VISUALIZATION OF NONLINEAR CLASSIFICATION MODELS IN NEUROIMAGING

Signed sensitivity maps

Peter M. Rasmussen[1,2], Tanya Schmah[3], Kristoffer H. Madsen[1,4], Torben E. Lund[2], Grigori Yourganov[5,6], Stephen C. Strother[5,7], Lars K. Hansen[1]

[1]*DTU Informatics, Technical University of Denmark, Kgs. Lyngby, Denmark.*

[2]*The Danish National Research Foundation's Center for Functionally Integrative Neuroscience, Aarhus University Hospital, Denmark.*

[3]*Department of Computer Science, University of Toronto, Toronto, Canada*

[4]*Danish Research Centre for Magnetic Resonance, Copenhagen University Hospital Hvidovre, Denmark.*

[5]*Rotman Research Institute, Baycrest Centre for Geriatric Care, Toronto, Canada*

[6]*Institute of Medical Science, University of Toronto, Canada*

[7]*Institute of Medical Biophysics, University of Toronto, Canada*

pmra@imm.dtu.dk, schmah@cs.toronto.edu, stoffer@drcmr.dk, torbenelund@me.com, gyourganov, sstrother@rotman-baycrest.on.ca, lkh@imm.dtu.dk

Keywords: neuroimaging, classification, multivariate analysis, model interpretation, model visualization, sensitivity map, NPAIRS resampling, functional magnetic resonance imaging

Abstract: Classification models are becoming increasing popular tools in the analysis of neuroimaging data sets. Besides obtaining good prediction accuracy, a competing goal is to interpret how the classifier works. From a neuroscientific perspective, we are interested in the brain pattern reflecting the underlying neural encoding of an experiment defining multiple brain states. In this relation there is a great desire for the researcher to generate brain maps, that highlight brain locations of importance to the classifiers decisions. Based on sensitivity analysis, we develop further procedures for model visualization. Specifically we focus on the generation of summary maps of a nonlinear classifier, that reveal how the classifier works in different parts of the input domain. Each of the maps includes sign information, unlike earlier related methods. The sign information allows the researcher to assess in which direction the individual locations influence the classification. We illustrate the visualization procedure on a real data from a simple functional magnetic resonance imaging experiment.

1 INTRODUCTION

1.1 Background

Interest in applying multivariate analysis techniques to functional neuroimaging data is increasing, see e.g., (Lautrup et al., 1994; Mørch et al., 1997; Strother et al., 2002; Cox and Savoy, 2003; LaConte et al., 2005; O'Toole et al., 2007). A comprehensive introduction to classification methods in function magnetic resonance imaging (fMRI) is provided in (Pereira et al., 2009). Widely used classification schemes include kernel methods such as support vector machines (SVMs) (Cox and Savoy, 2003; Davatzikos et al., 2005; LaConte et al., 2005; Mourão Miranda et al., 2005). In kernel based learning, the input data is implicitly mapped into a high-dimensional feature space, and the classification model finds a linear decision boundary in the feature space. Typical kernel based learning methods are capable of constructing arbitrary nonlinear decision boundaries in the input space (i.e. the space of the measurements). For additional discussion of nonlinear classification in neuroimaging, see (Mørch et al., 1997; Cox and Savoy, 2003; LaConte et al., 2005; Haynes and Rees, 2006; Pereira et al., 2009; Misaki et al., 2010; Schmah et al., 2010; Rasmussen et al., 2011). While linear methods are of limited used when faced with nonlinear data, nonlinear methods may require unavailable large samples to generalize well (Mørch et al., 1997). Another limitation that has hampered the application of nonlinear kernel methods is the lack of well es-

tablished simple deterministic visualization schemes (LaConte et al., 2005).

The aim of our present work is to develop further procedures for interpretation/visualization of nonlinear classification models.

1.2 Related work

The sensitivity analysis that we investigate for model interpretation is based on early work in (Zurada et al., 1994; Zurada et al., 1997). (Kjems et al., 2002) have developed probabilistic sensitivity maps as a generic technique that can be applied to any model in order to visualize the features with respect to their importance in classification. This sensitivity map technique has been used for extraction of a single *global* summary map of the features importances to a trained classifier. The procedure has been applied within the field of neuroimaging in linear discriminant analysis (Kjems et al., 2002), quadratic discriminant analysis (Yourganov et al., 2010), and nonlinear kernel models (Rasmussen et al., 2011), and in skin cancer detection by Raman spectroscopy using neural networks (Sigurdsson et al., 2004).

For model interpretation (Baehrens et al., 2010) recently proposed a general methodology for interpretation of classifiers by exploring "local explanation vectors" that are defined as class probability gradients. This procedure identifies features that are important for prediction at localized points in the data space. (Golland et al., 2005) proposed a similar localized interpretation approach for support vector machine (SVMs) in the context of analysis of differences in anatomical shape between populations. They aimed for a representation of the differences between two classes captured by the classifier in the neighborhood of data examples. These two procedures give a localized visualization of the classifier, since they provide measures (here images) of each feature's importance at particular points of interest in the data space.

1.3 Contribution of this work

In the present work we focus on extending the sensitivity map approach of (Kjems et al., 2002) to also allow for extraction of summary maps with sign information. Originally, this method of building sensitivity maps did not contain sign information: the voxel's weight reflected the the relative importance of a particular voxel to the classifiers decisions. However, it is relevant to also consider maps containing sign information, which indicates whether a voxel's signal should be increased or decreased in order to increase the probability of a particular data observation being assigned to a specific class. The approach of (Kjems et al., 2002) did not contain sign information due to the difficulty of cancellation of opposite signs in an average. In our approach, the cancellation problem is mitigated to some extent by an unsupervised clustering step, by which we derive maps as weighted averages.

We aim for the characterization of trained classifiers not by a single global summary map, but rather a series of localized summary maps containing sign information. The maps proposed in (Baehrens et al., 2010; Golland et al., 2005) contain sign information, and in principle these approaches provide one map per data observation. In the present study we were interested in deriving maps containing sign information, and also to extract only relatively few representative maps in order to maintain simplicity of the model's visualization. This has potential to enhance the interpretation of how trained classifiers function.

