build the learning curve as a continuous covariate into the discriminant analysis. Also, the dSPM source localizaiton method does not put priors on the regions of interest, nor does it emphasize the source points that have the highest discriminability, therefore another future direction is to build a one-step source space discriminant analysis model, like our one-step regression model, to further penalize the problem directly towards our statistical questions, and possibly improve the analysis results.

7 Conclusions

In this work, we analyzed the MEG recordings during a face-category learning experiment, and studied how the neural signals in face-sensitive ROIs changed with the behavioral learning curves. Firstly, we found the MEG sensor data had significant correlation with behavioral learning predominantly in 150-300 ms. In further regression analysis in the face-senstive ROIs, our STFT-R model produced smoother and thus more interpretable results than the alternative benchmark. In results by both methods, we observed that the correlation with behavioral learning was predominant in ventral face-sensitive regions, whereas regions in the temporal and frontal lobes did not show as strong effects. Our results revealed the spatio-temporal dynamics in the face-sensitive ROIs during

learning, on a finer-grained level than previous work, and also suggested the importance of visual processing in the ventral pathway in face learning.

Secondly with the discriminant analysis, we found that the neural signals recorded by MEG were able to differentiate the two face categories starting from 140 ms and lasting up to 560 ms. Such discriminability was localized in most of the face-sensitive regions, These results provided a spatio-temporal profile of face-category encoding in these regions. However, in comparisons between the early and late stages of learning in these ROIs, although there appeared to be a trend of discriminability decrease in some of the ventral regions, we did not find significant changes of discriminability that survived our permutation tests.

Acknowledgements

This work was funded by the Multi-Modal Neuroimaging Training Program (MNTP) fellowship from the NIH (5R90DA023420-08,5R90DA023420-09) and Richard King Mellon Foundation. We also thank Yang Xu and the MNE-python user group for their help.

Appendix

The Karush-Kuhn-Tucker conditions

Here we derive the Karush-Kuhn-Tucker (KKT) conditions for the hierarchical L_{21} problem. Since the term $f(\boldsymbol{z}) = \frac{1}{2}\sum_{r=1}^{q} ||\boldsymbol{M}^{(r)} - \boldsymbol{G}(\sum_{k=1}^{p} X_k^{(r)}\boldsymbol{Z}_k)\boldsymbol{\Phi}^{\boldsymbol{H}}||_F^2$ is essentially a sum of squared error of a linear problem, we can re-write it as $f(\boldsymbol{z}) = \frac{1}{2}||\boldsymbol{b} - \boldsymbol{A}\boldsymbol{z}||^2$, where $\boldsymbol{z}$ again is a vector concatenated by entries in $\boldsymbol{Z}$, $\boldsymbol{b}$ is a vector concatenated by $\boldsymbol{M}^{(1)}, \cdots, \boldsymbol{M}^{(q)}$, and $\boldsymbol{A}$ is a linear operator, such that $\boldsymbol{A}\boldsymbol{z}$ is the concatenated $\boldsymbol{G}(\sum_{k=1}^{p} X_k^{(r)}\boldsymbol{Z}_k)\boldsymbol{\Phi}^{\boldsymbol{H}}, r = 1, \cdots, q$. Note that although $\boldsymbol{z}$ is a complex vector, we can further reduce the problem into a real-valued problem by rearranging the real and imaginary parts of $\boldsymbol{z}$ and $\boldsymbol{A}$. Here for simplicity, we only derive the KKT conditions for the real case. Again we use $\{g_1, \cdots, g_h, \cdots, g_N\}$ to denote our ordered hierarchical group set, and λ_h to denote the corresponding penalty for group g_h. We also define diagonal matrices $\boldsymbol{D}_h$ such that

$$\boldsymbol{D}_h(l,l) = \begin{cases} 1 & \text{if } l \in g_h \\ 0 & \text{otherwise} \end{cases} \quad \forall h$$

therefore, the non-zero elements of $\boldsymbol{D}_h\boldsymbol{z}$ is equal to $\boldsymbol{z}|_{g_h}$. With the simplified notation, we re-cast the original problem into a standard formulation:

$$\min_{\boldsymbol{z}}(\frac{1}{2}\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{z}\|_2^2 + \sum_h \lambda_h \|\boldsymbol{D}_h\boldsymbol{z}\|_2) \tag{8}$$

