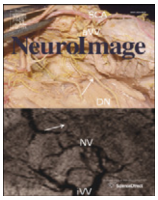
NeuroImage 55 (2011) 1120–1131

Contents lists available at ScienceDirect

NeuroImage

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



Visualization of nonlinear kernel models in neuroimaging by sensitivity maps

Peter Mondrup Rasmussen [a,b,c], Kristoffer Hougaard Madsen [a,c], Torben Ellegaard Lund [b], Lars Kai Hansen [a,*]

[a] DTU Informatics, Technical University of Denmark, Denmark
[b] The Danish National Research Foundation's Center for Functionally Integrative Neuroscience, Aarhus University Hospital, Denmark
[c] Danish Research Centre for Magnetic Resonance, Copenhagen University Hospital Hvidovre, Denmark

ARTICLE INFO

Article history:
Received 3 September 2010
Revised 26 November 2010
Accepted 2 December 2010
Available online 17 December 2010

Keywords:
Neuroimaging
Pattern analysis
Machine learning
Multivariate analysis
Kernel methods
Support vector machine
Nonlinear modeling
Model visualization
Sensitivity map

ABSTRACT

There is significant current interest in decoding mental states from neuroimages. In this context kernel methods, e.g., support vector machines (SVM) are frequently adopted to learn statistical relations between patterns of brain activation and experimental conditions. In this paper we focus on visualization of such nonlinear kernel models. Specifically, we investigate the sensitivity map as a technique for generation of global summary maps of kernel classification models. We illustrate the performance of the sensitivity map on functional magnetic resonance (fMRI) data based on visual stimuli. We show that the performance of linear models is reduced for certain scan labelings/categorizations in this data set, while the nonlinear models provide more flexibility. We show that the sensitivity map can be used to visualize nonlinear versions of kernel logistic regression, the kernel Fisher discriminant, and the SVM, and conclude that the sensitivity map is a versatile and computationally efficient tool for visualization of nonlinear kernel models in neuroimaging.

Introduction

Since the advent of functional neuroimaging, the dominant paradigm has been functional localization, i.e., the identification of brain regions with a significant regional change in activation with reference to a control condition, under sensory, motor, or cognitive experimental manipulation. With this aim analysis of functional neuroimaging data is conventionally implemented in terms of the mass-univariate general linear model (GLM), see e.g. Friston et al. (1995). Interest in applying multivariate analysis techniques to functional neuroimaging data has increased recently, 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), Hanson and Halchenko (2008). Multivariate brain image classification is also frequently referred to as mental state decoding (Haynes and Rees, 2006) or multivoxel pattern analysis (MVPA) (Norman et al., 2006). In this context different linear and nonlinear pattern recognition techniques have been used to learn the statistical relationship between patterns of "brain activation" and the experimental condition. While linear methods have in principle limitations, nonlinear methods may require too large samples to generalize well (Mørch et al., 1997), another limitation that has hampered the application of nonlinear kernel methods is the lack of a simple deterministic visualization scheme (Hanson et al., 2004; LaConte et al., 2005). In the present communication we offer such a generic model visualization technique applicable for nonlinear kernel based models.

Predictive modeling in neuroimaging

Pattern classification in functional neuroimaging was first proposed for PET imaging by Lautrup et al. (1994) and later for fMRI by Mørch et al. (1997), additional historical notes can be found in Hansen (2007). Predictive modeling is also a constituent part of the NPAIRS (non parametric prediction, activation, influence, and reproducibility resampling) framework proposed by Strother et al. (2002). In this framework the prediction accuracy together with pattern reproducibility provides empirical means for evaluation and optimization of data processing pipelines in neuroimaging. With the aim to link activation patterns to experimental manipulations, pattern classification has recently been applied to various types of fMRI experiments on visual (Cox and Savoy, 2003; Hanson et al., 2004; Mourão Miranda et al., 2007; Hanson and Halchenko, 2008; Sato et al., 2009), auditory (Formisano et al., 2008a), motor (LaConte et al., 2005; Sato et al., 2009), and cognitive tasks (Mourão Miranda et al., 2008). Other applications include lie detection (Davatzikos et al., 2005), real-time brain state classification (LaConte et al., 2007), and prediction of consumer behavior (Grosenick et al., 2008). Examples of pattern

[*] Corresponding author. DTU Informatics, Technical University of Denmark, Richard Petersens Plads, Kongens Lyngby, Denmark.
E-mail addresses: pmra@imm.dtu.dk (P.M. Rasmussen), lkh@imm.dtu.dk (L.K. Hansen).

1053-8119/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2010.12.035

classification with clinical perspectives include discrimination of patients with Alzheimer's disease from healthy aging subjects (Klöppel et al., 2008) and identification of individuals in at-risk mental states of psychosis (Koutsouleris et al., 2009).

