SINGLE-TRIAL DECODING OF SCALP EEG UNDER NATURALISTIC STIMULI

Greta Tuckute, Sofie Therese Hansen, Nicolai Pedersen, Dea Steenstrup, Lars Kai Hansen

Department of Applied Mathematics and Computer Science (DTU Compute)
Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark
{*grtu, lkai*}*@dtu.dk*

ABSTRACT

There is significant current interest in decoding mental states from electro-encephalography (EEG) recordings. We demonstrate inter-subject single-trial decoding of naturalistic stimuli based on scalp EEG acquired with user-friendly, portable, 32 dry electrode equipment. We show that Support Vector Machine (SVM) classifiers trained on a relatively small set of de-noised (averaged) trials perform at par with classifiers trained on a large set of noisy samples. We propose a novel method for computing sensitivity maps of EEG-based SVM classifiers for visualization of the EEG signatures exploited by SVM classifiers. We apply the NPAIRS framework for estimation of map uncertainty. We show that effect sizes of sensitivity maps are similar for classifiers trained on small samples of de-noised data and large samples of noisy data. We conclude that the average category classifier can successfully predict on single trial subjects, allowing for fast classifier training, parameter optimization and unbiased performance evaluation.

1. INTRODUCTION

Decoding of brain activity aims to reconstruct the perceptual and semantic content of neural processing based on activity measured in one or more brain imaging techniques, such as electro-encephalography (EEG), magneto-encephalography (MEG) and functional magnetic resonance imaging (fMRI). EEG-based decoding of human brain activity has significant potential due to excellent time resolution and a possibility of real-life acquisition, however, the signal is extremely diverse, subject-specific, sensitive to disturbances, and has a low signal to noise ratio, hence, posing a major challenge for both signal processing and machine learning [Nicolas-Alonso and Gomez-Gil, 2012]. Decoding studies based on fMRI have matured significantly during the last 15 years, see e.g. [Gerlach, 2007, Loula et al., 2018]. These studies successfully decode human brain activity from naturalistic image and movie stimuli [Kay et al., 2008, Prenger et al., 2009, Nishimoto et al., 2011, Huth et al., 2012, Huth et al., 2016, Güçlü and van Gerven, 2017].

In case of decoding of scalp EEG, the research area is still progressing, and relatively few studies document detection of brain states in regards to semantic categories (often discrimination between two high-level categories) [Simanova et al., 2010, Murphy et al., 2011, Taghizadeh-Sarabi et al., 2014, Kaneshiro et al., 2015, Zafar et al., 2017].

Due to before-mentioned challenges, previous studies have been performed in laboratory settings with high-grade EEG acquisition equipment [Simanova et al., 2010, Murphy et al., 2011, Stewart et al., 2014, Kaneshiro et al., 2015, Zafar et al., 2017]. Visual stimuli conditions can often not be described as naturalistic [Simanova et al., 2010, Murphy et al., 2011, Taghizadeh-Sarabi et al., 2014, Stewart et al., 2014, Kaneshiro et al., 2015] and individual trials are repeated multiple times [Simanova et al., 2010, Murphy et al., 2011, Stewart et al., 2014, Kaneshiro et al., 2015, Zafar et al., 2017] in order to average the event related potential (ERP) response for further classification purposes. Moreover, a number of participants are typically excluded from analysis due to artifacts and low classification accuracy [Taghizadeh-Sarabi et al., 2014, Zafar et al., 2017].

The motivation for the current study is to overcome the highlighted limitations in EEG-based decoding. Therefore, we acquired EEG data in a typical office setting using a portable, user-friendly, wireless EEG Enobio system with 32 dry electrodes. Stimuli consisted of naturalistic images from the Microsoft Common Objects in Context (MS COCO) image database [Lin et al., 2014], displaying complex everyday scenes and non-iconic views of objects from 23 different semantic categories. All images presented were unique and not repeated for the same subject throughout the experiment. We acquired data from 15 healthy participants (5 female). We are interested in exploring the limitations of inter-subject generalization, i.e., population models, hence no participants are excluded from analysis. Decoding ability is evaluated in a inter-subject design, i.e., in a leave-one-subject-out approach (as opposed to within-subject classification) to probe generalizability across participants.

The work in the current study is focused on the binary classification problem between two classes: brain processing of animate and inanimate image stimuli. Kernel meth-

ods, e.g., support vector machines (SVM) are frequently applied for learning of statistical relations between patterns of brain activation and experimental conditions. In classification of EEG data, SVMs have shown good performance in many contexts [Lotte et al., 2007, Murphy et al., 2011, Taghizadeh-Sarabi et al., 2014, Stewart et al., 2014, Andersen et al., 2017]. SVMs allow adoption of a nonlinear kernel function to transform input data into a high dimensional feature space, where it is possible to linearly separate data. The iterative learning process of the SVM will devise an optimal hyperplane with the maximal margin between each class in the high dimensional feature space. Thus, the maximum-margin hyperplane will form the decision boundary for distinguishing the brain response associated with animate and inanimate data [Saitta, 1995].

We adopt a novel methodological approach for computing and evaluating SVM classifiers based on two approaches: 1) Single-trial training and single-trial test classification, and 2) Training on a meaned response of each of the 23 image categories for each subject (corresponding to 23 trials per subject) and single-trial test classification. Furthermore, we open the black box and visualize which parts of the EEG signature are exploited by the SVM classifiers. In particular, we propose a method for computing sensitivity maps of EEG-based SVM classifiers based on a methodology originally proposed for fMRI [Rasmussen et al., 2011]. To evaluate effect sizes of ERP difference maps and sensitivity maps, we use a modified version of an NPAIRS resampling scheme [Strother et al., 2002]. Lastly, we investigate how the classifier based on the average EEG category response compares to the single-trial classifier in terms of prediction accuracy of novel subjects.

2. MATERIALS AND METHODS

2.1. Participants

A total of 15 healthy subjects with normal or corrected-to-normal vision (10 male, 5 female, mean age: 25, age range: 21-30), who gave written informed consent prior to the experiment, were recruited for the study. Participants reported no neurological or mental disorders. Non-invasive experiments on healthy subjects are exempt from ethical committee processing by Danish law [Den Nationale Videnskabsetiske Komité, 2014]. Among the 15 recordings, no participants were excluded, as we would like to generalize our results to a broad range of experimental recordings.

2.2. Stimuli

Stimuli consisted of 690 images from the Microsoft Common Objects in Context (MS COCO) dataset [Lin et al., 2014]. Images were selected from 23 semantic categories, with each category containing 30 images. All images presented were unique and not repeated for the same subject throughout the experiment. The initial selection criteria were 1) Image aspect ratio of 4:3, 2) Only a single super- and subcategory per image, and 3) Minimum 30 images within the category. Furthermore, we ensured that all 690 images had a relatively similar luminance and contrast to avoid the influence of low-level image features in the EEG signals. Thus, images within 77% of the brightness distribution and 87% of the contrast distribution were selected. Images that were highly distinct from standard MS COCO images were manually excluded (see Appendix B for exclusion criteria). Stimuli were presented using custom Python scripts built on PsychoPy2 software [Peirce, 2009].

2.3. EEG Data Collection

A user-friendly EEG equipment, Enobio (Neuroelectrics) with 32 channels and dry electrodes, was used for data acquisition. The EEG was electrically referenced using a CMS/DRL ear clip. The system recorded 24-bit EEG data with a sampling rate of 500 Hz, which was transmitted wirelessly using Wifi. LabRecorder was used for recording EEG signals. Lab Streaming Layer (LSL) was used to connect PsychoPy2 and LabRecorder for unified measurement of time series. The system was implemented on a Lenovo Legion Y520, and all recordings were performed in a normal office setting.

2.4. EEG Preprocessing

Preprocessing of the EEG was done using EEGLAB (sccn.ucsd.edu/eeglab). The EEG signal was bandpass filtered to 1-25 Hz and downsampled to 100 Hz. Artifact Subspace Reconstruction (ASR) [Mullen et al., 2015] was applied in order to reduce non-stationary high variance noise signals. Removed channels were interpolated from the remaining channels, and the data was subsequently re-referenced to an average reference. Epochs of 600 ms, 100 ms before and 500 ms after stimulus onset, similar to [Kaneshiro et al., 2015], were extracted for each trial. A sampling drift of 100 ms throughout the entire experiment was observed for all subjects and was corrected for offline.

Since the signal-to-noise ratio varied across trials and participants, all signals were normalized to z-score values (i.e., each trial and averaged trials from each participant was transformed so that it had a mean value of 0 and a standard deviation of 1 across time samples and channels).

2.5. Experimental Design

Participants were shown 23 blocks of trials composed of 30 images each. Of the 23 categories, 10 categories contained animals and the remaining 13 categories contained inanimate items, such as food or man-made objects. Each block corresponded to an image category (for categories and images used in the experiment, see Supplementary File 1), and the order of

categories and images within the categories was random for each subject. At the beginning of each category, a probe word denoting the category name was displayed for 5 s followed by 30 images from the corresponding category. Each image was displayed for 1 s, set against a mid-grey background. Inter-stimuli intervals (ISI) of variable length were displayed between each image. The ISI length was randomly sampled according to a uniform distribution from a fixed list of ISI values between 1.85 s and 2.15 s in 50 ms intervals, ensuring an average ISI duration of 2 s. To minimize eye movements between trials, the ISI consisted of a white fixation cross superimposed on a mid-grey background in the center of the screen.