To better describe the KKT conditions, we introduce some auxiliary variables, $\boldsymbol{u} = \boldsymbol{A}\boldsymbol{z}, \boldsymbol{v}_h = \boldsymbol{D}_h\boldsymbol{z}$. Then (8) is equivalent to

$$\min_{\boldsymbol{z},\boldsymbol{u},\boldsymbol{v}_h}(\frac{1}{2}\|\boldsymbol{b} - \boldsymbol{u}\|_2^2 + \sum_h \lambda_h \|\boldsymbol{v}_h\|_2)$$
$$\text{such that } \boldsymbol{u} = \boldsymbol{A}\boldsymbol{z}, \quad \boldsymbol{v}_h = \boldsymbol{D}_h\boldsymbol{z}, \forall h$$

The corresponding Lagrange function is

$$L(\boldsymbol{z}, \boldsymbol{u}, \boldsymbol{v}_h, \boldsymbol{\mu}, \boldsymbol{\xi}_h) = \frac{1}{2}\|\boldsymbol{b} - \boldsymbol{u}\|_2^2 + \sum_h \lambda_h \|\boldsymbol{v}_h\|_2 + \boldsymbol{\mu}^T(\boldsymbol{A}\boldsymbol{z} - \boldsymbol{u}) + \sum_h \boldsymbol{\xi}_h^T(\boldsymbol{D}_h\boldsymbol{z} - \boldsymbol{v}_h)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\xi}_h$'s are Lagrange multipliers. At the optimum, the following KKT conditions hold

$$\frac{\partial L}{\partial \boldsymbol{u}} = \boldsymbol{u} - \boldsymbol{b} - \boldsymbol{\mu} = 0 \tag{9}$$

$$\frac{\partial L}{\partial \boldsymbol{z}} = \boldsymbol{A}^T\boldsymbol{\mu} + \sum_h \boldsymbol{D}_h\boldsymbol{\xi}_h = 0 \tag{10}$$

$$\frac{\partial L}{\partial \boldsymbol{v}_h} = \lambda_h \partial\|\boldsymbol{v}_h\|_2 - \boldsymbol{\xi}_h \ni 0, \forall h \tag{11}$$

where $\partial\|\cdot\|_2$ is the subgradient of the L_2 norm. From (9) we have $\boldsymbol{\mu} = \boldsymbol{u} - \boldsymbol{b}$, then (10) becomes $\boldsymbol{A}^T(\boldsymbol{u} - \boldsymbol{b}) + \sum_h \boldsymbol{D}_h\boldsymbol{\xi}_h = 0$. Plugging $\boldsymbol{u} = \boldsymbol{A}\boldsymbol{z}$ in, we can see that the first term $\boldsymbol{A}^T(\boldsymbol{u} - \boldsymbol{b}) = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{z} - \boldsymbol{b})$ is the gradient of $f(\boldsymbol{z}) = \frac{1}{2}\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{z}\|_2^2$. For a solution $\boldsymbol{z}_0$, once we plug in $\boldsymbol{v}_h = \boldsymbol{D}_h\boldsymbol{z}_0$, the KKT conditions become

$$\nabla f(\boldsymbol{z})_{\boldsymbol{z}=\boldsymbol{z}_0} + \sum_h \boldsymbol{D}_h\boldsymbol{\xi}_h = 0 \tag{12}$$

$$\lambda_h \partial\|\boldsymbol{D}_h\boldsymbol{z}_0\|_2 - \boldsymbol{\xi}_h \ni 0, \forall h \tag{13}$$