A comprehensive introduction to classification methods in 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, 2007, 2008; Wang et al., 2007; Formisano et al., 2008a,b; Grosenick et al., 2008; Hanson and Halchenko, 2008; Martino et al., 2008; Ni et al., 2008; Sato et al., 2009). 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. Here the "input space" is the space where the measurements reside. 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). In a recent report ten classification methods were compared in a longitudinal fMRI study of stroke recovery (Schmah et al., 2010), three different two-class classification tasks were considered. The classes were heterogeneous in the sense that each class contained scans from four sub-classes. With respect to classification accuracy, the relative benefit of nonlinear methods compared to their linear counterparts varied over classification tasks. In two classification tasks the nonlinear methods proved superior performance, while in one task there was no significant benefit from applying nonlinear methods. Aiming at improved modeling of effective connectivity, nonlinear modeling has been introduced within the dynamic causal modeling framework (Stephan et al., 2008). In this context, nonlinear modeling allows for identification of models in which the connection between two brain regions is modulated by the activity in a third region. This argument extends to classification models, that is, nonlinear classification allows for signal detection, where *changes in interregional interactions* are related to the experimental variable.

Although nonlinear classification methods are less biased, we note that the far majority of previous pattern classification studies have resorted to linear kernel methods (where the feature space equals the input space). The main reasons are: i) Signal-to-noise ratios in fMRI data may not allow robust identification of nonlinear models at the available data set size as noted in Mørch et al. (1997), see also Cox and Savoy (2003), LaConte et al. (2005), Misaki et al. (2010)); ii) The lack of visualization technology for nonlinear kernel methods in neuroimaging (LaConte et al., 2005).

In the present paper we show that nonlinear modeling may indeed be relevant in the analysis neuroimaging data sets at realistic sample sizes, hence, generic simple visualization techniques for the most popular classifiers are needed.

Model visualization

While a visualization of the relative importance of brain locations to a linear kernel model can be determined simply from the corresponding input weights, it is not as straightforward to obtain a visualization of a nonlinear kernel model.

LaConte et al. (2005) provides an insightful and comprehensive discussion of four visualization schemes for the SVM. Specifically, a method termed feature space weighting (FSW) is proposed and analyzed. FSW comprises the following steps. First, an SVM is trained and a reduced data set is formed by removing scans corresponding to support vectors from the initial data set. Hereafter a summary map is generated by a univariate correlation analysis with the reference function (experimental variable). Hence, the FSW visualization strategy focuses on data points that do not contribute to the decision function. The FSW scheme does however not provide a measure of the relative importance of voxels to the classifier. Hanson and Halchenko (2008) used a sensitivity/perturbation approach for visualization of a linear SVM, adopting a procedure from feature variable selection for SVMs (Guyon et al., 2002; Guyon, 2003; Rakotomamonjy, 2003). These authors performed recursive feature elimination based on the squared values for the linear SVM weight vector.

In the machine learning literature previous work on visualization has focused on so-called pre-image analysis (Mika et al., 1999; Kwok and Tsang, 2004; Teixeira et al., 2008). Pre-image analysis concerns estimation of the inverse mapping from feature space to input space and provides a feature-space *localized* visualization of the kernel machine's mapping which has shown to be useful in, e.g., image denoising (Kwok and Tsang, 2004; Teixeira et al., 2008). In the context of neuroimaging, future applications of pre-image analysis may include image de-noising or visualization of class prototypes. Recently Baehrens et al. (2010) 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 the SVM 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 feature importance at particular points of interest in the data space. The sensitivity analysis that we here investigate provides a single *global* summary map of the features importance. As noted by Kjems et al. (2002) sensitivity analysis is a generic technique for extracting activation maps, that can be applied to any model. The resulting sensitivity maps have already been used for a variety of supervised models including linear regression and neural networks (Zurada et al., 1994, 1997; Kjems et al., 2002; Sigurdsson et al., 2004). To the best of our knowledge the sensitivity map as proposed by Zurada et al. (1994) has not previously been applied to kernel models in the analysis of data sets from functional neuroimaging. In this work we demonstrate that the sensitivity map is indeed useful for generation of global summary maps also for kernel classification methods.

In the following we briefly outline the concepts of supervised learning, kernel methods, and classification. To underline the generality of the visualization scheme we considered three classification models, the SVM, kernel logistic regression (KLR) and the kernel Fisher discriminant (KFD) classifier. These classifiers are described in Appendix B. Next, we introduce the theoretical framework of the sensitivity map for model visualization and present the sensitivity maps for both linear and nonlinear kernel models. Finally, we demonstrate the viability of the sensitivity map in an fMRI experiment.

Materials and methods

Classification models and kernel methods

Consider a labeled data set $\mathcal{D}=\{\mathbf{x}_n,t_n\}_{n=1}^{N}$, where $\mathbf{x}$ is a P dimensional input vector while t is the corresponding target variable. In fMRI $\mathbf{x}$ contains (part of) a brain scan volume. The target t may be class labels (classification) or continuous real values (regression) and typically encodes behavior. In terms of predictive modeling we limit our discussion to classification models, bearing in mind that the following theory is readily applied to regression models. We consider for simplicity a binary classification setup with only two classes, hence $t\in\{-1,1\}$, all expression may be generalized to multiple classes (Zurada et al., 1994).