Using the NPAIRS resampling framework (Strother et al., 2002) to assess the reliability/stability of the extracted maps is a key element in the analysis of our proposed procedure.

2 THEORY

In this section we briefly outline the concepts of supervised learning, and review the basics of kernel Fisher's discriminant (KFD) analysis that we use for classification. This is followed by an introduction to the theoretical framework of the sensitivity map for model visualization. The novel contribution in this section is the discussion of four procedures for obtaining brain maps from a trained classifier, and in particular one of these procedures based on clustering.

2.1 Classification setup

We consider a multi-class problem, where we have a labeled data set $\mathcal{D} = \{\mathbf{x}_n, c_n\}_{n=1}^{N}$, where $\mathbf{x}$ is a P dimensional input vector while c is the corresponding class label that groups $\mathbf{x}$ into C disjoint classes.

Kernel based methods work implicitly in a feature space $\mathcal{F}$ that is related to the input space $\mathcal{X}$ by a mapping $\phi : \mathcal{X} \rightarrow \mathcal{F}$, where $\phi(\cdot)$ is a function that returns a feature vector $\phi(\mathbf{x})$ corresponding to an input point $\mathbf{x}$. Rather than working in the feature space kernel based methods work on a kernel function $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$ that returns inner products in feature space. Examples of kernels are the linear kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$ and the RBF kernel

$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/q)$. For a further introduction to kernel based learning we refer to the literature, e.g. (Shawe-Taylor and Cristianini, 2004).

KFD analysis is a supervised dimensionality reduction technique (Mika et al., 1999), and a nonlinear generalization of Fisher's discriminant analysis. KFD seeks to find optimal projection directions along which the ratio of the between-class scatter to the total scatter is maximized. In the multi-class classification problem the Fisher's discriminant is given by the matrix $\mathbf{A}$, a $C-1$ column matrix, that optimizes the objective function

$$L(\mathbf{A}) = \frac{|\mathbf{A}^\top \mathbf{S}_B \mathbf{A}|}{|\mathbf{A}^\top (\mathbf{S}_T + \lambda \mathbf{I})\mathbf{A}|} \tag{1}$$

where $\mathbf{S}_B = \sum_{c=1}^{C} N_c (\mathbf{m}_c - \mathbf{m})(\mathbf{m}_c - \mathbf{m})^\top$ is the between-class scatter matrix, and $\mathbf{S}_T = \sum_{i=1}^{N} (\phi(\mathbf{x}_i) - \mathbf{m})(\phi(\mathbf{x}_i) - \mathbf{m})^\top$ is the total scatter matrix, with N_c denoting the number of samples in class c and $\mathbf{m}_c$ and $\mathbf{m}$ class means and grand mean respectively. Note that we here consider regularized Fisher's discriminant analysis, where λ is a regularization parameter (Friedman, 1989). There exist several ways to solve the above optimization problem. One is to consider the following generalized eigenvalue problem (Zhang et al., 2009b)

$$\mathbf{S}_B \mathbf{A} = (\mathbf{S}_T + \lambda \mathbf{I})\mathbf{A}\mathbf{D}, \tag{2}$$

where $\mathbf{A}$ holds the eigenvectors in the columns and $\mathbf{D}$ holds the corresponding eigenvalues in the diagonal. Since the eigenvectors can be expressed as $\mathbf{A} = \sum_{n=1}^{N} (\phi(\mathbf{x}_i) - \mathbf{m})\mathbf{b}_i^\top$, where $\mathbf{b}_i$ is a coefficient vector with $C-1$ rows, we can reformulate eq. (2) in terms of the kernel matrix

$$\mathbf{CWCB} = (\mathbf{CC} + \lambda \mathbf{C})\mathbf{BD}, \tag{3}$$

where $\mathbf{C} = \mathbf{HKH}$ is the centered kernel matrix $\mathbf{K}$, with the centering matrix defined as $\mathbf{H} = \mathbf{I}_N - \mathbf{1}_N \mathbf{1}_N'/N$, and $\mathbf{W}$ is an $(N \times N)$ positive symmetric weight matrix with elements $W_{i,j} = 1/N_c$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ belongs to the same class and $W_{i,j} = 0$ otherwise. Based on the solution to eq. (3) we can obtain the projection of a feature space vector $\phi(\mathbf{x})$ onto $\mathbf{A}$ as

$$\mathbf{z}_\mathbf{x} = \mathbf{A}^\top (\phi(\mathbf{x}) - \mathbf{m}) = \mathbf{B}^\top \mathbf{H}(\mathbf{k}_\mathbf{x} - \frac{1}{N}\mathbf{K}\mathbf{1}_N), \tag{4}$$

where $\mathbf{k}_\mathbf{x} = (k(\mathbf{x}_1, \mathbf{x}), ..., k(\mathbf{x}_N, \mathbf{x}))^\top$, and $\mathbf{B}$ is an $(N \times (C-1))$ matrix.

On top of the projection of data points onto the subspace identified by KFD, we implement a simple classifier assuming a Gaussian noise model. The likelihood function is

$$p(\mathbf{z}_\mathbf{x}|\mu_c, \sigma^2) = (2\pi\sigma^2)^{-\frac{C-1}{2}} \exp(-\frac{1}{2\sigma^2}||\mathbf{z}_\mathbf{x} - \mu_c||^2), \tag{5}$$

with μ_c denoting the mean of projections of class c and variance σ^2 shared across classes. Note that we here use the notation $\mathbf{z}_\mathbf{x}$ to emphasize that $\mathbf{z}$ is a vector valued function of $\mathbf{x}$. Classification is then performed according to Bayes' rule

$$p(c|\mathbf{z}_\mathbf{x}) = \frac{p(\mathbf{z}_\mathbf{x}|c)p(c)}{\sum_{c'=1}^{C} p(\mathbf{z}_\mathbf{x}|c')p(c')}. \tag{6}$$

In the following we will also refer to $p(c|\mathbf{z}_\mathbf{x})$ as the classifier's c'th *output channel*.