In (13), we have the following according to the definition of subgradients

$$\boldsymbol{\xi}_h = \lambda_h \frac{\boldsymbol{D}_h\boldsymbol{z}_0}{\|\boldsymbol{D}_h\boldsymbol{z}_0\|_2} \text{ if } \|\boldsymbol{D}_h\boldsymbol{z}_0\|_2 > 0$$
$$\|\boldsymbol{\xi}_h\|_2 \leq \lambda_h \text{ if } \|\boldsymbol{D}_h\boldsymbol{z}_0\|_2 = 0$$

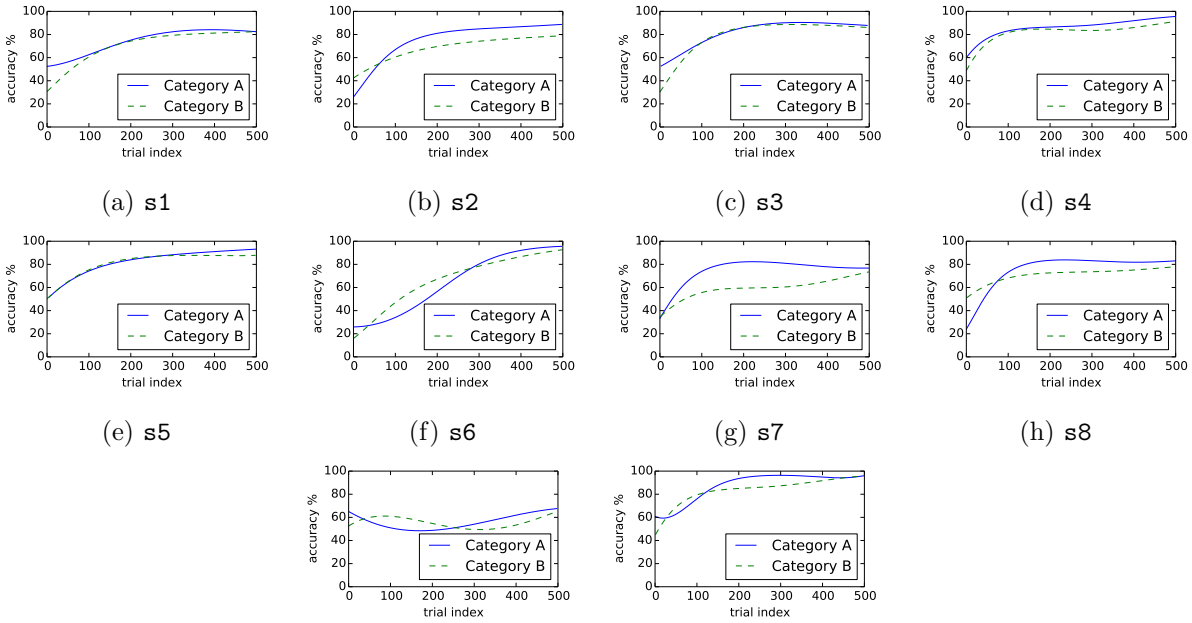
Therefore we can determine whether (12) and (13) hold by solving the following problem.

$$\min_{\boldsymbol{\xi}_h} \frac{1}{2}\|\nabla f(\boldsymbol{z})_{\boldsymbol{z}=\boldsymbol{z}_0} + \sum_h \boldsymbol{D}_h\boldsymbol{\xi}_h\|_2^2$$
$$\text{subject to } \boldsymbol{\xi}_h = \lambda_h \frac{\boldsymbol{D}_h\boldsymbol{z}_0}{\|\boldsymbol{D}_h\boldsymbol{z}_0\|_2} \text{ if } \|\boldsymbol{D}_h\boldsymbol{z}_0\|_2 > 0$$
$$\|\boldsymbol{\xi}_h\|_2 \leq \lambda_h \text{ if } \|\boldsymbol{D}_h\boldsymbol{z}_0\|_2 = 0$$

which is a standard group lasso problem with no overlap. We can use coordinate-descent to solve it. We define $\frac{1}{2}\|\nabla f(\boldsymbol{z})_{\boldsymbol{z}=\boldsymbol{z}_0} + \sum_h \boldsymbol{D}_h\boldsymbol{\xi}_h\|_2^2$ at the optimum as a measure of violation of the KKT conditions.