In a linear classifier the objective is to estimate model parameters $\boldsymbol{\theta}=\{\mathbf{w},b\}$, such that the discriminant function

$$y(\mathbf{x};\theta) = \mathrm{w}^{\top}\mathrm{x} + b, \tag{1}$$

generalizes well to future data. Objects are classified according to the sign of y in Eq. (1). A more flexible classifier can be implemented by a nonlinear projection of the observations $\mathbf{x}_n$ into a larger space $\mathcal{F}$ (often referred to as a feature space) (Shawe-Taylor and Cristianini, 2004). Let $\phi:\mathcal{X}\rightarrow\mathcal{F}$ be a mapping from the input space $\mathcal{X}$ to $\mathcal{F}$. Kernel-based algorithms seek to find a linear decision boundary of the same form as in Eq. (1) in feature space

$$y(\mathbf{x};\theta) = \mathrm{w}^{\top}\phi(\mathbf{x}) + b. \tag{2}$$

If the weight vector $\mathbf{w}$ can be expressed as a linear combination of the training points $\mathbf{w}=\sum_{n=1}^{N}\alpha_n\phi(\mathbf{x}_n)$ we can use the *kernel trick* to express the discriminant function as

$$y(\mathbf{x};\boldsymbol{\theta}) = \sum_{n=1}^{N}\alpha_n k(\mathbf{x}_n,\mathbf{x}) + b, \tag{3}$$

with the model now parametrized by the smaller set of parameters $\boldsymbol{\theta}=\{\boldsymbol{\alpha},b\}$ (Lautrup et al., 1994). The kernel is a function that returns the inner product $k(\mathbf{x}_i,\mathbf{x}_j)=\phi(\mathbf{x}_i)^{\top}\phi(\mathbf{x}_j)$ in the feature space and must satisfy certain conditions (Shawe-Taylor and Cristianini, 2004).

By far the most used kernel method in neuroimaging is the SVM classifier. In neuroimaging applications the SVM is usually used along with the linear kernel $k(\mathbf{x}_i,\mathbf{x}_j) = \mathrm{x}_i^{\top}\mathbf{x}_j$ (here the feature space equals the input space). In terms of model interpretation it is convenient to work with the linear kernel, since the model can be visualized via the weight vector $\mathbf{w}$ as a "discriminating volume" that shows the relative importance of each voxel to the classifier (LaConte et al., 2005; Mourão Miranda et al., 2005). Note that $\mathbf{w}$ in the linear case is simply a weighted average of the training examples $\mathbf{w}=\sum_{n=1}^{N}\alpha_n\mathbf{x}_n$. Examples of other and more flexible kernels are the polynomial kernel $k(\mathbf{x}_i,\mathbf{x}_j)=(\mathrm{x}_i^{\top}\mathbf{x}_j+q)^2$ and the RBF kernel $k(\mathbf{x}_i,\mathbf{x}_j)=\exp(-||\mathbf{x}_i-\mathbf{x}_j||^2/q)$, where q is a kernel parameter. Such kernels allow for implementation of a nonlinear decision boundary in the input space. However this is at the expense of model interpretation, since a weight cannot be assigned to each input feature in the same way as with the linear kernel.

Probabilistic sensitivity maps

In the context of discriminative models in neuroimaging we are interested in the importance of the different input features, i.e., voxels to the classifier. Sensitivity analysis or the *sensitivity map* is a simple measurement of this importance. For early definitions, see Zurada et al. (1994, 1997), and for recent applications of the sensitivity map in functional neuroimaging (Kjems et al., 2002; Strother et al., 2002) and in skin cancer detection by Raman spectroscopy (Sigurdsson et al., 2004).

We 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 $p(\mathbf{x})$. This can be accomplished by the sensitivity map which is defined as the expected value of the squared derivatives of the function with respect to its arguments,

$$s_j = \int\left(\frac{\partial f(\mathbf{x})}{\partial x_j}\right)^2 p(\mathbf{x})d\mathbf{x}, \tag{4}$$

where s_j denotes the sensitivity measure at the j'th voxel, thus the set $\{s_j, j=1,\ldots,P\}$ is a volume. Since we aim at the global sensitivity over the input space we square the derivative to avoid cancelation of positive and negative terms, as some input regions may have positive sensitivity while others have negative sensitivity. Different choices for the function $f(\mathbf{x})$ exist.

The simplest choice of the visualization function $f(\mathbf{x})$ in Eq. (4) is the discriminant function $y(\mathbf{x})$ in Eq. (2), hence the sensitivity map becomes

$$s_j = \int\left(\frac{\partial y(\mathbf{x})}{\partial x_j}\right)^2 p(\mathbf{x})d\mathbf{x}. \tag{5}$$

Due to the potential complexity of performing the integral we resort to the empirical estimate

$$\hat{s}_j = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\partial\alpha^{\top}\mathbf{k}\mathbf{x}}{\partial x_j}\right)^2_{\mathbf{x}=\mathbf{x}_n}, \tag{6}$$

where $\mathbf{k}_{\mathbf{x}}$ is a $(N\times1)$ vector that holds the elements $\mathbf{k}_{\mathbf{x}}^{(n)} = k(\mathbf{x}_n,\mathbf{x})$, where $\mathbf{x}_n$ is the n'th training example.

Thus, the sensitivity map requires the derivative of the kernel function. In the following we consider the sensitivity maps for kernel models with linear, polynomial and RBF kernels.

I — Sensitivity map for the linear kernel

First we note that $\mathbf{k}_{\mathbf{x}}$ holds the elements $k(\mathbf{x}_n,\mathbf{x}) = \mathrm{x}_n^{\top}\mathbf{x}$, where n is the index of a specific training example. The derivative in Eq. (6) is then calculated as

$$\begin{aligned}\frac{\partial\alpha^{\top}k\mathbf{x}}{\partial x_j} &= \frac{\partial\sum_n\alpha_n x_n^{\top}x}{\partial x_j}\\ &= \sum_n\alpha_n x_{n,j},\end{aligned} \tag{7}$$

where $x_{n,j}$ is the j'th voxel in training example n. Hence the sensitivity map for a linear kernel model is equal to the square of the conventional input space *weight map* given by $\tilde{w}_j = \sum_n\alpha_n x_{n,j}$.