2.2 Model visualization

2.2.1 General definition

Sensitivity analysis is a simple measure of the relative importance of the different input features (voxels, in the present context) to the classifier. We follow the approach on (Kjems et al., 2002; Strother et al., 2002) and aim for a visualization of the relative importance of the input data for a given function $f(\mathbf{x})$ in a stochastic environment with a distribution over the inputs given by the probability density function $p(\mathbf{x})$. We define the signed sensitivity map by

$$\mathbf{s} = \int_A \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} p(\mathbf{x}) d\mathbf{x}, \tag{7}$$

where $\mathbf{s}$ is a P dimensional vector, where the j'th element holds the sensitivity measure corresponding to the j'th voxel. Here A denotes the region of integration (some region of the image space). Hence, the map summarizes the gradient field of the function $f(\mathbf{x})$ in some region A.

If $f(\cdot)$ is a nonlinear function, cancellation may occur since some regions of the input space can have positive sensitivity while other regions can have negative sensitivity. To avoid such cancellation we also consider a map based on the squared sensitivities

$$\mathbf{s} = \int_A \left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \right)^2 p(\mathbf{x}) d\mathbf{x}. \tag{8}$$

In practice the maps are approximated as a finite sum over observations, for example eq. (7) is approximated by

$$\mathbf{s} = \frac{1}{N_I} \sum_{n \in I} \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}|_{\mathbf{x}=\mathbf{x}_n}, \tag{9}$$

where the sum is calculated based on a set I containing N_I observations.

2.2.2 Choice of visualization function

Different choices for the visualization function $f(\mathbf{x})$ exist.

One choice is to use the posterior probability in eq. (6)

$$f(\mathbf{z}_\mathbf{x}) = p(c|\mathbf{z}_\mathbf{x}), \quad (10)$$

and this is the approach in (Baehrens et al., 2010).

A variant hereof is

$$f(\mathbf{z}_\mathbf{x}) = \log(p(c|\mathbf{z}_\mathbf{x})), \quad (11)$$

that is the approach of (Kjems et al., 2002). In the present paper we follow this approach and use eq. (11) as a visualization function.

Another possibility is to use the decision function(s) as a visualization function. This is the approach in (Yourganov et al., 2010; Rasmussen et al., 2011). For example, if we build a nearest mean classifier on top of the subspace identified by KDF we have C output channels where the c'th channel is given by

$$f(\mathbf{z}_\mathbf{x}) = ||\mathbf{z}_\mathbf{x} - \mu_c||^2. \quad (12)$$

2.2.3 Choice of summation region and output channel(s)

The set I over which the summation in eq. (9) is done and one or more output channels must be selected in order to construct the map. Many possible variations exist, and in the following we will outline just a few.

Procedure I
Let all observations be members of I. Furthermore we consider all output channels, and hence obtain a grand average map as

$$\mathbf{s}^{ga} = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_I}\sum_{n\in I}\left(\frac{\partial \log(p(c|\mathbf{z}_\mathbf{x}))}{\partial \mathbf{x}}|_{\mathbf{x}=\mathbf{x}_n}\right)^2, \quad (13)$$

where we square the single sensitivities to avoid potential cancellations.

Procedure II
Let all observations in a class c be members of I_c and let only members of I_c contribute to the sum over output channel c. We obtain a grand average map as

$$\mathbf{s}^{ga} = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_{I_c}}\sum_{n\in I_c}\left(\frac{\partial \log(p(c|\mathbf{z}_\mathbf{x}))}{\partial \mathbf{x}}|_{\mathbf{x}=\mathbf{x}_n}\right)^2, \quad (14)$$

which is the procedure in (Kjems et al., 2002).

Procedure III
Here we are interested in a map reflecting an inter-class contrast. Let all observations in a class c' be members of $I_{c'}$ and consider an output channel c. We obtain a map as

$$\mathbf{s}^{c|c'} = \frac{1}{N_{I_{c'}}}\sum_{n\in I_{c'}}\frac{\partial \log(p(c|\mathbf{z}_\mathbf{x}))}{\partial \mathbf{x}}|_{\mathbf{x}=\mathbf{x}_n}, \quad (15)$$

where we use the notation $\mathbf{s}^{c|c'}$ to indicate that we consider the sensitivity with respect to output channel c of the classifier, and the summation is performed over the members class c'. This map describes how observations in an input space region characterized by the members in class c' should be changed in general in order to increase the posterior probability of class c. Note that this procedure can be problematic if the derivatives have different sign across the members in c'. One way to deal with this issue is to use the definition eq. (8) to obtain a sensitivity map with no sign information.

Procedure IV
If there exists a considerable heterogeneity in the sign of the gradients within class c', as discussed above, we propose the following as an alternative to using the squared gradients in order to derive signed sensitivity maps: We consider specific classes c and c', as in the Procedure III, with $I_{c'}$ comprised of all observations in c'. We then perform a soft clustering of all pairs $(x_n, s(x_n))$ for $n \in I_{c'}$, where $s(x_n)$ equals $\partial \log p(c|z_x)/\partial x|_{x=x_n}$. Hence, we are interested in patterns based on observations that are similar both with respect to their input space location and their sensitivity measure. The soft clustering takes the form of a Gaussian mixture model (GMM) with the number of clusters optimized by cross-validation. Finally, we construct a sensitivity map for each cluster, by weighting the sensitivities $s(x_n)$ by weights w_n^k defined to be the posterior probability of $(x_n, s(x_n))$ being in the given cluster k

$$\mathbf{s}^{c|c',k} = \frac{1}{N_{I_{c'}}}\sum_{n\in I_{c'}} w_n^k \frac{\partial \log(p(c|\mathbf{z}_\mathbf{x}))}{\partial \mathbf{x}}|_{\mathbf{x}=\mathbf{x}_n}. \quad (16)$$

If we assume that $p(c|z_x)$ is smooth, then taking a large enough number of clusters will result in the vectors $s(x)$ within each cluster being similar, which will reduce the sign cancellation problem and also obscure less structure.