Subjects viewed images on a computer monitor with a viewing distance of 57 cm. The size of stimuli was 4 x 3 degrees of the visual angle. Duration of the experiment was 39.3 min, which included five 35 s breaks interspersed between the 23 blocks. Before the experimental start, participants underwent a familiarization phase with two blocks of reduced length (103 s).

2.6. Support Vector Machines

Support vector machines (SVM) were implemented to classify the EEG data into two classes according to animate and inanimate trials. $y_i \in \{-1, 1\}$ is the identifier of the category, and an observation is defined to be the EEG response in one epoch ($[-100, 500]$ ms w.r.t. stimulus onset).

The SVM classifier is implemented by a nonlinear projection of the observations $\mathbf{x_n}$ into a high-dimensional feature space $\mathcal{F}$.

Let $\phi : \mathcal{X} \rightarrow \mathcal{F}$ be a mapping from the input space $\mathcal{X}$ to $\mathcal{F}$. The weight vector $\mathbf{w}$ can be expressed as a linear combination of the training points $\mathbf{w} = \sum_{n=1}^{N} \alpha_n \phi(\mathbf{x}_n)$ and the kernel trick is used to express the discriminant function as:

$$y(\mathbf{x}; \boldsymbol{\theta}) = \alpha^\top \mathbf{k_x} + b = \sum_{n=1}^{N} \alpha_n k(\mathbf{x}_n, \mathbf{x}) + b \qquad (1)$$

with the model now parametrized by the smaller set of parameters $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, b\}$ [Lautrup et al., 1994]. The Radial Basis Function (RBF) kernel allows for implementation of a non-linear decision boundary in the input space. The RBF kernel $\mathbf{k_x}$ holds the elements:

$$k(\mathbf{x}_n, \mathbf{x}) = \exp\left(-\gamma ||\mathbf{x}_n - \mathbf{x}||^2\right) \qquad (2)$$

where γ is a tunable parameter.

The SVM algorithm works by identifying a hyperplane in the feature space that optimally separates the two classes in the training data. Often it is desirable to allow a few misclassifications in order to obtain a better generalization error. This trade-off is controlled by a tunable regularization parameter c.

Two overall types of SVM classifiers were implemented: 1) Single-trial classifier, and 2) Average category level classifier. Both classifiers decode the supercategories, animate versus inanimate, and they both classify between subjects. The single-trial classifier is trained on 690 trials for each subject included in the training set. The average category classifier averages within the 23 categories for each subject, such that the classifier is trained on 23 epochs for each subject included in the training set, instead of 690 epochs.

The performance of the single-trial classifier was estimated using 14 participants as the training set, and the remaining participant was used as the test set (SVM parameters visualized in Figure S7). Cross-validation was performed on 10 parameter values in ranges $c = [0.05; 10]$ and $\gamma = [2.5 \times 10^{-7}; 5 \times 10^{-3}]$, thus cross-validating across 100 parameter combinations for each held out subject.

For an debiased estimate of test accuracy, the single-trial classifier was trained on 13 subjects, with 2 participants held out for validation and test in each iteration. Fifteen classifiers were trained with different subjects held out in each iteration. An optimal parameter set of c and γ was estimated using participants 1-7 as validation subjects (mean parameter value), which was used to estimate the test accuracy for subjects 8-15 and vice versa. Thus, two sets of optimal parameters were found (Figure S9). Cross-validation was performed on 10 parameter values in ranges $c = [0.25; 15]$ and $\gamma = [5 \times 10^{-7}; 2.5 \times 10^{-2}]$, i.e. 100 combinations.

The average category level classifier is much faster to train and was built using a basic nested leave-one-subject-out cross-validation loop. In the outer loop, one subject was held out for testing while the remaining 14 subjects entered the inner loop. The inner loop was used to estimate the optimum c and γ parameters for the SVM classifier. The performance of the model was calculated based on the test set. Each subject served as test set once. A permutation test was performed to check for significance. For each left out test subject, the animacy labels were permuted and compared to the predicted labels. This was repeated 1000 times, and the accuracy scores of the permuted sets were compared against the accuracy score of the non-permuted set. The upper level of performance was estimated by choosing the parameters based on the test set. Cross-validation was performed on 10 parameter values in ranges $c = [0.25; 15]$ and $\gamma = [5 \times 10^{-7}; 2.5 \times 10^{-2}]$, i.e 100 combinations.

2.7. Sensitivity mapping

To visualize the SVM RBF kernel, the approach proposed by [Rasmussen et al., 2011] was adapted. The sensitivity map is computed as the derivative of the RBF kernel, c.f. Eq. (2)

$$\frac{\partial \alpha^\top \mathbf{k_x}}{\partial x_j} = \sum_n \alpha_n 2\gamma(x_{n,j} - x_j) \exp(-\gamma \left\|\mathbf{x_n} - \mathbf{x}\right\|^2) \qquad (3)$$

Pseudo code for computing the sensitivity map across time samples and trials is found in Appendix A.

2.8. Effect size evaluation

The NPAIRS (nonparametric prediction, activation, influence, and reproducibility resampling) framework [Strother et al., 2002] was implemented to evaluate effect sizes of the SVM sensitivity map and animate/inanimate ERP differences. The sensitivity map and the ERP differences based on all subjects were thus scaled by the average difference of sub-sampled partitions.

The scaling was calculated based on $S = 100$ splits. In each split, two partitions of the data set were randomly selected without replacement. A partition consisted of 7 subjects, thus achieving two partitions of 7 subjects each (leaving a single, random subject out in each iteration).

For evaluation of the ERP difference map, a difference map was calculated for each partition ($\mathbf{M}_1$ and $\mathbf{M}_2$). Similarly, for evaluation of the sensitivity map, a SVM classifier was trained on each partition, and sensitivity maps were computed for both SVM classifiers (corresponding to $\mathbf{M}_1$ and $\mathbf{M}_2$ for the ERP difference map evaluation). The sensitivity map for the single-trial SVM classifier was computed based on optimal model parameters, while the sensitivity map of the average category classifier was based on the mean parameters as chosen by the validation sets. The maps from the two partitions were contrasted and squared. Across time samples ($t = 1, \ldots, T$) and trials ($n = 1, \ldots, N$) an average standard deviation of the average difference between partitions was calculated

$$\sigma^2 = \frac{1}{STN} \sum_{i,t,n=1}^{S,T,N} \left(\mathbf{M}^i_{1,t,n} - \mathbf{M}^i_{2,t,n}\right)^2 \tag{4}$$

The full map, $\mathbf{M}_{\text{full}}$ (based on 15 subjects) was then divided by the standard deviation to produce the effect size

$$\widehat{\mathbf{M}} = \frac{\mathbf{M}_{\text{full}}}{\sigma}. \tag{5}$$

3. RESULTS

We classify the recorded EEG using SVM RBF models such that trials are labeled with the high-level category of their presented stimuli, i.e., either animate or inanimate. We first report results using an average category classifier followed by a single trial classifier, and then apply the average category classifier for prediction of single trial EEG responses. Also, we report effect sizes of ERP difference maps and sensitivity maps for evaluation of both SVM classifiers.

3.1. Time Dependency

There will naturally be a temporal component in EEG. This unwanted non-stationarity of the signal can, for example, arise from electrodes gradually loosing or gaining connection to the scalp, an increasing tension of facial muscles or other artifactual currents [Delorme et al., 2007] [Rowan and Tolunsky, 2003]. If the data are epoched, the drift may misleadingly appear as a pattern reproducible over trials, a tendency that may be further reinforced by component analysis techniques that emphasize repeatable components [de Cheveigné and Arzounian, 2018].

Slow linear drifts can be removed by employing high pass filters, however more complicated temporal effects are harder to remove. We investigated the temporal trend both before and after Artifact Subspace Reconstruction (ASR) for each subject, see Figures S2 and S3. Generally, the time dependencies are reduced by ASR. However, the variance across trials remains time correlated for subjects 3 and 14 after ASR.

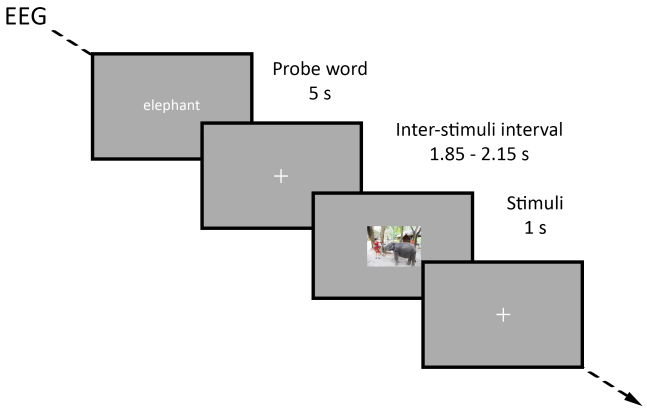

Fig. 1. Experimental design of the visual stimuli presentation paradigm. The time-course of the events is shown. Participants were shown a probe word before each category, and jittered inter-stimuli intervals consisting of a fixation cross was added between stimuli presentation. The experiment consisted of 690 trials in total, 23 categories of 30 trials, ordered randomly (both category- and image-wise) for each subject.