II — Sensitivity map for the polynomial kernel

For the polynomial kernel $\mathbf{k}_{\mathbf{x}}$ holds the elements $k(\mathbf{x}_n,\mathbf{x}) = (\mathrm{x}_n^{\top}\mathbf{x}-q)^2$. The derivative is calculated as

$$\begin{aligned}\frac{\partial\alpha^{\top}\mathbf{k}_{\mathbf{x}}}{\partial x_j} &= \frac{\partial\sum_n\alpha_n\left(\mathbf{x}_{\mathbf{n}}^{\top}\mathbf{x}-\mathbf{q}\right)^2}{\partial x_j}\\ &= \sum_n\alpha_n 2\left(\mathbf{x}_{\mathbf{n}}^{\top}\mathbf{x}-\mathbf{q}\right)x_{n,j},\end{aligned} \tag{8}$$

where x_j again is the j'th feature in $\mathbf{x}$.

III — Sensitivity map for the RBF kernel

For the RBF kernel $\mathbf{k}_{\mathbf{x}}$ holds the elements $k(\mathbf{x}_n,\mathbf{x})=\exp\left(-\frac{||\mathbf{x}_n-\mathbf{x}||^2}{q}\right)$. The derivative is calculated as

$$\begin{aligned}\frac{\partial\alpha^{\top}\mathbf{k}\mathbf{x}}{\partial x_j} &= \frac{\partial\sum_n\alpha_n\exp\left(-\frac{||\mathbf{x}_{\mathbf{n}}-\mathbf{x}||^2}{q}\right)}{\partial x_j}\\ &= \sum_n\alpha_n 2\frac{\left(x_{n,j}-x_j\right)}{q}\exp\left(-\frac{||\mathbf{x}_{\mathbf{n}}-\mathbf{x}||^2}{q}\right).\end{aligned} \tag{9}$$

Since the functional form of the discriminant function $y(\mathbf{x})$ is identical for all classification models we consider here, there is no difference in how sensitivity maps are generated for the different classifiers. Hence the algorithm used is the same. The maps will of course in general differ since different classifiers provide different estimates of model parameters according to model assumptions. In

Appendix A we provide code for calculation of the sensitivity map for a classifier with an RBF kernel.

Data sets

I — Simple illustrative simulation

For illustration of the properties of nonlinear classifiers and their visualization via the sensitivity map we considered a simple simulation setup inspired by Pereira and Botvinick (in press). Consider a binary classification task with a feature vector comprising five to six voxels. Each voxel holds samples drawn from a Gaussian distribution. We generated training, validation, and test sets all with 300 examples. In the following the i'th voxel is referred to as x_i. The scenarios were as follows:

- Scenario 1: x_1 and x_2 are equal informative with respect to the class labels, i.e. class distributions are separated with the same distance along the two dimensions.
- Scenario 2: x_1 is more informative than x_2, i.e. class distributions are separated more along x_1 than along x_2.
- Scenario 3: x_1 and x_2 are informative with respect to the class labels through an XOR coupling. Hence their product is informative, but there is no difference in class means along the two dimensions.
- Scenario 4: x_1 is informative (same strength as in Scenario 1). x_2 and x_3 are informative through an XOR coupling.
- Scenario 5: Same as Scenario 4, but where x_1 is less informative (class means are reduced to half the distance).

Additionally, we add three uninformative voxels in each scenario with the same distribution as the one of the informative voxels (but with no class mean difference or coupling between voxels). We trained an SVM classifier with both linear (linear SVM) and RBF (nonlinear SVM) kernels to discriminate between the classes. The relative voxel importance in the SVM classifier with the linear kernel was measured in terms of the absolute weight vector value. For the SVM classifier with the RBF kernel we measured voxel importance by the square root of the sensitivity map.

II — fMRI 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 3 T (Siemens Magnetom Trio) scanner using an 8-channel head coil (Invivo, Florida, USA). The data set was collected using an EPI (echo planar imaging) GRE (gradient-echo pulse) sequence with 28 slices acquired in interleaved order with the following acquisition parameters: Repetition time (TR) 1670 ms, echo time (TE) 30 ms, flip angle (FA) 90°, field of view (FOV) 192 × 192 mm, 64 × 64 acquisition matrix, the voxel size was 3 × 3 × 4.5 mm. For visualization purposes a high resolution anatomical scan was obtained using a magnetization prepared rapid gradient echo (MPRAGE) sequence with 192 sagittal slices and 1 mm isotropic resolution. Additional sequence parameters were as follows: TR = 15.4 ms, TE = 3.93 ms, flip angle FA = 9°, FOV = 256 × 256. 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), and (4) reversing checkerboard on both halfs of the screen (BOTH). In order to maintain attention the participants were instructed to keep focus on a small circle presented in the center of the screen during the experiment, and to respond with a right hand button press to a change in the color of the circle. Each stimuli condition was presented for 15 s followed by 5.04 s 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.