Specifically we estimate the weights in eq. (16) as follows: (1) For each of the observations $\mathbf{x}$ we construct a feature vector $\mathbf{f}$ by stacking the corresponding single sensitivity $\mathbf{s}$ and the observation $\mathbf{x}$ itself $\mathbf{f} = [\mathbf{x}; \mathbf{s}]$ ($\mathbf{x}$ and $\mathbf{s}$ were both scaled to unit norm in order to put them on the somewhat same scale), so that $\mathbf{f}$ is a $2P$-dimensional vector. (2) We perform principal component analysis (PCA) and project the feature vectors $\mathbf{f}$ onto a PCA subspace. (3) Based on the low dimensional feature representation, we build a Gaussian mixture model (GMM). To estimate the number of components/clusters k we use cross validation, where we estimate the generalization error of the GMM by evaluating the likelihood function on the left out fold.

3 MATERIALS & METHODS

This section provides a description of the fMRI data set used for illustration. This is followed by a description of our classification setup and the resampling scheme used for model evaluation.

3.1 Functional MRI data set - visual paradigm

Six healthy subjects were enrolled after informed consent as approved by the local Ethics Committee. The fMRI data set was acquired on a 3T (Siemens Magnetom Trio) scanner using a 8-channel head coil (Invivo, Florida, USA). The participants were subjected to four visual conditions presented on a screen with the following sequence: (1) no visual stimulation (no), (2) reversing checkerboard on the left half of the screen (left), (3) reversing checkerboard on the right half of the screen (right), (4) reversing checkerboard on both halfs of the screen (both). Each stimulus condition was presented for 15 seconds followed by 5.04 seconds of rest with no visual stimulation. The stimulation sequence was repeated 12 times in the experimental run, and 576 scan volumes were acquired in total.

Pre-processing of the fMRI time series was conducted using the SPM8 software package (http://www.fil.ion.ucl.ac.uk/spm) and comprised the following steps for each subject: (1) Rigid body realignment of the EPI images to the mean volume in the time series. (2) Spatial normalization of the mean EPI image to the EPI template in SPM8. (3) The estimated warp field was applied to the individual EPI images. The normalized images were written with 3 mm isotropic voxels. (4) For visualization purposes the anatomical scan was spatially normalized to the T1 template in SPM8, using the same settings as for the EPI images. (5) The EPI images were smoothed with an isotropic 8 mm full-width half-maximum Gaussian kernel. (6) The data were masked with a rough whole-brain mask (75257 voxels). (7) To remove low frequency components from the time series, a set of discrete cosine basis functions up to a cut-off period of 128 seconds were projected out of the data. (8) Within each subject the individual voxel time series were standardized (each voxel subtracted the mean and scaled by the standard deviation of the voxel's time series). Further details on the data acquisition and pre-processing are provided in (Rasmussen et al., 2011).

Figure 1 shows the average fMRI images for each of the four experimental conditions, and is shown in order to assist the reader in interpreting the subsequent figures. Negative signal occurs due to the standardization of data within each subject. The maps are thresholded to show the upper 10 percentile of the voxel absolute value distribution (as in all subsequent maps shown).

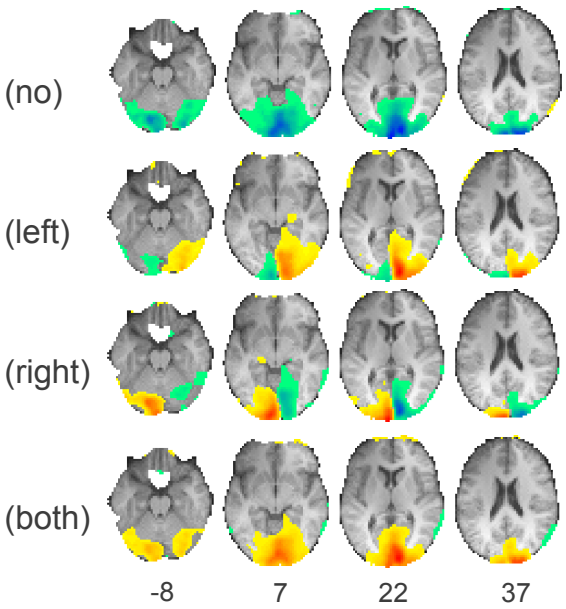

Figure 1: Average EPI images across all six subjects for the four conditions in the fMRI data set. The data was standardized within each subject. Warm and cold colors are positive and negative values respectively. The maps are thresholded to show the upper 10 percentile of the voxel absolute value distribution, projected onto an average structural scan of the six subjects included in the analysis. Numbers under the slices denote z coordinates in MNI space. Slices are displayed according to neurological convention (right side of a brain slice is the right side of the brain).

3.2 Classification setup and model evaluation

In the analysis of the fMRI data set, we considered subjects as the basic resampling unit, where the classifier was trained on data from a subset of subjects, while the target labels were inferred for scans from subjects in the out-of-sample subset. Scans 7-11 in each epoch were used in the predictive modeling, and the remaining volumes were discarded to avoid contaminating effects of the BOLD signal. We aimed for a "whole-brain" and single block classification with temporal averaging of scans within the same block. No feature selection prior to the application of the classification models was performed. Two binary classification task were formulated:

- Classification task I: We considered a four class

classification task, where scans from the four conditions {no, left, right, both} were assigned to the classes {1,2,3,4}. We expect observations belonging to the same class to be relatively homogeneous in this classification task.

- Classification task II: Scans from condition (no) and (both) were assigned to class 1, while scans from condition (left) and (right) were assigned to class 2. By this labeling we intended to introduce an artificial coupling between brain regions, equivalent to computer science's *xor* function. In this classification task we expect a relatively large degree of heterogeneity between observations belonging to that same class.

For classification we used KFD analysis with the RBF kernel. The model estimation requires selection of the regularization parameter λ in KFD and selection of the q parameter that controls the kernel width of the RBF kernel. We let λ range over the interval $2^{-15} \ldots 2^{30}$ relative to the average non-zero eigenvalue of the data covariance matrix, while q was varied in the range $2^{-5} \ldots 2^{16}$ relative to the average input-space distance to the nearest 25% points across all training examples.