Let f_J be the function f constrained on a set J. Because the gradient of f is linear, if $\boldsymbol{z_0}$ only has non-zero entries in J, then the entries of $\nabla f(\boldsymbol{z})$ in J are equal to $\nabla f_J(\boldsymbol{z}|_J)$ at $\boldsymbol{z} = \boldsymbol{z}_0$. In addition, $\boldsymbol{\xi}_h$'s are separate for each group. Therefore if $\boldsymbol{z}_0$ is an optimal solution to the problem constrained on J, then the KKT conditions are already met for entries in J (i.e. $(\nabla f(\boldsymbol{z})_{\boldsymbol{z}=\boldsymbol{z}_0} + \sum_h \boldsymbol{D}_h\boldsymbol{\xi}_h)\,|_J = 0$); for $g_h \not\subset J$, we use $(\ \frac{1}{2}\|\,(\nabla f(\boldsymbol{z})_{\boldsymbol{z}=\boldsymbol{z}_0} + \sum_h \boldsymbol{D}_h\boldsymbol{\xi}_h)\,|_{g_h}\|^2)$ at the optimum as a measurement of how much the elements in group g_h violate the KKT conditions, which is a criterion when we greedily add groups (see Algorithm 2).

Additional figures

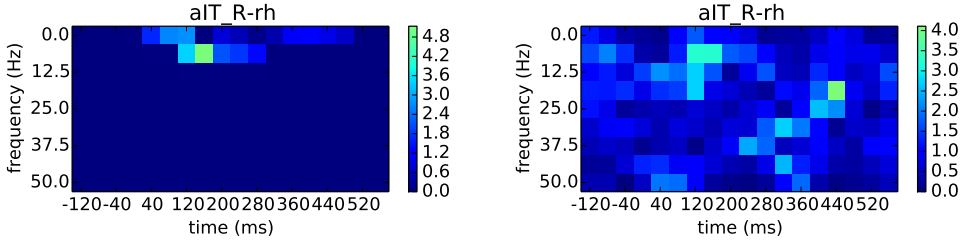


(a) s1 (b) s2 (c) s3 (d) s4

(e) s5 (f) s6 (g) s7 (h) s8

(i) s9 (j) s10

Figure A.1: Behavioral learning curves for individual participants. Subject s9 had almost flat learning curves, and was excluded for the analysis.



(a) STFT-R (b) MNE-R

Figure A.2: Sum of squared regression coefficients of STFT components within an ROI, divided by bootstrapped standard errors, by STFT-R (a) and MNE-R (b), for the example subject s1, Category A.

References

[Babadi et al., 2014] Babadi, B., Obregon-Henao, G., Lamus, C., Hamalainen, M. S., Brown, E. N., and Purdon, P. L. (2014). A subspace pursuit-based iterative greedy hierarchical solution to the neuromagnetic inverse problem. *NeuroImage*, 87(0):427 – 443.

[Bach et al., 2011] Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. (2011). Optimization with sparsity-inducing penalties. *CoRR*, abs/1108.0775.

[Barrett and Rugg, 1990] Barrett, S. E. and Rugg, M. D. (1990). Event-related potentials and the semantic matching of pictures. *Brain and Cognition*, 14(2):201 – 212.

[Beck and Teboulle, 2009] Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202.

[Dale et al., 2000] Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., and Halgren, E. (2000). Dynamic statistical parametric mapping: combining fmri and meg for high-resolution imaging of cortical activity. *Neuron*, 26(1):55–67.

[Deffke et al., 2007] Deffke, I., Sander, T., Heidenreich, J., Sommer, W., Curio, G., Trahms, L., and Lueschow, A. (2007). Meg/eeg sources of the 170-ms response to faces are co-localized in the fusiform gyrus. *NeuroImage*, 35(4):1495 – 1501.

[DeGutis and D'Esposito, 2007] DeGutis, J. and D'Esposito, M. (2007). Distinct mechanisms in visual category learning. *Cognitive, Affective, & Behavioral Neuroscience*, 7(3):251–259.