Image pre-processing

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 using a two-step procedure (6-parameter affine transformation with least squares objective function). (2) Spatial normalization of the mean EPI image to the EPI template in SPM8 (nonlinear frequency cutoff: 25, nonlinear iterations: 16, nonlinear regularization: 1). (3) The estimated warp field was applied to the individual EPI images. The normalized images were written with 3 mm isotropic voxels (4-th degree B-spline interpolation). (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 was masked with a rough whole-brain mask (75,257 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 s were projected out of the data. (8) Within each subject the individual voxel time series were standardized.

Unsupervised analysis

To provide the reader with an understanding of the underlying structure of the data set we performed an initial principal component analysis (PCA) of the data set. This analysis was invoked to illustrate the complexities of the classification tasks that can be formulated in the data set — no dimensionality reduction was performed in the classification analysis. To estimate the intrinsic dimensionality in the data we used the procedure of Hansen et al. (1999). In this framework PCA is defined in terms of a cost function under the assumptions that the data can be modeled by a multivariate Gaussian distribution. For a particular number of principal components (PCs) K the vector space is split into a signal space spanned by the first K eigenvectors of the data covariance matrix, sorted according to the size of eigenvalues, and a noise space defined by the remaining eigenvectors with non-zero eigenvalues. The covariance matrix of the Gaussian distribution is then modeled as the sum of a contribution from the K largest PCs and a noise contribution. After parameter estimation we test how well the Gaussian model explains the test data set. By varying K we observe the generalization error as a function of subspace dimensionality. Subjects were considered as the resampling unit, where the PCA was based on data from three subjects, and the generalization error was measured on the out-of-sample subjects. This was repeated for all possible splits of the data set.

Classification and model visualization

In the analysis of the fMRI data set we considered subjects as the basic resampling unit, where the statistical classifiers were trained on data from a subset of subjects, while the target labels were inferred for scans from subjects in the out-of-sample subset. Scan 7–11 in each epoch were used in the predictive modeling, and the remaining volumes were discarded to avoid contaminating effects of the hemodynamic BOLD signal. We aimed for a “whole-brain” and single block classification with temporal compression of scans within the same block (Mourão Miranda et al., 2006). No feature selection prior to the application of the classification models was performed. Two binary classification tasks were formulated:

- Classification task I: scans from condition (LEFT) were assigned to class −1 while (RIGHT) was assigned to class 1. We expected this classification task to be relatively easy for the linear methods to solve. Intuitively, this classification task can be compared to Scenario 1 in the illustrative simulation.

- 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 1. This task was expected to be harder for the linear methods and relatively easy for the more flexible nonlinear methods to solve. By this labeling we intended to introduce an artificial coupling between brain regions, equivalent to computer science's *xor* function. Intuitively, this classification task can be compared to Scenario 3 in the illustrative simulation.

Classification models were established in terms of the SVM, KLR, and KFD. Both the linear and the RBF kernels were considered. In the following we will refer to models with linear kernels as *linear* models and the models with an RBF kernel as *nonlinear* models. All models have a regularization parameter that needs to be selected (C for the SVM and λ for KLR and KFD). Additionally the RBF kernel also has the parameter q that controls the kernel width. Model performance was assessed for C and λ ranging over the interval $2^{-40}2^{40}$ relative to the average non-zero eigenvalue of the data covariance matrix. q was varied in the range $2^{-5}2^{15}$ relative to the average input-space distance to the nearest 25% points across all training examples. These heuristics were chosen to give the model hyperparameter values that are somewhat on the same scale as the data. In order to construct maps of voxel importance to the classifiers we used the absolute weight map for the linear models and the square root of the sensitivity map for the nonlinear models.

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. These prediction accuracies were averaged and considered as the prediction metric (p) of the NPAIRS scheme. In addition, the Pearson's correlation coefficient between spatial maps derived from the two models was calculated as the 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 activation volume (rSPI). This procedure was repeated 10 times, where subjects were distributed between splits according to all possible combinations.

We use the p and r metrics for model optimization, where we choose model parameter that maximized both metrics jointly over the entire space of cross-validated results. We refer to such a model as a *pr-optimized* model.

Additionally, to assess the consensus between visualizations of the linear and the nonlinear classifiers we adopted the strategy proposed in Hansen et al. (2001). First, the histograms of the average rSPI for the nonlinear SVM were transformed to match the average rSPI of the linear SVM. Hereafter, scatter plots of the average rSPIs were generated. For visualization the average rSPIs were thresholded to include supra threshold voxels in the upper 10 percentile of the distribution of the average rSPIs and projected onto the average anatomical scan.