For model evaluation we used the NPAIRS resampling scheme (Strother et al., 2002). In this cross-validation framework the data were split into two partitions of equal size (three subjects in each partition). The model was trained on the first split and the prediction accuracy was estimated from the second split and vice versa, yielding two estimates of the prediction accuracy. We measure the prediction accuracy by the posterior probability for the correct class label. These prediction accuracies were averaged and considered as the prediction metric (p) of the NPAIRS scheme. In addition we extracted a grand average sensitivity map according to eq. (13) for each of the two classification models. The Pearson's correlation coefficient between spatial maps derived from the two models was calculated as a spatial reproducibility metric (r). Each map vector was scaled to unit norm, and the scatter plot of the maps from each model was projected onto a signal axis and an uncorrelated noise axis as described in (Strother et al., 2002). The projection onto the signal axis was scaled by the standard deviation of the noise projection, which gave a reproducible statistical parametric image (rSPI). This procedure was repeated for all possible splits of the subjects (10 resampling runs).

We use the p and r metrics for model optimization, where we choose model parameters that maximize both metrics jointly, in terms of the Euclidean distance to the point (p,r)=(1,1) over the entire space of cross-validated results (Strother et al., 2002; Zhang et al., 2009a).

In order to interpret the classifier based on brain maps containing sign information we also derived maps according to Procedure III in section (2.2).

Furthermore, for classification task II we derived maps according to Procedure IV in Section (2.2), since we expected a relatively large heterogeneity between single sensitivities of observations within the same class. We performed Steps 1-3 of Procedure IV (Section (2.2)) within the NPAIRS resampling scheme. For the PCA subspace in Step 2, we used the space spanned by the first two PCs (this was a heuristic choice). The most likely number of clusters $k \in [1,...,6]$ in the GMM was found by maximizing the mean likelihood over the 10 resampling splits. We then performed a second pass through the NPAIRS procedure, where we fixed k across all splits and resampling runs in order to obtain the same number of clusters across models. For each split we fitted a GMM and derived sensitivity maps according to eq. (16). In order to derive rSPIs the clusters labels must be aligned across splits. We used a simple reference filtering procedure, where the cluster's labels of a particular split was permuted in order to maximize the correlation between sensitivity maps across splits.

4 RESULTS AND DISCUSSION

Figure 2 shows average rSPIs derived from the classification models in classification task I (four classes). The average prediction accuracy was 0.9229 in terms of average posterior probability of the correct class. Figure 2(A) shows the average rSPI across NPAIRS splits based on the grand average maps $\mathbf{s}^{ga}$ derived according to Procedure I in Section 2.2. The map identifies that voxels in the visual cortex contribute the most with relevant information to the classifiers. The average reproducibility of the map across splits was 0.81. Note that the rSPI only contains positive values since it is based on squared sensitivities. Figure 2(B) shows average rSPIs that are based on the interclass contrast (Procedure III in Section 2.2). According to the map $\mathbf{s}^{left|no}$ a signal increase primarily in the right visual cortex will increase the posterior probability of the scans in class (no) being classified as belonging to class (left). Figure 1 indicates that this is reasonable, since the condition (left) is characterized by a larger signal in the right visual cortex relative to the (no) condition. Likewise, the map $\mathbf{s}^{left|right}$ indicates that lowering the signal in the left visual cortex and increasing signal in the right visual cortex will increase the posterior probability of the scans in class (right) being classified as (left). The reproducibility

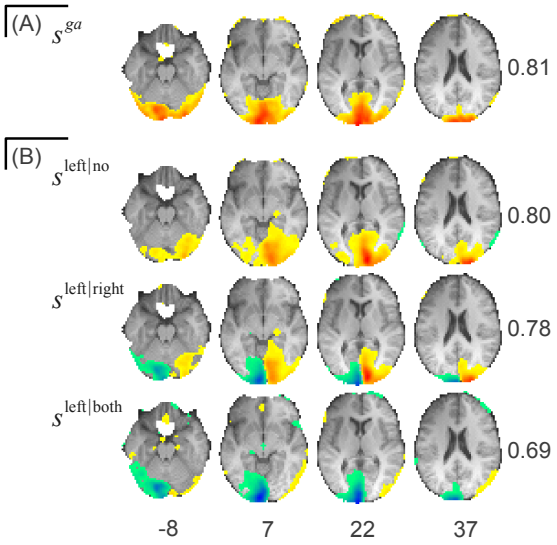

Figure 2: Classification task I with a four class classification task. Different brain maps were extracted from the trained classifier. Panel (A) shows an average rSPI based on the grand average map (Procedure I in Section 2.2). Panel (B) shows signed rSPIs based on Procedure III in Section 2.2. The notation e.g. $\mathbf{s}^{left|no}$ means that the map indicates how scans in class 'no' should be changed in order to increase the posterior probability of class 'left'. The numbers right to the brain slices denote the reproducibility estimated within the NPAIRS resampling framework. The rSPIs are thresholded to show the upper 10 percentile of the distribution of the absolute voxel values, projected onto an average structural scan of the six subjects included in the analysis. Numbers under the slices denote z coordinates in MNI space. Slices are displayed according to neurological convention.

of the maps in Figure 2 are fairly high ranging from 0.69-0.80.

Figure 3(C-E) shows average rSPIs derived from the classification models in classification task II (xor-problem). The average prediction accuracy was 0.905. Figure 3(C) shows the average rSPI across NPAIRS splits based on the grand average maps $\mathbf{s}^{ga}$ derived according to Procedure I in Section 2.2. The maps identify that voxels in the visual cortex that contribute the most relevant information to the classifiers. The average reproducibility of the maps across splits was 0.81. Note that Figure 3(C) resembles Figure 2(A), which means that the regions identified as being relevant to classification in tasks I and II are similar. Figure 3(D) shows average rSPIs based on interclass contrasts (Procedure III in Section 2.2). Due to the potential heterogeneity within the classes, the sensitivity maps were based on squared single sensitivities. Hence, no sign information is present in the maps.