3.2. Event Related Potential analysis

After EEG data preprocessing, we confirmed that our visual stimuli presentation elicited a visually evoked potential response. The ERPs for the trials of animate content and the trials of inanimate content are compared in Figure 2. The grand average ERPs across subjects (thick lines) are shown along with the average animate and inanimate ERPs of each subject.

The average animate and inanimate ERPs were most different 310 ms after stimuli onset. The average scalp map at this time point can also be seen in Figure 2 for the two supercategories as well as the difference between them.

Inspection of Figure 2 shows that visual stimuli presentation elicited a negative ERP component at 80-100 ms poststimulus onset followed by a positive deflection at around 140

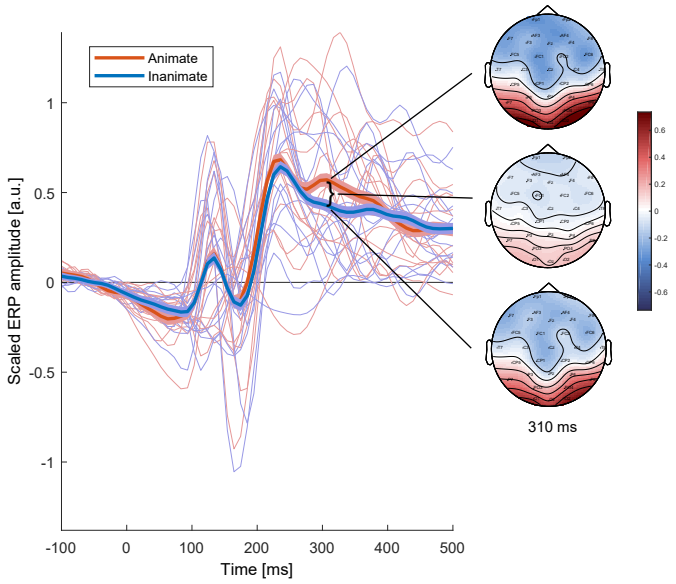

Fig. 2. Average animate and inanimate ERPs across subjects (thick lines) and for each subject (thin lines). ERP analysis was performed on the occipital/-parietal channels O1, O2, Oz, PO3 and PO4. Scalp maps are displayed for the animate/inanimate ERPs and difference thereof at 310 ms.

ms post-stimulus onset. A P300 subcomponent, P3a, was evident around 250 ms and a P3b component around 300 ms [Polich, 2007]. It is evident that the P3b component is more prominent for the animate category. The observed temporal ERP dynamics were comparable to prior ERP studies of the temporal dynamics of visual object processing [Cichy et al., 2014].

Mean animate/inanimate ERPs responses for each subject separately can be found in Figure S1.

3.3. Support Vector Machines

We sought to determine whether EEG data in our experiment can be automatically classified using SVM models. The Python toolbox Scikit-learn [Pedregosa et al., 2011] was used to implement RBF SVM models.

We specifically trained two different types of SVM classifiers, a single-trial, and an average category classifier, and assessed the classifiers' accuracy on labeling EEG data in a leave-one-subject-out approach.

SVMs are regarded efficient tools for high-dimensional binary as well as non-linear classification tasks, but their ultimate classification performance depends heavily upon the selection of appropriate parameters of c and γ [M Bishop, 2006]. Parameters for the upper level of performance for the single-trial classifier were found using cross-validation in a leave-one-subject-out approach, resulting in a penalty parameter $c = 1.5$ and $\gamma = 5 \times 10^{-5}$ based on the optimum mean parameters across test subjects (Figure S8). From Figure S7 it is evident that the optimum parameters were different for each subject, underlining inter-subject variability in the EEG responses.

To reduce bias of the performance estimate of the single-trial classifier, parameters were selected based on two validation partitions, resulting in $c = 0.5$ and $\gamma = 5 \times 10^{-4}$ for the first validation set, and $c = 1.5$ and $\gamma = 5 \times 10^{-5}$ for the second validation set (Figure S9).

The average category classifier also showed inter-subject variability with respect to the model parameters, see Figures S4-S6. The classifier had an average penalty parameter of $c = 7.2$, and an average $\gamma = 3.7 \times 10^{-4}$ when based on the validation sets. The average optimum parameters when based on test sets with averaged categories and single-trials were in the same range, with $c = 4.4$ and $\gamma = 3.0 \times 10^{-4}$ and $c = 6.7$ and $\gamma = 2.2 \times 10^{-4}$, respectively.

Figures 4-5 show the SVM classification performances using the two types of classifiers. Based on the leave-one-subject-out classification, we note the large variability of single subject performance. While different performances are obtained using the single-trial and average trial classifiers on single-trial test sets, the overall accuracies are similar, with an average of 0.574 and 0.575, respectively (Figure 5).

For some subjects, low accuracy is caused by a parameter mismatch between trials belonging to that subject and its validation sets. For other subjects, the SVM model is not capable of capturing their data even when parameters are based on that subject, due to poor signal to noise level.

A standard error of the mean of 0.01 was found for both the debiased performance measure of the single-trial classifier and for the unbiased single-trial classifier (corrected for the leave-one-subject-out approach). **GT: REF til dette/yderlige beskrivelse?**

3.4. Event Related Potential Difference Map and Sensitivity Map

We investigated the raw ERP difference map between animate and inanimate categories, as well as the sensitivity map for the single-trial and average category SVM classifiers. The sensitivity map reveals EEG time points and channels that are of relevance to the SVM decoding classifiers.

For effect size evaluation we implement an NPAIRS resampling scheme [Strother et al., 2002]. In this cross-validation framework, the data were split into two partitions of equal size (7 subjects in each partition randomly selected without replacement). This procedure was repeated 100 times to obtain standard errors of the maps for computing effect sizes (Section 2.8).

From inspection of Figure 3 it is evident that occipital and parietal channels (O1, O2, P7, P8) were relevant for SVM classification at time points comparable to the ERP difference map. Frontal channels (Fp1, Fp2) were exploited by

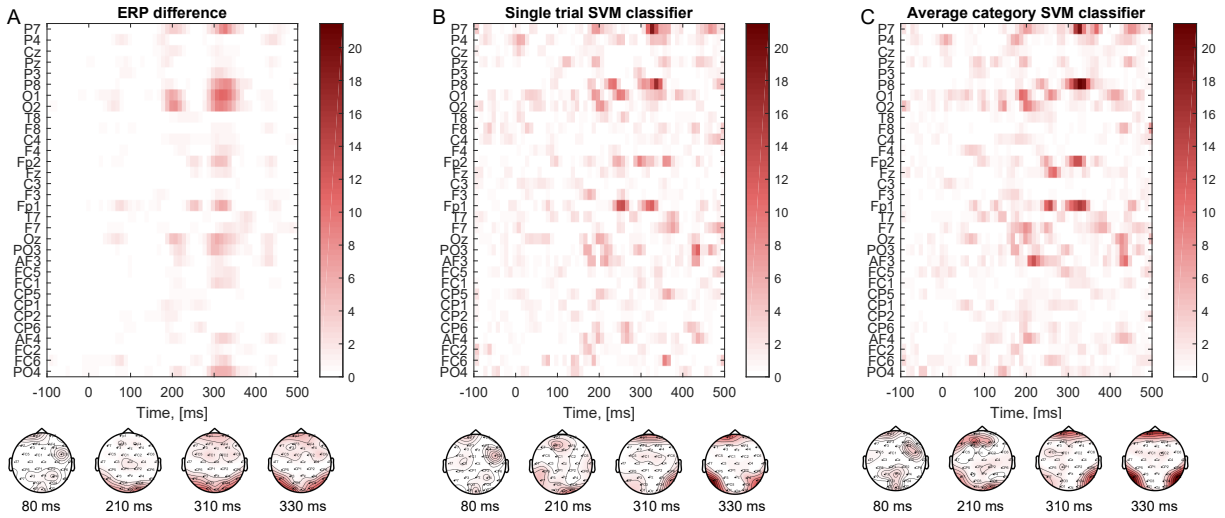

Fig. 3. Effect sizes for ERP animate/inanimate difference map and single-trial and average category SVM classifiers. Effect sizes were computed based on 100 NPAIRS resampling splits.