Results

I — Simple illustrative simulation

Fig. 1 shows the performances of the classifiers in the different scenarios. In the plots the bars show the relative size of the voxel-wise importance measure. The bars are based on 250 realizations of the simulation. For each realization we scaled the “weight/importance” vector to unit variance. In the plots we also provide the average test set accuracy. In Scenario 1 x_1 and x_2 are given the same importance for linear SVM as expected and in accordance with Pereira and Botvinick (in press). In the visualization of the nonlinear SVM via the sensitivity map we obtain a similar result, hence all relevant voxels are identified and they have the same magnitude with respect to the importance measure. In Scenario 2 there is also agreement between the visualization of the linear and nonlinear SVM. In Scenario 3 x_1 and x_2 are correctly identified by the nonlinear SVM as important. Also note the poor performance of the linear SVM with respect to test set accuracy compared to the nonlinear SVM. In Scenario 4 the linear SVM only identifies x_1 as informative, while the nonlinear SVM also identifies x_2 and x_3 as informative. In Scenario 5 the relative importance of x_2 and x_3 increases when x_1 become less informative. Also note, the nonlinear SVM have greater prediction accuracy than

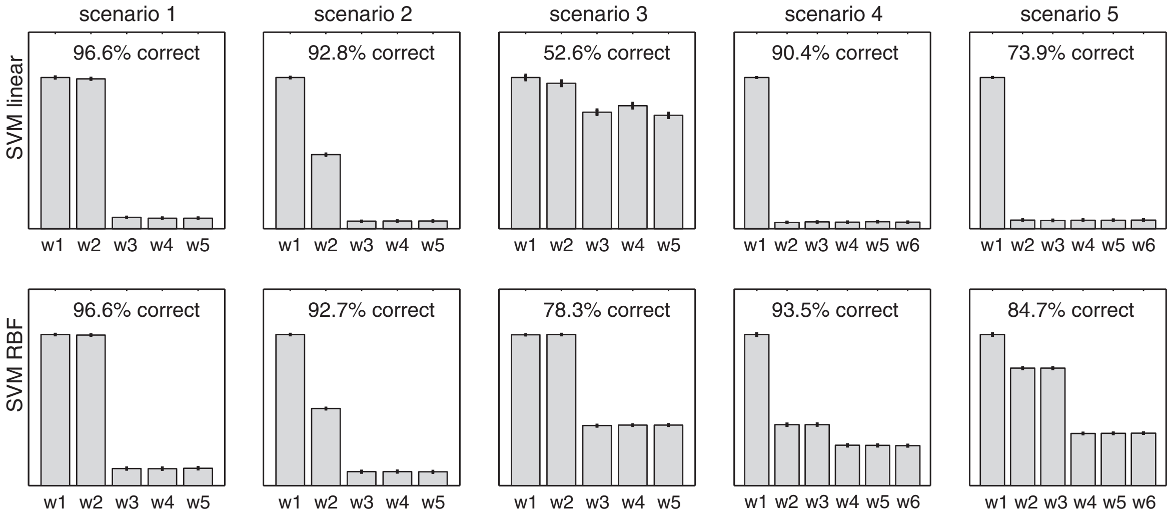


Fig. 1. Illustration of measurement of voxel importance in a classifier. We consider five scenarios with a data vector comprising five or six voxels. See the text for description of the different scenarios. In the simulation the classifiers were trained on 300 examples and tested on 300 examples. Model hyperparameters were chosen based on performance on a validation set also with 300 examples. The relative voxel importances in the SVM classifier with the linear kernel (SVM linear) were measured in terms of the absolute weight vector value. For the SVM classifier with the RBF kernel (SVM RBF) we measured voxel importance by the sensitivity map. The bars show the relative size of the voxel-wise importance measure. The bars and prediction accuracies are based on 250 realizations of the simulation. Error bars depict one standard deviation of the mean.

the linear SVM in Scenario 4 and 5, since the nonlinear model uses information from x_2 and x_3 in addition to the information in x_1.

II — fMRI data set — visual paradigm

Unsupervised analysis

Within each NPAIRS split we estimated the dimensionality of the PCA subspace that minimized the generalization error on the test sets according to the multivariate Gaussian model of the data. Optimal dimensionality ranged from three to seven across the different splits, with four PCs retained on average. Fig. 2 shows an example of the three first PCs estimated from the training set in a single NPAIRS split. The scatter plots show both training and test examples projected onto the PCs. The plots of PC projections show both training and test points, where training points are with filled markers. The blue and red voxels on the brain slices correspond to negative and positive PC loadings respectively. The maps are thresholded to show the 5 upper positive and negative percentiles. The first two PCs had just a few positive elements, hence it appears mainly with "blue" loadings. The score plots illustrate the complexity of classification tasks that can be formulated within the present data set. For example, (LEFT) vs. (RIGHT) can relatively easily be separated with a linear decision boundary within the space spanned by any combination of two of the first three PCs. (NO) and (BOTH) vs. (LEFT) and (RIGHT) cannot be separated along any of the combinations of PCs by a linear decision boundary. However, it appears fairly easy for a nonlinear method to perform such a separation within the combination of PC1, PC2 and PC3. Hence, the flexibility of nonlinear methods allows for separation of any categorization of the clusters (four classes) as illustrated in the PC1/PC2/PC3 plot. The unsupervised analysis is only invoked as an exploratory tool, no dimensionality reduction was performed and all classification models were based on the original data space.

Classification task I

To summarize the classification setup, the classifiers were trained to distinguish between the conditions (LEFT) and (RIGHT). Both training and test sets contained scans from three subjects. Results are based on 10 NPAIRS resampling splits. Fig. 3 shows the model performance in terms of prediction accuracy and pattern reproducibility over a range of parameter values for models with the RBF kernel. We observe that for a fixed q there is a tendency of an increased prediction accuracy at large values of C. Also both metrics increase at increasing kernel width. For the nonlinear classifier type we considered two models for further analysis. One was chosen with $q = 2^{15}$, since models with RBF kernels become increasingly linear when the kernel width becomes large. The other model selected for each classifier was the pr-maximizing model indicated by a cross in the top panel of Fig. 3. For classifiers with the linear kernel we also selected the pr-maximizing model. Performances of the pr-maximizing models are summarized in Table 1. All models achieved ~100% in prediction accuracy in test data. The model visualization also provides a high performance in terms of pattern reproducibility.