For comparison, we derived signed class specific sensitivities directly as in classification task I Figure 2(B) (without squaring), and found that the reproducibilities of the maps were reduced to 0.36. This decrease in reproducibility is presumably due to cancellation effects.

Figure 3(E) shows the rSPIs obtained from Procedure IV in Section 2.2. The first and second rows of Figure 3(E) can be seen as refinements of the first row of Figure 3(D), obtained by dividing the class $\{left, right\}$ into two clusters. Similarly, the third and fourth rows of Figure 3(E) are refinements of the second row of Figure 3(E).

The clustering process is illustrated in Figure 3(A-B). Figure 3(B) shows an example of clustering of observations within class $\{left, right\}$. In both class $\{no, both\}$ and class $\{left, right\}$ we found evidence of two clusters based on generalization error measured as the negative likelihood of the Gaussian mixture model (Figure 3(A)). Based on the cluster analysis we derive two maps for each of the two classes according to Procedure IV in Section 2.2. By this procedure we obtained maps with sign information. Figure 3(E) shows the resulting four rSPI (two for each class). For example $\mathbf{s}^{\{no,both\}|\{left,right\},cluster1}$ denotes that the sensitivity map is based on the output class $\{no, both\}$, and the sum in eq. (16) is calculated over the members of class $\{left, right\}$, where the weighting factor w_n^k is based on the posterior probability for observations belonging to cluster 1. The reproducibility values of the sensitivity maps are moderate, ranging from 0.53-0.61. Note that these reproducibility values are lower than for Procedure III based on squared single sensitivities (see Figure 3(D)), but higher than for Procedure III with signed sensitivities (mean reproducibility 0.36, as stated earlier). To further interpret the maps in Figure 3(E) we calculated the average weight factor w_n^k in eq. (16) for each of the sub clusters. For the map $\mathbf{s}^{\{no,both\}|\{left,right\},cluster1}$ we found that the members of condition (left) had an average weighting factor of 0.0038 in the sum, while the members of the condition (right) had an average weighting factor of 0.9857. Hence, members of the condition (right) contribute the most to this map. Likewise, the members of condition (left) contributed the most to the map $\mathbf{s}^{\{no,both\}|\{left,right\},cluster2}$. For the map $\mathbf{s}^{\{left,right\}|\{no,both\},cluster1}$ we found that the members of condition (no) had an average weighting factor of 0.0000 in the sum, while the members of the condition (both) had an average weighting factor of 0.9986. Hence, members of the condition (both) contribute the most to this map. Likewise, the members of condition (no) contribute the most to the map $\mathbf{s}^{\{left,right\}|\{no,both\},cluster2}$.

Note the similarity of the maps in the first row of Figure 3(E) ($\mathbf{s}^{\{no,both\}|\{left,right\},cluster1}$, where cluster

1 corresponds to class *right*, as just explained) and the second row of Figure 2(B) ($\mathbf{s}^{left|right}$). Also the fourth row of Figure 3(E) ($\mathbf{s}^{\{left,right\}|\{no,both\},cluster2}$, where cluster 2 corresponds to class *no*) is similar to the first row of Figure 2(B) ($\mathbf{s}^{left|both}$); and the third row of Figure 3(E) ($\mathbf{s}^{\{left,right\}|\{no,both\},cluster1}$, where cluster 1 corresponds to class *both*) is similar to the third row of Figure 2(C) ($\mathbf{s}^{left|both}$). The significance of these observations is that Procedure IV has succeeded in providing similar information when applied to Classification Task II (the 2-class xor-problem) as the information given by Procedure III applied to Classification Task I (the 4-class classification), even though the available class labels in Task II are less informative and the two classes *no, both* and *left, right* are heterogeneous. This suggests that Procedure IV can provide useful information about the nature of nonlinear classifiers when applied to complex, heterogeneous classes.

5 CONCLUSION

The established probabilistic sensitivity map procedure provides a global summary map of the relative importance of voxels to a trained classifier (Kjems et al., 2002). However, no sign information is present in such a map. In the present work we have proposed a procedure to allow for generation of summary maps with sign information. Furthermore, we have proposed a clustering procedure that is applicable in cases where relatively large heterogeneity between observation exists which may degrade the performance of the model visualization due to cancellation effects.

As a proof of concept, we have illustrated the approach on a data set from a simple fMRI experiment, with classes deliberately defined to be heterogeneous. Our procedure successfully recovered known structure in the classes. We also found that the maps produced for this data set are robust, in the sense that they are reproducible as judged by the NPAIRS resampling framework. We showed that reproducibility is improved by the new clustering procedure.

Our results suggest that our new method of model visualization may be useful in visualizing nonlinear classifiers trained on heterogeneous classes. Further work is needed to compare variations of the method, in particular different possible choices of the visualization function (see Section 2.2.2), and to validate the method on a larger variety of real or synthetic data.

ACKNOWLEDGEMENTS

This work is partly supported by the Danish Lundbeck Foundation through the program www.cimbi.org. The Simon Spies Foundation is acknowledged for donation of the Siemens Trio scanner. Kristoffer H. Madsen was supported by the Danish Medical Research Council (grant no. 09-072163) and the Lundbeck Foundation (grant no. R48-A4846).

REFERENCES

Baehrens, D., Schroeter, T., Harmeling, S., Kawanab, M., Hansen, K., and Müller, K.-R. (2010). How to explain individual classification decisions. *Journal of Machine Learning Research*, 11:1803–1831.

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

Davatzikos, C., Ruparel, K., Fan, Y., Shen, D., Acharyya, M., Loughead, J., Gur, R., and Langleben, D. (2005). Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. *NeuroImage*, 28(3):663–668.

Friedman, J. H. (1989). Regularized discriminant analysis. *J. Am. Statistical Assoc.*, 84:165 – 175.

Golland, P., Grimson, W. E. L., Shenton, M. E., and Kikinis, R. (2005). Detection and analysis of statistical differences in anatomical shape. *Medical Image Analysis*, 9:69–86.