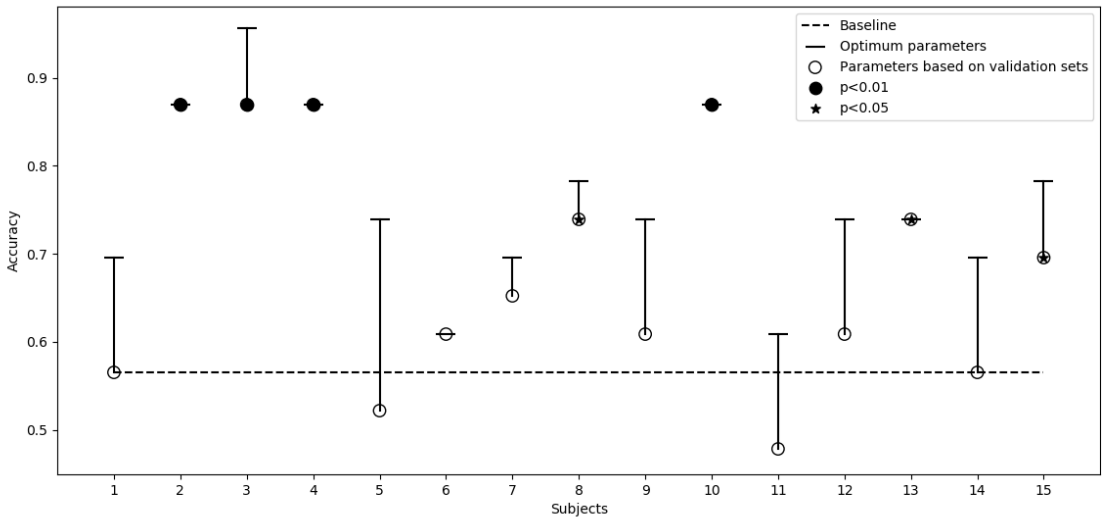

Fig. 4. SVM classifier trained on average categories and tested on average categories. Significance estimated by permutation testing.

both SVM classifiers, but to a larger extent by the average category classifier (Figure 3C). Furthermore, the average category classifier exploited a larger proportion of earlier time points compared to the single-trial classifier. The sensitivity maps for the single-trial and average category classifiers suggest that despite the difference in number and type of trials, the classifiers are similar.

4. DISCUSSION

In the current work, we approach the challenges of EEG-based decoding: non-laboratory settings, user-friendly EEG acquisition equipment with dry electrodes, naturalistic stimuli, no repeated presentation of stimuli and no exclusion of participants. The potential benefits of mitigating these challenges is to study the brain dynamics in natural settings.

We aim to increase the EEG applicability in terms of cost, mobility, set-up and real-time acquisition by using commercial-grade EEG equipment. Moreover, it has recently

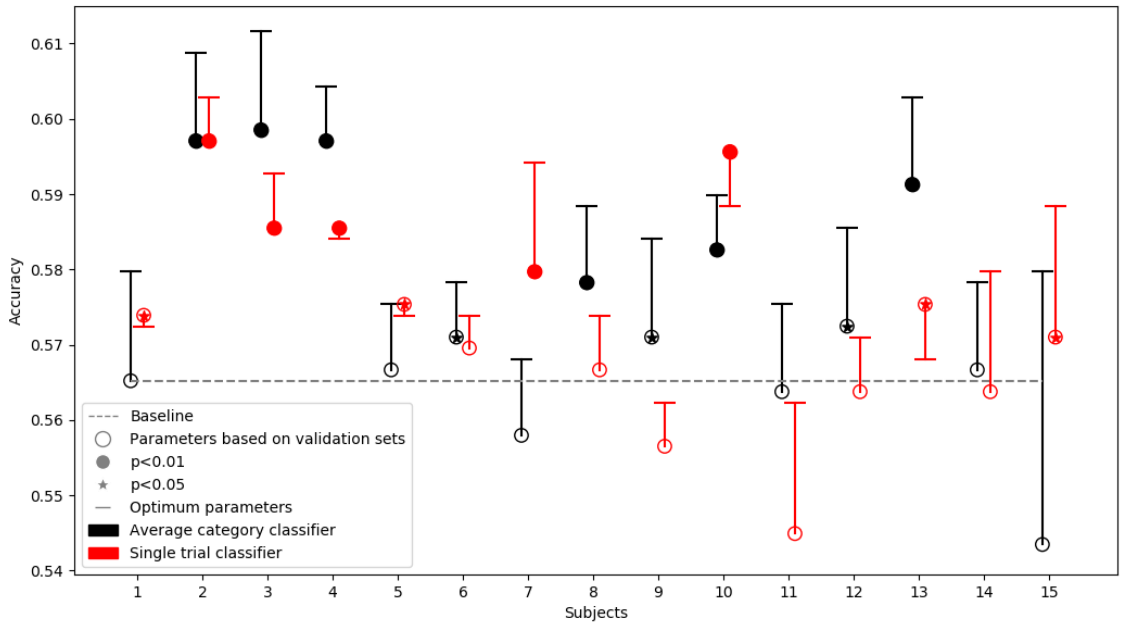

Fig. 5. Classifier trained on average categories (black) or trained on single trials (red) and tested on single trials. Note: In some cases the "optimum parameters" are not found to be optimum, which can be explained by different training phases of the two single trial classifiers. The classifier based on validation sets was trained on 13 subjects while the classifier with parameters based on the test set was trained on 14 subjects. For 5 out of 15 subjects the classifier based on 13 subjects was able to obtain higher accuracies.

been demonstrated that commercial-grade EEG equipment compares to high-grade equipment in laboratory settings in terms of neural reliability as quantified by inter-subject correlation [Poulsen et al., 2017]. Furthermore, a systematic comparison shows similar quality of wet and dry electrodes [Kam et al., 2019].

The stimuli in our experimental paradigm consisted of complex everyday scenes and non-iconic views of objects [Lin et al., 2014]. Animate and inanimate images were similar in composition, i.e., an object or animal in its natural surroundings. The presented images were not close-ups of animals/objects as in many previous ERP classification studies, for instance when distinguishing "tools" from "animals" [Simanova et al., 2010, Murphy et al., 2011] or "faces" from "objects" [Kaneshiro et al., 2015].

Our ultimate goal is to decode actual semantic differences between categories; thus we perform low-level visual feature standardization on experimental trials prior to the experiment, investigate time dependency of the EEG response throughout the experiment (Figures S2-S3) and perform ASR to reduce this dependency.

4.1. Event Related Potential Analysis

Previous work on visual stimuli decoding demonstrate that visual input reaches the frontal lobe as early as 40-65 ms after image presentation and participants can make an eye movement 120 ms after onset in a category detection task. Evidence of category specificity has been found at both early ($\sim$ 150 ms) and late ($\sim$ 400 ms) intervals of the visually evoked potential [Rousselet et al., 2004, Rousselet et al., 2007]. ERP studies indicate that category-attribute interactions (natural/non-natural) emerge as early as 116 ms after stimulus onset over fronto-central scalp regions, and at 150 and 200 ms after stimulus onset over occipitoparietal scalp regions [Hoenig et al., 2008]. For animate versus inanimate images, ERP differences have been demonstrated detectable within 150 ms of presentation [Thorpe et al., 1996]. Kaneshiro et al., 2015, demonstrate that the first 500 ms of single-trial EEG responses contain information for successful category decoding between human faces and objects, and above chance object classification as early as 48-128 ms after stimulus onset [Kaneshiro et al., 2015].

However, there appears to be uncertainty whether these early ERP differences represent low-level visual stimuli or actual high-level differences. We observe the major differ-

ence between animate/inanimate ERPs at around 310 ms (Figures 2-3). Moreover, we find ERP signatures different among subjects (comparable to [Simanova et al., 2010]), which challenges the across-subject model generalizability with our sample size of 15 subjects.

4.2. Support Vector Machine Classification

In this study, we adopted non-linear RBF kernel SVM classifiers to classify between animate/inanimate naturalistic visual stimuli in a leave-one-subject-out approach.

SVM classifiers have previously been implemented for EEG-based decoding. SVM in combination with independent component analysis data processing has been used to classify whether a visual object is present or absent from EEG [Stewart et al., 2014]. Zafar et al., 2017, propose a hybrid algorithm using convolutional neural networks (CNN) for feature extraction and likelihood-ratio-based score fusion for prediction of brain activity from EEG [Zafar et al., 2017]. Taghizadeh-Sarabi et al., 2015, extract wavelet features from EEG, and selected features are classified using a "one-against-one" SVM multiclass classifier with optimum SVM parameters set separately for each subject [Taghizadeh-Sarabi et al., 2014].

We found very similar performance of the single-trial and average trial classifiers (Figure 5). As the average classifier is significantly faster to train, a full nested cross-validation scheme was feasible. The fact that the two classifiers have similar performance indicates that the reduced sample size in the average classifier is offset by these (averaged) samples better signal to noise ratio. The fast training of the average trial classifier allows for parameter optimization and unbiased performance evaluation.

Based on the leave-one-subject-out classification performance (Figures 4-5), it is evident that there is a difference in how well the classifier generalizes across subjects, which partly is due to the diversity of ERP signatures across subjects (Figure S1). This can be explained by a manifestation of inherent inter-subject variability in EEG. Across-participant generalization in EEG is complicated by many factors: the signal to noise ratio at each electrode is affected by the contact to the scalp which is influenced by local differences in skin condition and hair, the spatial location of electrodes relative to underlying cortex will vary according to anatomical head differences, and there may be individual differences in functional localization across participants. **Motivation/discussion about inter-subject approach?**

Both SVM classifiers utilized a relatively large number of support vectors. The single-trial SVM classifier used for computing the sensitivity map had model coefficients $\alpha = -1.5, \ldots, 1.5$, where 1204 α values out of 10350 were equal to 0 (9146 support vectors). The average category classifier had model coefficients in the range $\alpha = -7.2, \ldots, 7.2$, and 46 out of 345 coefficients were zero (299 support vectors). **korrekt?** In both cases a high number of support vectors is obtained, which indicates the complexity of the problem as well as a poor EEG signal-to-noise ratio.