[DeGutis and D'Esposito, 2009] DeGutis, J. and D'Esposito, M. (2009). Network changes in the transition from initial learning to well-practiced visual categorization. *Frontiers in human neuroscience*, 3.

[DiCarlo and Cox, 2007] DiCarlo, J. J. and Cox, D. D. (2007). Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341.

[Galka et al., 2004] Galka, A., Ozaki, O. Y. T., Biscay, R., and Valdes-Sosa, P. (2004). A solution to the dynamical inverse problem of eeg generation using spatiotemporal kalman filtering. *NeuroImage*, 23:435–453.

[Gao et al., 2013] Gao, Z., Goldstein, A., Harpaz, Y., Hansel, M., Zion-Golumbic, E., and Bentin, S. (2013). A magnetoencephalographic study of face processing: M170, gamma-band oscillations and source localization. *Human brain mapping*, 34(8):1783–1795.

[Ghuman et al., 2014] Ghuman, A. S., Brunet, N. M., Li, Y., Konecky, R. O., Pyles, J. A., Walls, S. A., Destefino, V., Wang, W., and Richardson, R. M. (2014). Dynamic encoding of face information in the human fusiform gyrus. *Nature communications*, 5.

[Gramfort et al., 2012] Gramfort, A., Kowalski, M., and Hamaleinen, M. (2012). Mixed-norm estimates for the m/eeg inverse problem using accelerated gradient methods. *Physics in Medicine and Biology*, 57:1937–1961.

[Gramfort et al., 2014] Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., and Hmlinen, M. S. (2014). Mne software for processing meg and eeg data. *NeuroImage*, 86(0):446 – 460.

[Gramfort et al., 2013] Gramfort, A., Strohmeier, D., Haueisen, J., Hamalainen, M., and Kowalski, M. (2013). Time-frequency mixed-norm estimates: Sparse m/eeg imaging with non-stationary source activations. *NeuroImage*, 70(0):410 – 422.

[Hamalainen et al., 1993] Hamalainen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J., and Lounasmaa, O. V. (1993). Magnetoencephalography–theory, instrumentation, to noninvasive studies of the working human brain. *Reviews of Modern Physics*, 65:414–487.

[Hamalainen and Ilmoniemi, 1994] Hamalainen, M. and Ilmoniemi, R. (1994). Interpreting magnetic fields of the brain: minimum norm estimates. *Med. Biol. Eng. Comput.*, 32:35–42.

[Henson et al., 2011] Henson, R. N., Wakeman, D. G., Litvak, V., and Friston, K. J. (2011). A parametric empirical bayesian framework for the eeg/meg inverse problem: generative models for multi-subject and multi-modal integration. *Frontiers in human neuroscience*, 5.

[Ishai, 2008] Ishai, A. (2008). Lets face it: Its a cortical network. *NeuroImage*, 40(2):415 – 419.

[Itier and Taylor, 2004] Itier, R. J. and Taylor, M. J. (2004). Effects of repetition learning on upright, inverted and contrast-reversed face processing using erps. *Neuroimage*, 21(4):1518–1532.

[Itz et al., 2014] Itz, M. L., Schweinberger, S. R., Schulz, C., and Kaufmann, J. M. (2014). Neural correlates of facilitations in face learning by selective caricaturing of facial shape or reflectance. *NeuroImage*, 102:736–747.

[Jenatton et al., 2011] Jenatton, R., Mairal, J., Obozinski, G., and Bach, F. (2011). Proximal methods for hierarchical space coding. *J. Mach. Learn. Res*, 12:2297–2334.

[Kanwisher et al., 1997] Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, 17(11):4302–4311.

[Kriegeskorte et al., 2007] Kriegeskorte, N., Formisano, E., Sorger, B., and Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences*, 104(51):20600–20605.

[Lamus et al., 2012] Lamus, C., Hamalainen, M. S., Temereanca, S., Brown, E. N., and Purdon, P. L. (2012). A spatiotemporal dynamic distributed solution to the meg inverse problem. *NeuroImage*, 63:894–909.