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

Kjems, U., Hansen, L. K., Anderson, J., Frutiger, S., Muley, S., Sidtis, J., Rottenberg, D., and Strother, S. C. (2002). The quantitative evaluation of functional neuroimaging experiments: mutual information learning curves. *NeuroImage*, 15(4):772–786.

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

Lautrup, B., Hansen, L., Law, I., Mørch, N., Svarer, C., and Strother, S. (1994). Massive weight sharing: A cure for extremely ill-posed problems. *Proceedings of the Workshop on Supercomputing in Brain Research: From Tomography to Neural Networks. World Scientific, Ulich, Germany*, pages 137–148.

Mika, S., Rätsch, G., Schölkopf, B., Smola, A., Weston, J., and Müller, K.-R. (1999). Invariant feature extraction and classification in kernel spaces. *Advances in Neural Information Processing Systems*, 12:526–532.

Misaki, M., Kim, Y., Bandettini, P., and Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage*, 53(1):103–118.

Mørch, N., Hansen, L. K., Strother, S. C., Svarer, C., Rottenberg, D. A., Lautrup, B., Savoy, R., and Paulson, O. B. (1997). Nonlinear versus Linear Models in Functional Neuroimaging: Learning Curves and Generalization Crossover. *IPMI '97: Proceedings of the 15th International Conference on Information Processing in Medical Imaging*, pages 259–270.

Mourão Miranda, J., Bokde, A., Born, C., Hampel, H., and M (2005). Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data. *NeuroImage*, 28:980 – 995.

O'Toole, A. J., Jiang, F., Abdi, H., P Nard, N., Dunlop, J. P., and Parent, M. A. (2007). Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *Journal of Cognitive Neuroscience*, 19:1735–1752.

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

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.

Schmah, T., Yourganov, G., Zemel, R., Hinton, G., Small, S., and Strother, S. (2010). A Comparison of Classification Methods for Longitudinal fMRI Studies. *Neural Computation*, 22:2729–2762.

Shawe-Taylor, J. and Cristianini, N. (2004). Kernel methods for pattern analysis. Cambridge.

Sigurdsson, S., Philipsen, P., Hansen, L., Larsen, J., Gniadecka, M., and Wulf, H. (2004). Detection of skin cancer by classification of Raman spectra. *IEEE Transactions on Biomedical Engineering*, 51(10):1784–1793.

Strother, S., Anderson, J., Hansen, L., 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.

Yourganov, G., Schmah, T., Small, S., Rasmussen, P., and Strother, S. (2010). Functional connectivity metrics during stroke recovery. *Arch Ital Biol.*, 148(3):259–270.

Zhang, J., Anderson, J. R., Liang, L., Pulapura, S. K., Gatewood, L., Rottenberg, D. A., and Strother, S. C. (2009a). Evaluation and optimization of fMRI single-subject processing pipelines with NPAIRS and second-level CVA. *Magn Reson Imaging*, 27:264–278.

Zhang, Z., Dai, G., and Jordan, M. I. (2009b). A flexible and efficient algorithm for regularized fisher discriminant analysis. In *ECML PKDD '09: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 632–647, Berlin, Heidelberg. Springer-Verlag.

Zurada, J., Malinowski, A., and Cloete, I. (1994). Sensitivity analysis for minimization of input data dimension forfeedforward neural network. *1994 IEEE International Symposium on Circuits and Systems, 1994. ISCAS'94.*, 6:447–450.

Zurada, J., Malinowski, A., and Usui, S. (1997). Perturbation method for deleting redundant inputs of perceptron networks. *Neurocomputing*, 14(2):177–193.

APPENDIX

In the following we show how to calculate the derivative used in eq. (7,8,9). First we calculate the derivative in eq. (4)

$$\frac{\partial \mathbf{z_x}}{\partial \mathbf{x}} = \frac{\partial \mathbf{k_x}}{\partial \mathbf{x}} \mathbf{HB}. \tag{17}$$

We then calculate the derivative of the visualization function

$$\begin{aligned}
&\frac{\partial \log(p(c|\mathbf{z_x}))}{\partial \mathbf{x}} = \\
&\frac{\partial \log(p(\mathbf{z_x}|c)P(c))}{\partial \mathbf{x}} - \frac{\partial \log(\sum_{c'=1}^{C} p(\mathbf{z_x}|c')P(c'))}{\partial \mathbf{x}} = \\
&-\frac{\partial ||\mathbf{z_x} - \mu_c||^2}{\partial \mathbf{x}} + \sum_{c'=1}^{C} p(c'|\mathbf{z_x}) \frac{\partial ||\mathbf{z_x} - \mu_{c'}||^2}{\partial \mathbf{x}} = \\
&-\frac{\partial \mathbf{z_x}}{\partial \mathbf{x}} (\mathbf{z_x} - \mu_c) + \frac{\partial \mathbf{z_x}}{\partial \mathbf{x}} \sum_{c'=1}^{C} p(c'|\mathbf{z_x})(\mathbf{z_x} - \mu_{c'}) = \\
&-\frac{\partial \mathbf{z_x}}{\partial \mathbf{x}} \left((\mathbf{z_x} - \mu_c) - \sum_{c'=1}^{C} p(c'|\mathbf{z_x})(\mathbf{z_x} - \mu_{c'}) \right) = \\
&-\frac{\partial \mathbf{k_x}}{\partial \mathbf{x}} \mathbf{HB} \left((\mathbf{z_x} - \mu_c) - \sum_{c'=1}^{C} p(c'|\mathbf{z_x})(\mathbf{z_x} - \mu_{c'}) \right).
\end{aligned} \tag{18}$$

For example, for the linear kernel we have

$$\frac{\partial \mathbf{k_x}}{\partial \mathbf{x}} = \mathbf{X}, \tag{19}$$

where $\mathbf{X}$ holds training observations in the columns. For the RBF kernel we have

$$\frac{\partial \mathbf{k_x}}{\partial \mathbf{x}} = \mathbf{M} \mathbf{D_k}, \tag{20}$$

where $\mathbf{M}$ in a $(P \times N)$ matrix that holds the elements $M_{i,j} = x_i^j - x_i$, with x_i^j referring to the i'th element in the j'th training example. $\mathbf{D_k}$ is an $(N \times N)$ diagonal matrix with the elements of $\mathbf{k_x}$ in the diagonal.