4.3. Sensitivity Mapping

In the current work, we ask which parts of the EEG signatures are being exploited by the SVM decoding classifiers. We investigated the probabilistic sensitivity map for single-trial and average trial SVM classifiers based on a binary classification task. We identified regions where discriminative information resides and found this information comparable to the difference map between ERP responses for animate and inanimate trials.

While previous studies demonstrate that high-level categories in EEG are distinguishable before 150 ms after stimulus onset [Rousselet et al., 2007, Hoenig et al., 2008, Kaneshiro et al., 2015], we find the most prominent difference in animate/inanimate ERPs around 210 ms and 320 ms, which is also exploited by the SVM classifiers (Figure 3). **Frontal electrode discussion?**

Based on the similarity between the sensitivity maps for single-trial and average category classifiers (Figure 3), we conclude that these classifiers exploit the same EEG features to a large extent. We therefore investigated whether the average category classifier successfully is able to predict on single-trial test subjects. We show that classifiers trained on averaged trials perform at par with classifiers trained on a large set of noisy single-trial samples.

4.4. Conclusion

We investigate scalp EEG recorded with a user-friendly, portable 32 dry electrode equipment from healthy subjects under naturalistic stimuli. We accomplish unbiased decoding of single-trial EEG using SVM models trained on de-noised (averaged) trials, thus facilitating fast classifier training, parameter optimization and unbiased performance evaluation. The SVM classifiers were evaluated in a inter-subject approach, thus probing generalizability across participants. We propose a novel methodology for evaluating and computing sensitivity maps for EEG-based SVM classifiers, allowing for visualization of discriminative SVM classifier information. Finally, by linking temporal and spatial features of EEG to training of SVM classifiers, we take an essential step in understanding how machine learning techniques exploit neural signals.

5. OVERVIEW SUPPLEMENTARY MATERIAL

Appendix A: Sensitivity map Python code
Appendix B: Manual exclusion criteria for image selection

Supplementary file 1: Image IDs, supercategories and categories for all images used in the experiment from Microsoft

Common Objects in Context (MS COCO) image database.

Figures S1-S9 contain supplementary material, and are used for reference in the main manuscript.

6. DATA AVAILABILITY

Code available:
https://github.com/gretatuckute/DecodingSensitivityMapping.

7. CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest regarding the publication of this paper.

8. FUNDING STATEMENT

This work was supported by the Novo Nordisk Foundation Interdisciplinary Synergy Program 2014 (Biophysically adjusted state-informed cortex stimulation (BASICS)) [NNF14OC0011413].

9. AUTHORS' CONTRIBUTIONS

G.T, N.P and L.K.H. designed research; G.T and N.P acquired data; D.S, G.T and N.P performed initial data analyses; G.T., S.T.H and L.K.H performed research; G.T, S.T.H and L.K.H wrote the paper.

10. REFERENCES

[Andersen et al., 2017] Andersen, R. S., Eliasen, A. U., Pedersen, N., Andersen, M. R., Hansen, S. T., and Hansen, L. K. (2017). EEG source imaging assists decoding in a face recognition task. *arXiv preprint arXiv:1704.05748*.

[Cichy et al., 2014] Cichy, R. M., Pantazis, D., and Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, 17(3):455–462.

[de Cheveigné and Arzounian, 2018] de Cheveigné, A. and Arzounian, D. (2018). Robust detrending, rereferencing, outlier detection, and inpainting for multichannel data. *NeuroImage*, 172(December 2017):903–912.

[Delorme et al., 2007] Delorme, A., Sejnowski, T., and Makeig, S. (2007). Enhanced detection of artifacts in eeg data using higher-order statistics and independent component analysis. *Neuroimage*, 34(4):1443–1449.

[Den Nationale Videnskabsetiske Komité, 2014] Den Nationale Videnskabsetiske Komité (2014). Vejledning om anmeldelse, indberetning mv. (sundhedsvidenskablige forskningsprojekter). (Januar):116.

[Gerlach, 2007] Gerlach, C. (2007). A review of functional imaging studies on category specificity. *Journal of Cognitive Neuroscience*, 19(2):296–314.

[Güçlü and van Gerven, 2017] Güçlü, U. and van Gerven, M. A. (2017). Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, 145:329–336.

[Hoenig et al., 2008] Hoenig, K., Sim, E.-J., Bochev, V., Herrnberger, B., and Kiefer, M. (2008). Conceptual Flexibility in the Human Brain: Dynamic Recruitment of Semantic Maps from Visual, Motor, and Motion-related Areas. *Journal of Cognitive Neuroscience*, 20(10):1799–1814.

[Huth et al., 2016] Huth, A. G., Lee, T., Nishimoto, S., Bilenko, N. Y., Vu, A. T., and Gallant, J. L. (2016). Decoding the Semantic Content of Natural Movies from Human Brain Activity. *Frontiers in Systems Neuroscience*, 10(October):1–16.

[Huth et al., 2012] Huth, A. G., Nishimoto, S., Vu, A. T., and Gallant, J. L. (2012). A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. *Neuron*, 76(6):1210–1224.

[Kam et al., 2019] Kam, J. W., Griffin, S., Shen, A., Patel, S., Hinrichs, H., Heinze, H. J., Deouell, L. Y., and Knight, R. T. (2019). Systematic comparison between a wireless EEG system with dry electrodes and a wired EEG system

with wet electrodes. *NeuroImage*, 184(August 2018):119–129.

[Kaneshiro et al., 2015] Kaneshiro, B., Perreau Guimaraes, M., Kim, H.-S., Norcia, A. M., and Suppes, P. (2015). A Representational Similarity Analysis of the Dynamics of Object Processing Using Single-Trial EEG Classification. *Plos One*, 10(8):e0135697.

[Kay et al., 2008] Kay, K., Naselaris, T., Prenger, R., and Gallant, J. (2008). Identifying natural images from human brain activity. *Nature*, 452(7185):352–355.

[Lautrup et al., 1994] Lautrup, B., Hansen, L. K., Law, I., Mørch, N., Svarer, C., and Strother, S. (1994). Massive Weight-Sharing: A Cure for Extremely Ill-Posed Problems. *Workshop on Supercomputing in Brain Research: From Tomography to Neural Networks, J{ü}lich, Germany*, page 137.

[Lin et al., 2014] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8693 LNCS(PART 5):740–755.

[Lotte et al., 2007] Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., and Arnaldi, B. (2007). A review of classification algorithms for EEG-based brain-computer interfaces. *J. Neural Eng.*, 4(R1):R1–R13.

[Loula et al., 2018] Loula, J., Varoquaux, G., and Thirion, B. (2018). Decoding fmri activity in the time domain improves classification performance. *NeuroImage*, 180:203–210.

[M Bishop, 2006] M Bishop, C. (2006). *Pattern recognition and machine learning*. Springer-Verlag New York.

[Mullen et al., 2015] Mullen, T., Kothe, C., Chi, M., Ojeda, A., Kerth, T., Makeig, S., Jung, T.-P., and Cauwenberghs, G. (2015). Real-time Neuroimaging and Cognitive Monitoring Using Wearable Dry EEG. *IEEE Transactions on Biomedical Engineering*, 62(11):2553–2567.

[Murphy et al., 2011] Murphy, B., Poesio, M., Bovolo, F., Bruzzone, L., Dalponte, M., and Lakany, H. (2011). EEG decoding of semantic category reveals distributed representations for single concepts. *Brain and Language*, 117(1):12–22.

[Nicolas-Alonso and Gomez-Gil, 2012] Nicolas-Alonso, L. F. and Gomez-Gil, J. (2012). Brain computer interfaces, a review. *Sensors*, 12(2):1211–1279.

[Nishimoto et al., 2011] Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Peirce, 2009] Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. 2(January):1–8.

[Polich, 2007] Polich, J. (2007). Updating P300: An Integrative Theory of P3a and P3b. *Clin Neurophysiol.*, 118(10):2128–2148.

[Poulsen et al., 2017] Poulsen, A. T., Kamronn, S., Dmochowski, J., Parra, L. C., and Hansen, L. K. (2017). EEG in the classroom: Synchronised neural recordings during video presentation. *Scientific Reports*, 7:1–9.

[Prenger et al., 2009] Prenger, R. J., Gallant, J. L., Kay, K. N., Naselaris, T., and Oliver, M. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915.

[Rasmussen et al., 2011] Rasmussen, P. M., Madsen, K. H., Lund, T. E., and Hansen, L. K. (2011). Visualization of nonlinear kernel models in neuroimaging by sensitivity maps. *NeuroImage*, 55(3):1120–1131.

[Rousselet et al., 2007] Rousselet, G. A., Husk, J. S., Bennett, P. J., and Sekuler, A. B. (2007). Single-trial EEG dynamics of object and face visual processing. *NeuroImage*, 36(3):843–862.