[Liu et al., 2000] Liu, J., Higuchi, M., Marantz, A., and Kanwisher, N. (2000). The selectivity of the occipitotemporal m170 for faces. *Neuroreport*, 11(02):337–341.

[Maris and Oostenveld, 2007] Maris, E. and Oostenveld, R. (2007). Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, 164(1):177–190.

[Mattout et al., 2006] Mattout, J., Phillips, C., Penny, W. D., Rugg, M. D., and Friston, K. J. (2006). Meg source localization under multiple constraints: an extended bayesian framework. *NeuroImage*, 30(3):753–767.

[Mishkin et al., 1983] Mishkin, M., Ungerleider, L. G., and Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends in neurosciences*, 6:414–417.

[Nestor et al., 2008] Nestor, A., Vettel, J. M., and Tarr, M. J. (2008). Task-specific codes for face recognition: how they shape the neural representation of features for detection and individuation. *PloS one*, 3(12):e3978.

[Pascual-Marqui, 2002] Pascual-Marqui, R. (2002). Standardized low resolution brain electromagnetic tomography (sloreta): technical details. *Methods Find. Exp. Clin. Pharmacol.*, 24:5–12.

[Pierce et al., 2011] Pierce, L. J., Scott, L. S., Boddington, S., Droucker, D., Curran, T., and Tanaka, J. W. (2011). The n250 brain potential to personally familiar and newly learned faces and objects. *Frontiers in human neuroscience*, 5.

[Pitcher et al., 2011] Pitcher, D., Walsh, V., and Duchaine, B. (2011). The role of the occipital face area in the cortical face perception network. *Experimental Brain Research*, 209(4):481–493.

[Pyles et al., 2013] Pyles, J. A., Verstynen, T. D., Schneider, W., and Tarr, M. J. (2013). Explicating the face perception network with white matter connectivity. *PloS one*, 8(4):e61611.

[Rajimehr et al., 2009] Rajimehr, R., Young, J. C., and Tootell, R. B. (2009). An anterior temporal face patch in human cortex, predicted by macaque maps. *Proceedings of the National Academy of Sciences*, 106(6):1995–2000.

[Scherg and Von Cramon, 1985] Scherg, M. and Von Cramon, D. (1985). Two bilateral sources of the late aep as identified by a spatio-temporal dipole model. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 62(1):32–44.

[Schweinberger et al., 2004] Schweinberger, S. R., Huddy, V., and Burton, A. M. (2004). N250r: a face-selective brain response to stimulus repetitions. *Neuroreport*, 15(9):1501–1505.

[Stine, 1985] Stine, R. A. (1985). Bootstrap prediction intervals for regression. *Journal of the American Statistical Association*, 80:1026–1031.

[Su et al., 2013] Su, J., Tan, Q., and Fang, F. (2013). Neural correlates of face gender discrimination learning. *Experimental brain research*, 225(4):569–578.

[Tanaka et al., 2006] Tanaka, J. W., Curran, T., Porterfield, A. L., and Collins, D. (2006). Activation of preexisting and acquired face representations: the n250 event-related potential as an index of face familiarity. *Journal of Cognitive Neuroscience*, 18(9):1488–1497.

[Taulu and Simola, 2006] Taulu, S. and Simola, J. (2006). Spatiotemporal signal space separation method for rejecting nearby interference in meg measurements. *Physics in medicine and biology*, 51(7):1759.

[Wasserman, 2010] Wasserman, L. (2010). *All of Statistics: A Concise Course in Statistical Inference*. Springer Publishing Company, Incorporated.

[Xu, 2013] Xu, Y. (2013). Cortical spatiotemporal plasticity in visual category learning (doctoral dissertation).

[Xu et al., 2011] Xu, Y., Sudre, G. P., Wang, W., Weber, D. J., and Kass, R. E. (2011). Characterizing global statistical significance of spatiotemporal hot spots in magnetoencephalography/electroencephalography source space via excursion algorithms. *Statistics in medicine*, 30(23):2854–2866.