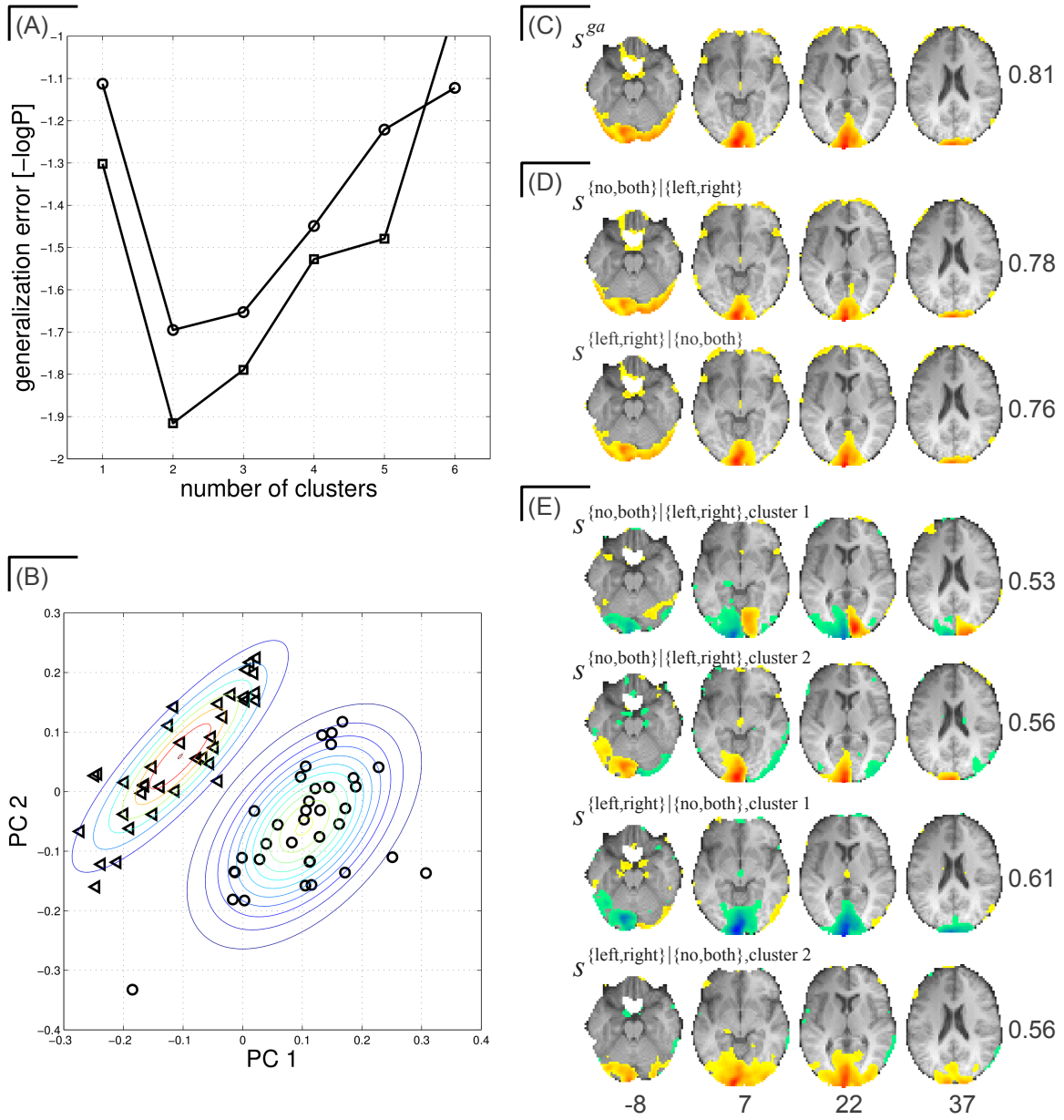

Figure 3: Classification task II. **Panel A-B:** Structure of single sensitivities. See Procedure IV in Section 2.2 for further details. Modeling of the data distribution was based on a Gaussian mixture model (GMM). **Panel A** shows the generalization error as a function of the number of components (averaged over 10 resampling runs) in the GMM. The generalization error was calculated by evaluation of the likelihood function of the GMM based on four-fold cross validation performed as a nested loop within each NPAIRS split half. The circles corresponds to the class $\mathbf{s}^{\{no,both\}|\{left,right\}}$, and the squares corresponds to the $\mathbf{s}^{\{left,right\}|\{no,both\}}$ single sensitivity measures. In both classes we observe evidence for two clusters/components. **Panel B** shows contours of the probability densities of a GMM fitted to $\mathbf{s}^{\{no,both\}|\{left,right\}}$ sensitivities of a single split of the data. The triangles corresponds to the observations in condition (left) and the circles corresponds to observations in the condition (right). Note that this labeling was not "visible" to the modeling procedure. **Panel C-E:** Interpretation of the trained classifier. Different brain maps were extracted from the classifier. **Panel (C)** shows an average rSPI based on the grand average sensitivity map (Procedure I in Section 2.2). **Panel (D)** shows rSPIs based on single class sensitivities. The notation eg. $\mathbf{s}^{\{no,both\}|\{left,right\}}$ means that the map indicates how scans in class $\{left, right\}$ should be changed in order to increase the posterior probability of class $\{no, both\}$. Note that the maps does not contain sign information (Procedure III in Section 2.2). **Panel (E)** shows rSPIs based on single class sensitivity, where each class is characterized by two sub-clusters (Procedure IV in Section 2.2). The numbers right to the brain slices denote the reproducibility estimated within the NPAIRS resampling framework. The rSPIs are thresholded to show the upper 10 percentile of the distribution of the absolute voxel values, projected onto an average structural scan of the six subjects included in the analysis. Numbers under the slices denote z coordinates in MNI space. Slices are displayed according to neurological convention.