[Rousselet et al., 2004] Rousselet, G. A., Mace, M. J.-M., and Fabre-Thorpe, M. (2004). Animal and human faces in natural scenes: How specific to human faces is the N170 ERP component? *Journal of Vision*, 4(1):2–2.

[Rowan and Tolunsky, 2003] Rowan, A. J. and Tolunsky, E. (2003). *A primer of EEG: with a mini-atlas*. Butterworth-Heinemann Medical.

[Saitta, 1995] Saitta, L. (1995). Support-Vector Networks SVM.pdf. 297:273–297.

[Simanova et al., 2010] Simanova, I., van Gerven, M., Oostenveld, R., and Hagoort, P. (2010). Identifying object categories from event-related EEG: Toward decoding of conceptual representations. *PLoS ONE*, 5(12).

[Stewart et al., 2014] Stewart, A. X., Nuthmann, A., and Sanguinetti, G. (2014). Single-trial classification of EEG in a visual object task using ICA and machine learning. *Journal of Neuroscience Methods*, 228:1–14.

[Strother et al., 2002] Strother, S. C., Anderson, J., Hansen, L. K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., and Rottenberg, D. (2002). The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework. *NeuroImage*, 15(4):747–771.

[Taghizadeh-Sarabi et al., 2014] Taghizadeh-Sarabi, M., Daliri, M. R., and Niksirat, K. S. (2014). Decoding Objects of Basic Categories from Electroencephalographic Signals Using Wavelet Transform and Support Vector Machines. *Brain Topography*, 28(1):33–46.

[Thorpe et al., 1996] Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381.

[Zafar et al., 2017] Zafar, R., Dass, S. C., and Malik, A. S. (2017). Electroencephalogram-based decoding cognitive states using convolutional neural network and likelihood ratio based score fusion. *PLoS ONE*, 12(5):1–23.

Supplementary materials

Appendix A
The following piece of code shows how to compute the sensitivity map for a SVM classifier with an RBF kernel across all trials using Python and NumPy (np).

```
                                                map =
np.matmul(X,np.matmul(np.diag(alpha),k))-(np.matmul(X,(np.diag(np.matmul(alpha,k)))))
                    s = np.sum(np.square(map),axis=1)/np.size(alpha)
```

`k` denotes the $(N \times N)$ RBF training kernel matrix from equation 2, with N as the number of training examples. `alpha` denotes a $(1 \times N)$ vector with model coefficients. `X` denotes a $(P \times N)$ matrix with training examples in columns. `s` is a $(P \times 1)$ vector with estimates of channel sensitivities for each time point, which can be re-sized into a matrix of size [no. channels, no. time points] for EEG-based sensitivity map visualization.

Appendix B
Manual exclusion criteria for MS COCO images [Lin et al., 2014] for the experimental paradigm:
- Object unidentifiable
- Object not correctly categorized
- Different object profoundly more in focus
- Color scale manipulation
- Frame or text overlay on image
- Distorted photograph angle
- Inappropriate image

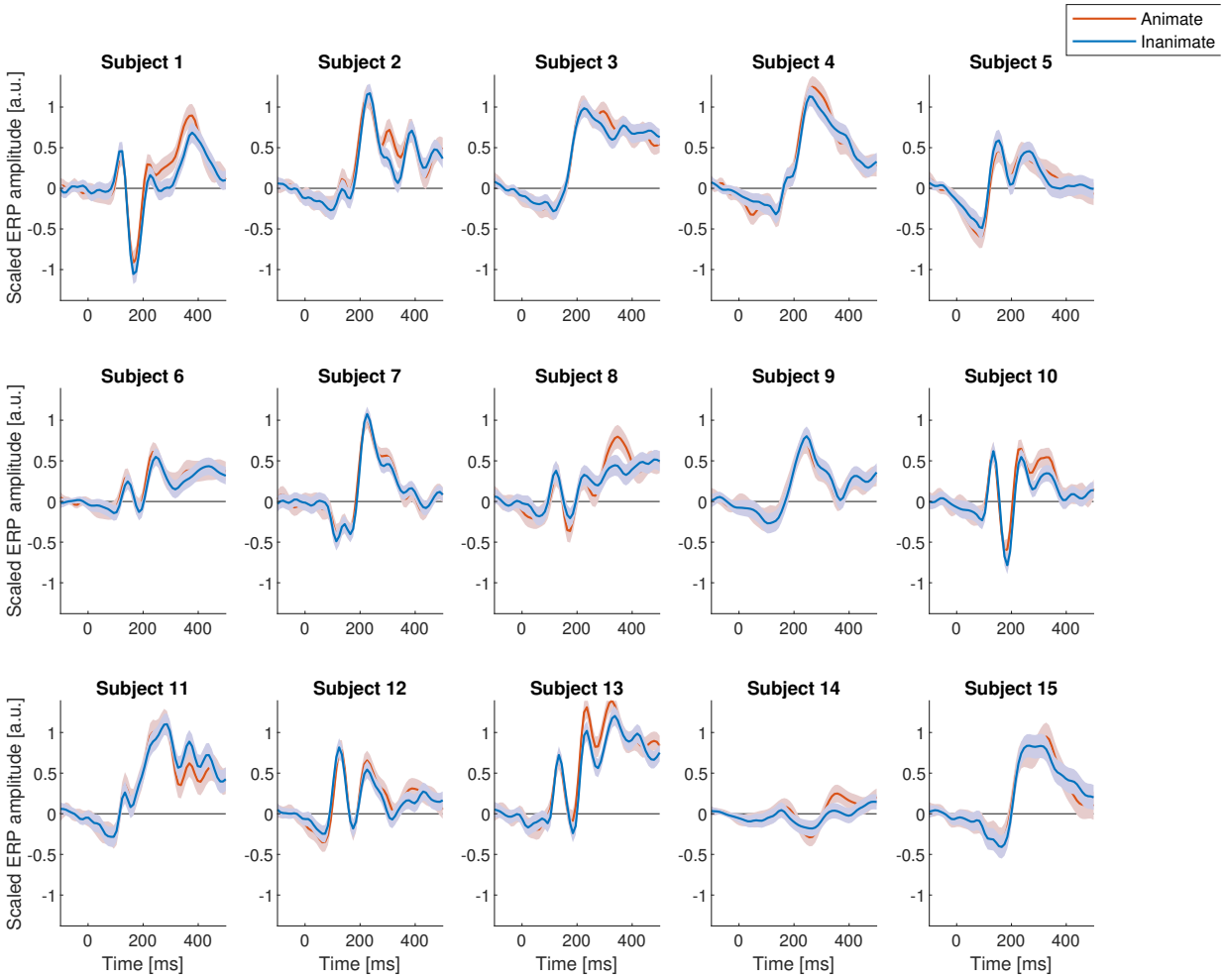

Fig. S1. Animate and inanimate ERPs for each subject separately with two standard errors around the mean.

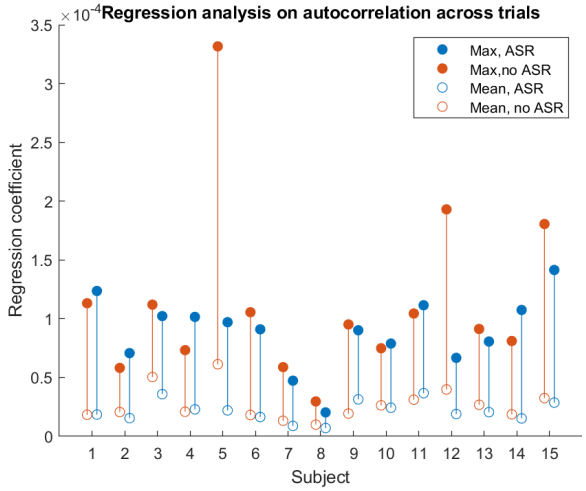

Fig. S2. Time dependency as quantified by autocorrelation. A high regression coefficient means that a channel (or an average of all channels) had an autocorrelation which increased or decreased linearly with time lags, and is thus indicative of high time dependency.

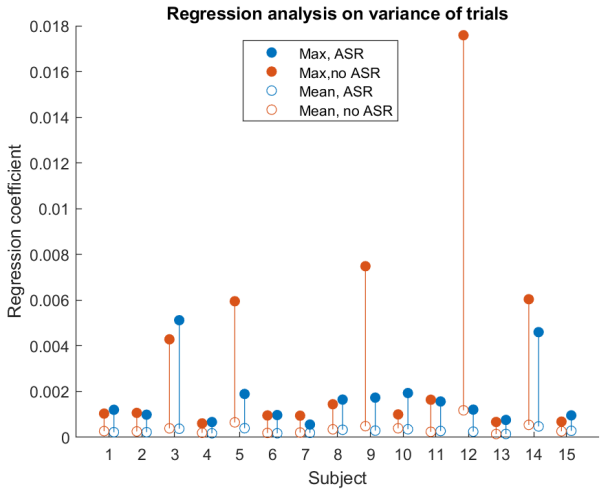

Fig. S3. Time dependency as quantified by variance of trials. A high regression coefficient means that a channel (or an average of all channels) had a trial variance which increased or decreased linearly with trial number, and is thus indicative of high time dependency.