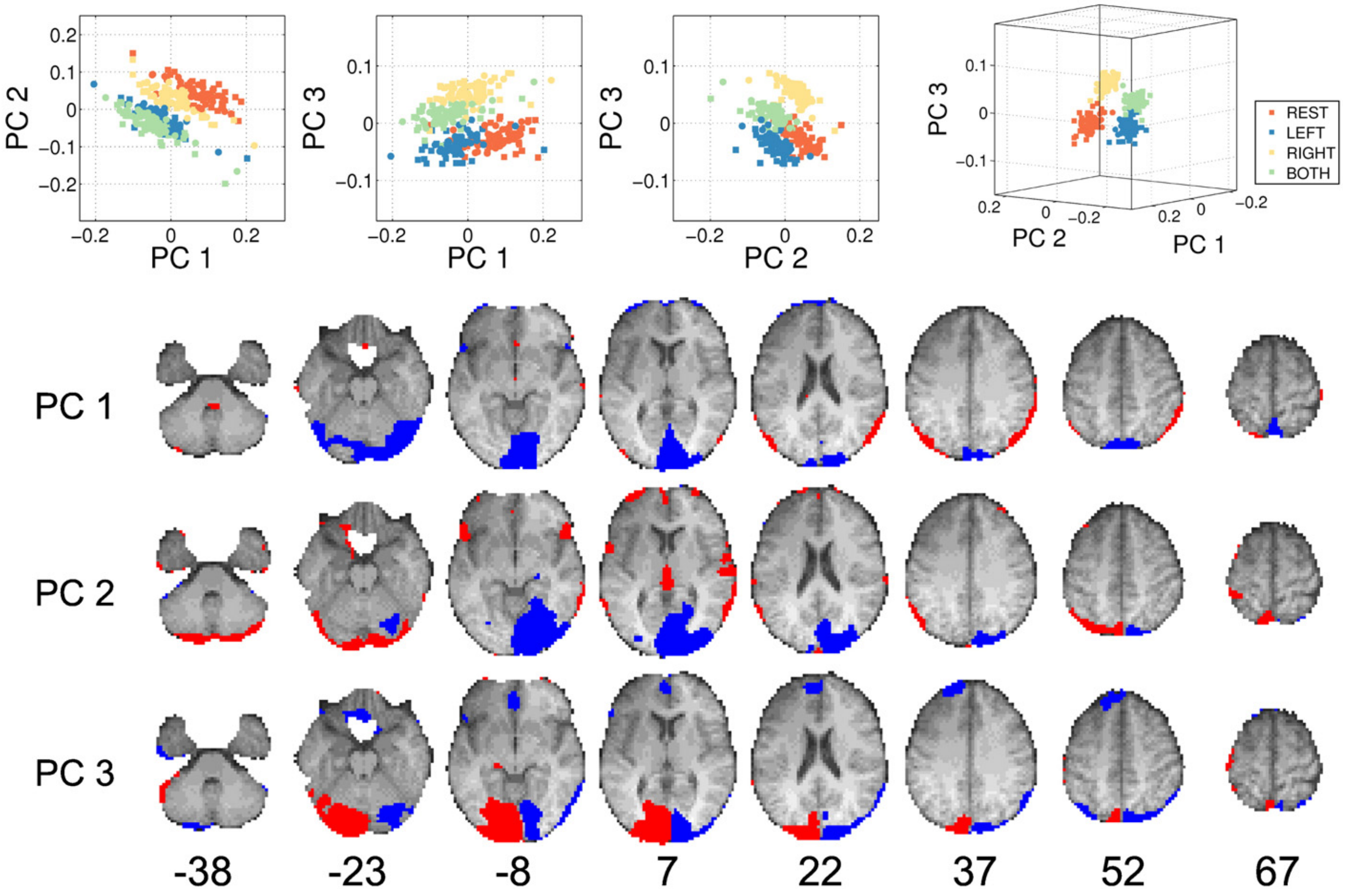


Fig. 2. PCA analysis of the fMRI data set. An example of the three first PCs estimated from the training set in a NPAIRS split. The scatter plots show both training (square markers) and test examples (circle markers) projected onto the PCs. Both the training and test sets comprised three subjects. The blue and red voxels on the brain slices correspond to negative and positive PC loadings respectively. The maps are thresholded to show the 5 upper positive and negative percentiles. The maps are projected onto an average anatomical scan of the six subjects included in the analysis. Numbers under the slices denote z coordinates in MNI space.