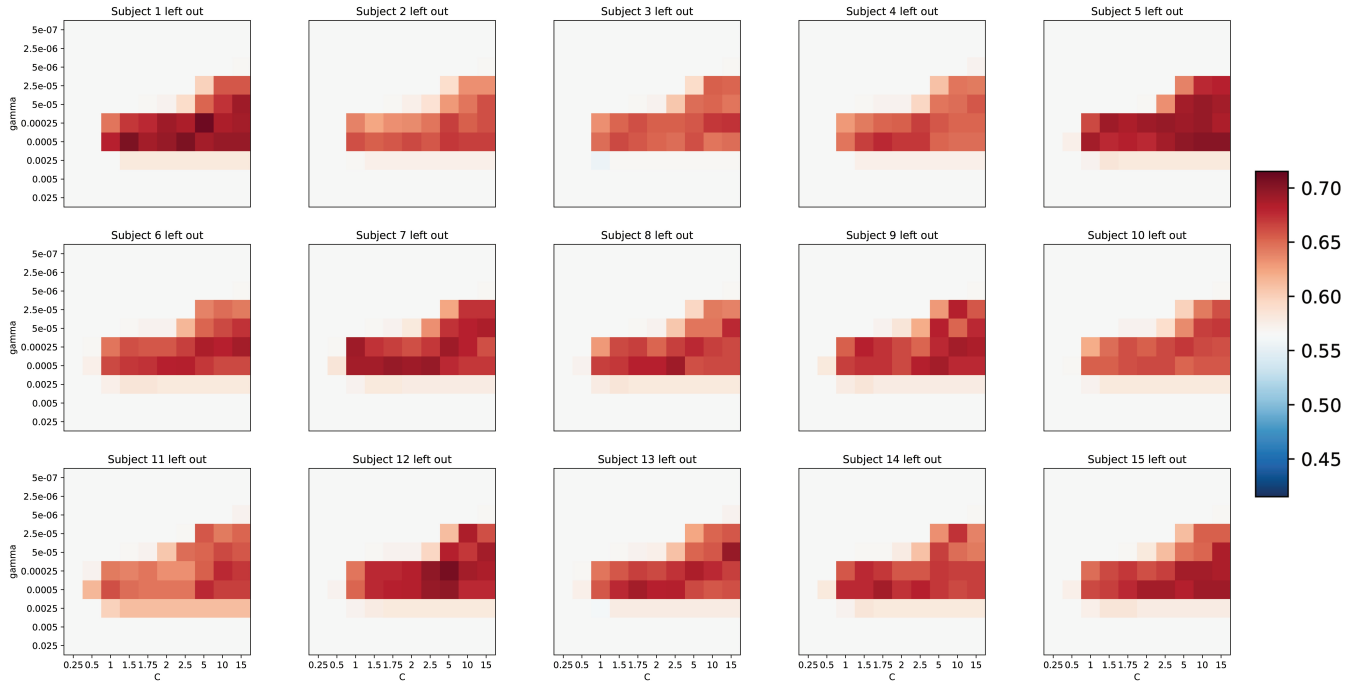

Fig. S4. Validation accuracies (mean over validation sets) for the average category classifier. c values are displayed on the x-axis, and consisted of values: $[0.25, 0.5, 1, 1.5, 1.75, 2, 2.5, 5, 10, 15]$. γ values are displayed on the y-axis, and consisted of values: $[5 \times 10^{-7}, 2.5 \times 10^{-6}, 5 \times 10^{-6}, 2.5 \times 10^{-5}, 5 \times 10^{-5}, 2.5 \times 10^{-4}, 5 \times 10^{-4}, 2.5 \times 10^{-3}, 5 \times 10^{-3}, 2.5 \times 10^{-2}]$. Same scaling for all subjects.

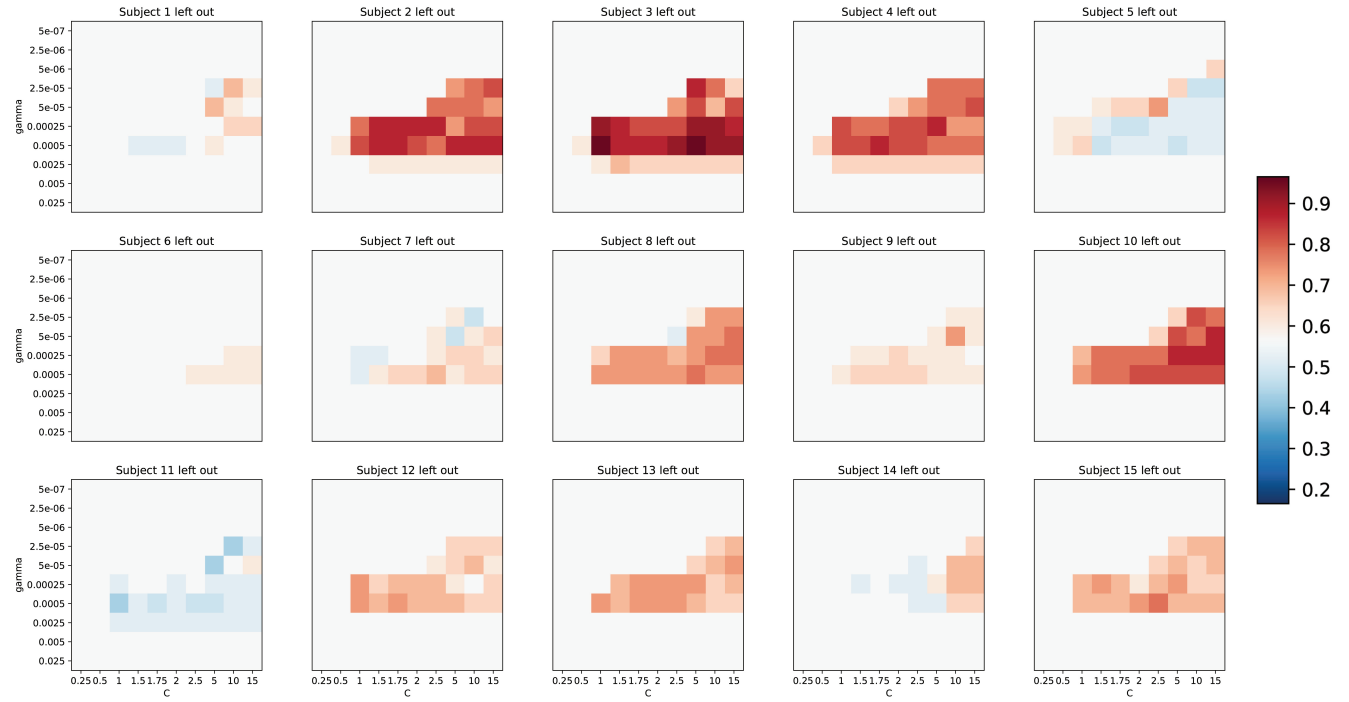

Fig. S5. Test accuracies for average category classifier. Tested on the averaged categories of the withheld subject. Same cross-validation parameter values as in Figure S4.

.

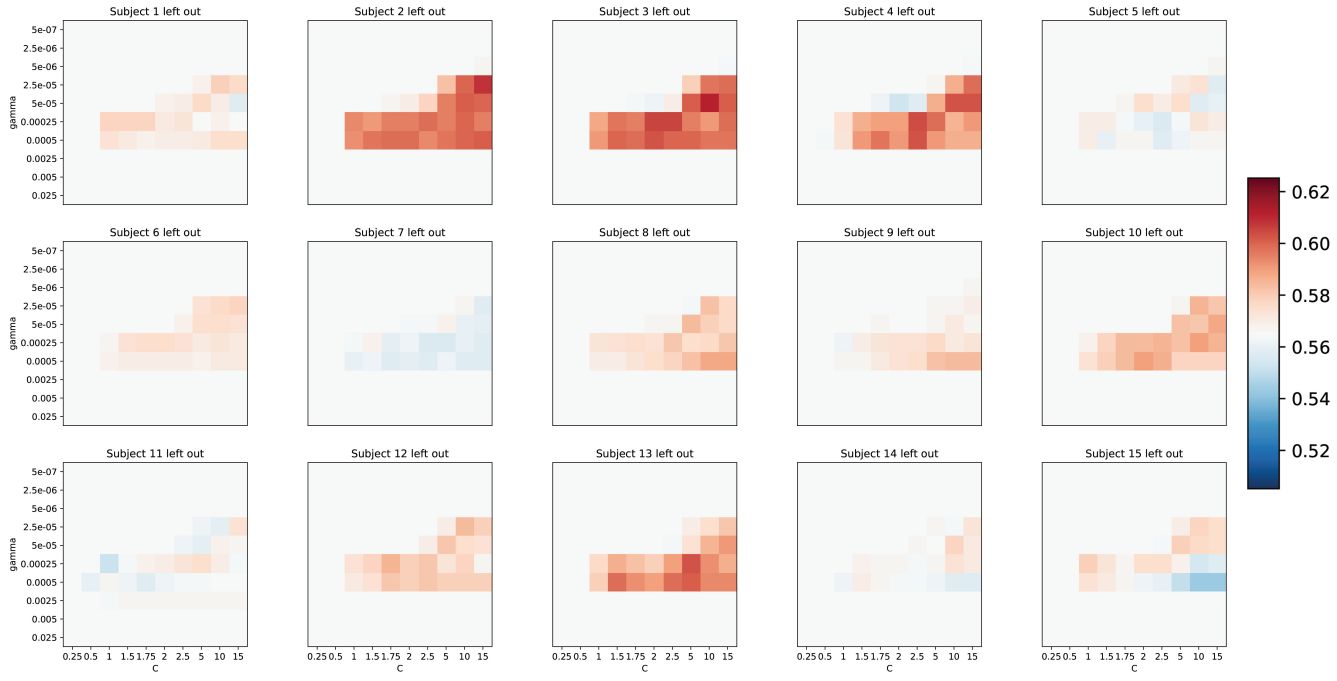

Fig. S6. Test accuracies for average category classifier. Tested on the individual trials of the withheld subject. Same cross-validation parameter values as in Figure S4.

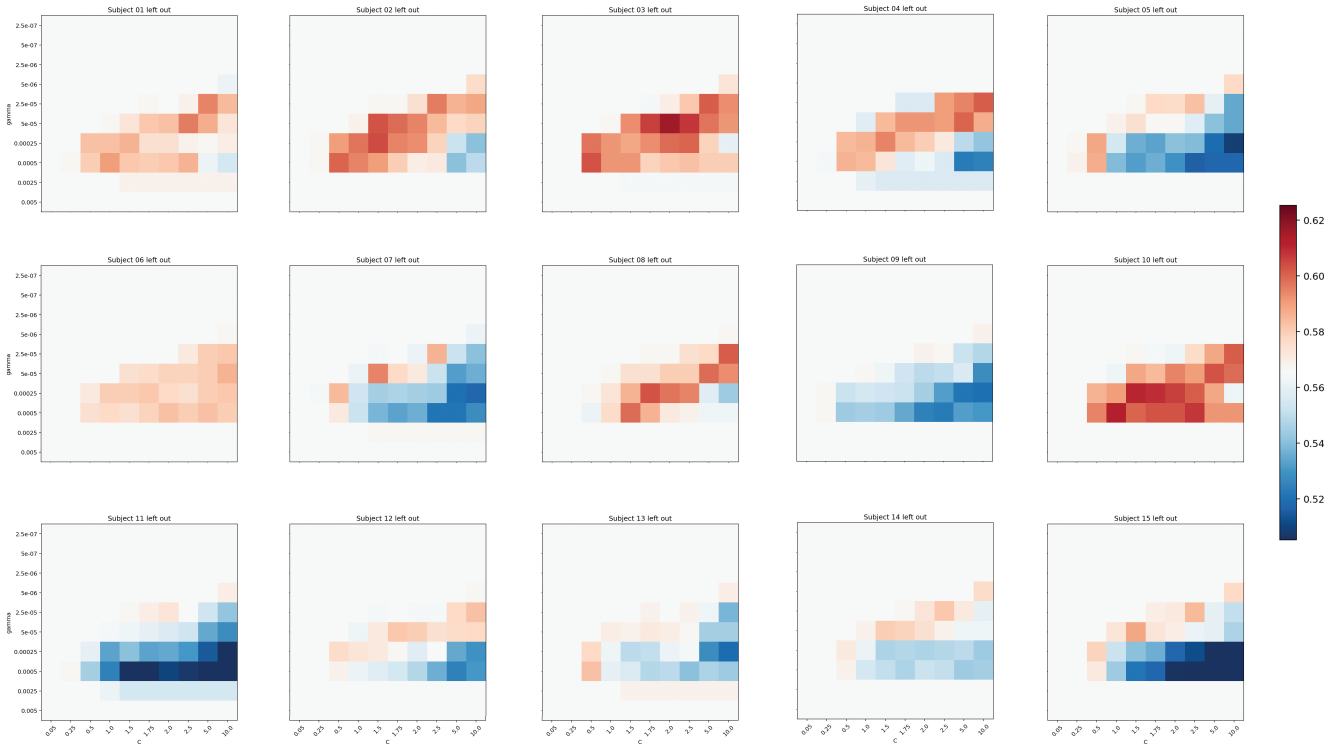

Fig. S7. Cross-validation with a single held out subject to estimate parameters for the upper level performance single-trial SVM classifier. c values are displayed on the x-axis, and consisted of values: $[0.05, 0.25, 0.5, 1, 1.5, 1.75, 2, 2.5, 5, 10]$. γ values are displayed on the y-axis, and consisted of values: $[2.5 \times 10^{-7}, 5 \times 10^{-7}, 2.5 \times 10^{-6}, 5 \times 10^{-6}, 2.5 \times 10^{-5}, 5 \times 10^{-5}, 2.5 \times 10^{-4}, 5 \times 10^{-4}, 2.5 \times 10^{-3}, 5 \times 10^{-3}]$.

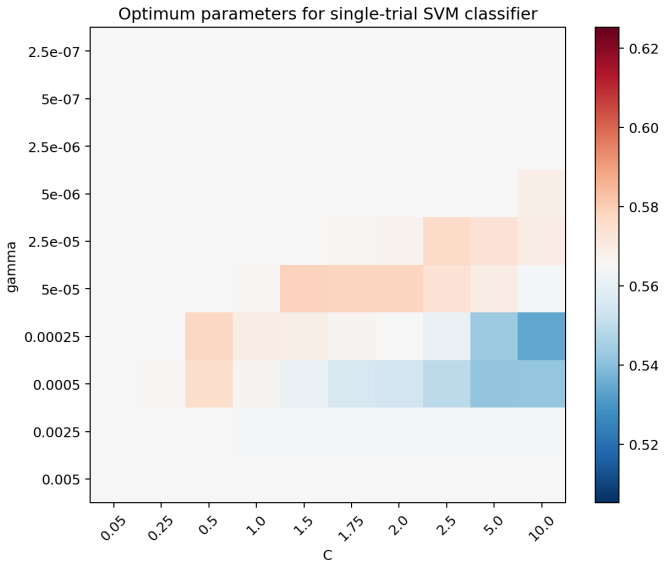

Fig. S8. Upper level performance parameters for the single-trial SVM classifier based on the mean parameters for held out subjects in Figure S7.

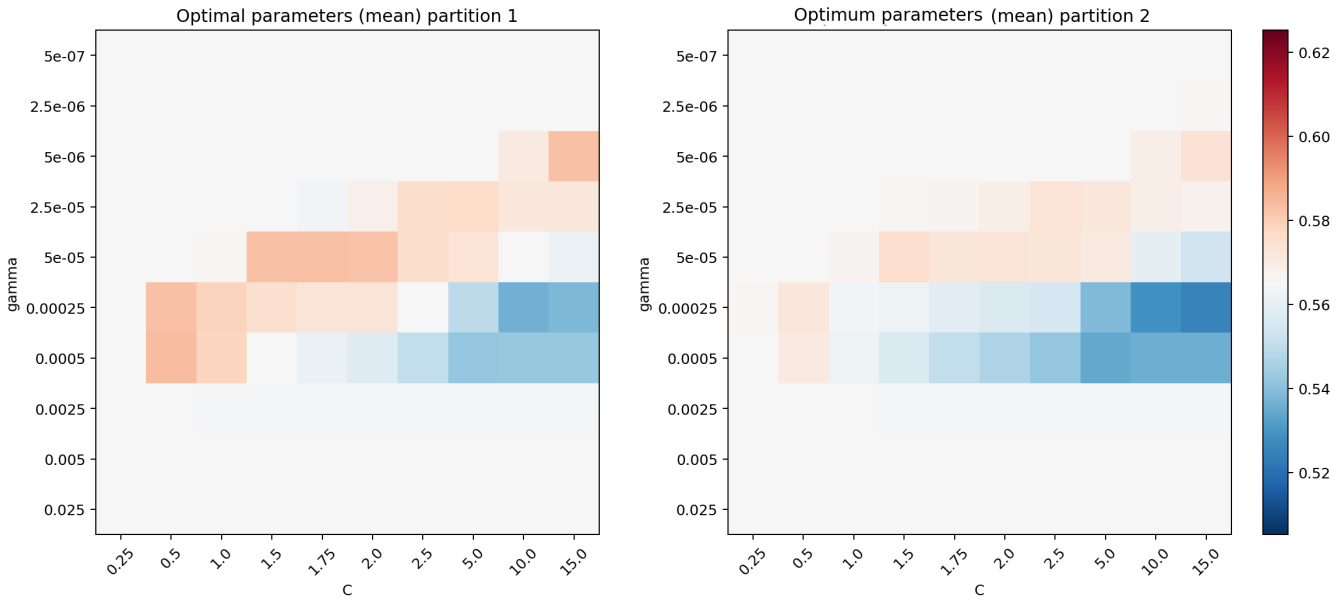

Fig. S9. Optimum parameters for the single-trial SVM classifier based on the mean parameters of validation partition 1 (subjects 1-7) and partition 2 (subjects 8-15). Same cross-validation parameter values as in Figure S4.