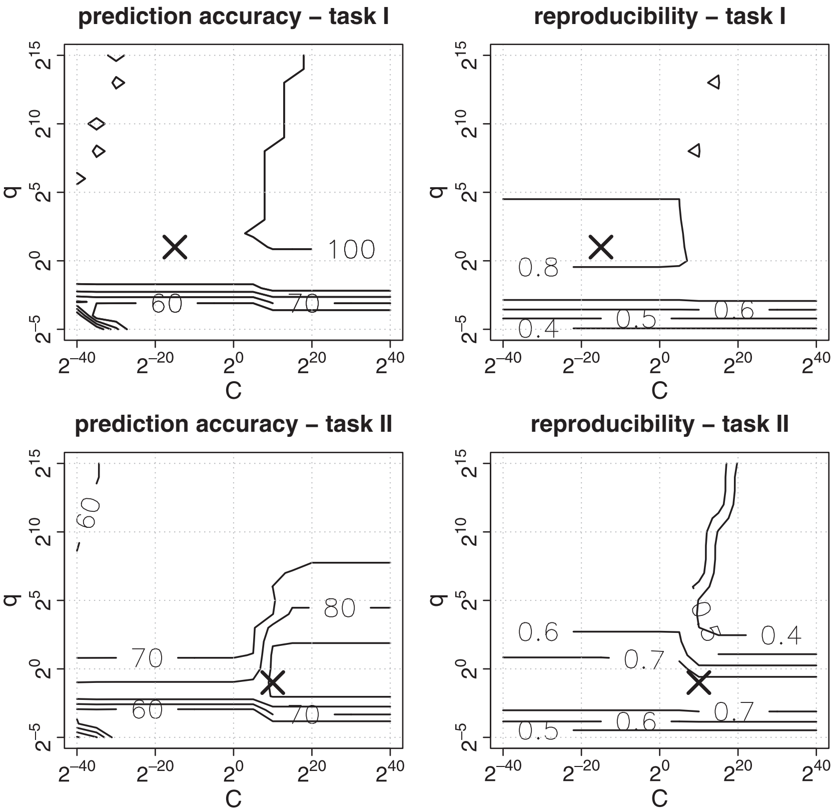


Fig. 3. Parameter optimization grid for the SVM with an RBF kernel. *C* is the "complexity" parameter of the SVM, and *q* is the kernel width. Models were optimized against both prediction accuracy and pattern reproducibility. The plots are based on mean values of 10 NPAIRS splits. Top row is classification task I and bottom row is classification task II. The crosses indicate models selected according to pr-maximization (see Materials and methods section).

Fig. 4 provides a comparison between brain maps extracted from linear and nonlinear SVM classifiers. In the scatter plot each dot corresponds to a voxel. The lines indicate thresholds for the upper 10 percentiles. Points above the horizontal line are supra threshold voxels in the average rSPI of the nonlinear SVM. Points right to the vertical line are supra threshold voxels on in the average rSPI of the linear SVM. Orange points indicate agreement between both models, while blue and yellow are supra threshold voxels in the linear SVM and nonlinear SVM respectively. For the nonlinear SVM with $q = 2^{15}$ there is a large extent of consensus with the linear SVM. Hence, both models and their visualization identify that voxels in the visual cortex contribute with relevant information to the classifiers. Also the nonlinear SVM classifier that is selected according to the pr-maximization criterion shows great similarities with the linear SVM. While there is a high degree of consensus in the upper part of the scatter plot the models tend to disagree in the lower part of the scatter plot. The nonlinear SVM also identifies small frontal regions as important among the 10% fraction of voxels. Unthresholded versions of the average rSPIs are found in Supplementary materials Fig. 1.

Table 1
Results for classification task I and II. For all models performances are reported for parameter settings that optimize both prediction accuracy (p) and reproducibility (r) jointly. The table reports the mean values and standard errors based on 10 NPAIRS splits.

	Task I		Task II	
	p	r	p	r
SVM linear	100.0% (0.00)	0.79 (0.022)	67.8% (1.12)	0.57 (0.010)
SVM RBF	99.7% (0.49)	0.81 (0.028)	92.2% (1.73)	0.75 (0.021)
KLR linear	100.0% (0.00)	0.80 (0.025)	57.4% (1.71)	0.57 (0.010)
KLR RBF	100.0% (0.00)	0.81 (0.024)	92.0% (1.78)	0.76 (0.021)
KFD linear	100.0% (0.00)	0.80 (0.025)	57.4% (1.71)	0.57 (0.010)
KFD RBF	99.9% (0.22)	0.81 (0.026)	92.3% (1.76)	0.75 (0.020)

Classification task II

The classifiers were trained to distinguish between the conditions (NO) and (BOTH) vs. (LEFT) and (RIGHT). Fig. 3 shows the model performance of the SVM in terms of prediction accuracy and pattern reproducibility over a range of parameter values for models with the RBF kernel. A small kernel width compared to that in classification task I was supported by both performance metrics. Results in terms of prediction accuracy and pattern reproducibility are summarized in Table 1. For all classifiers the linear models showed a decreased performance in terms of both prediction accuracy and pattern reproducibility compared to classification task I, while the nonlinear model show relatively good performance. According to Fig. 4 there is less consensus between the linear and nonlinear models about which voxels that should be included in the upper 10% fraction in the average rSPI. Voxels identified as relevant by the nonlinear SVM largely resemble the voxels identified as relevant by both classifiers in classification task I. The linear model captures voxels in the visual cortex and relative large blobs that were not included among the upper 10 percentile identified in classification task I. Note that the linear model did not identify a relative large fraction of the voxel in the right visual cortex among the upper 10 percentile. Unthresholded versions of the average rSPIs are found in Supplementary materials Fig. 1.

Model consensus across classifiers

Fig. 5 provides a consensus analysis between the average rSPIs of all models in both data sets. Model parameters were selected according to pr-optimization. First we observe that within each kernel

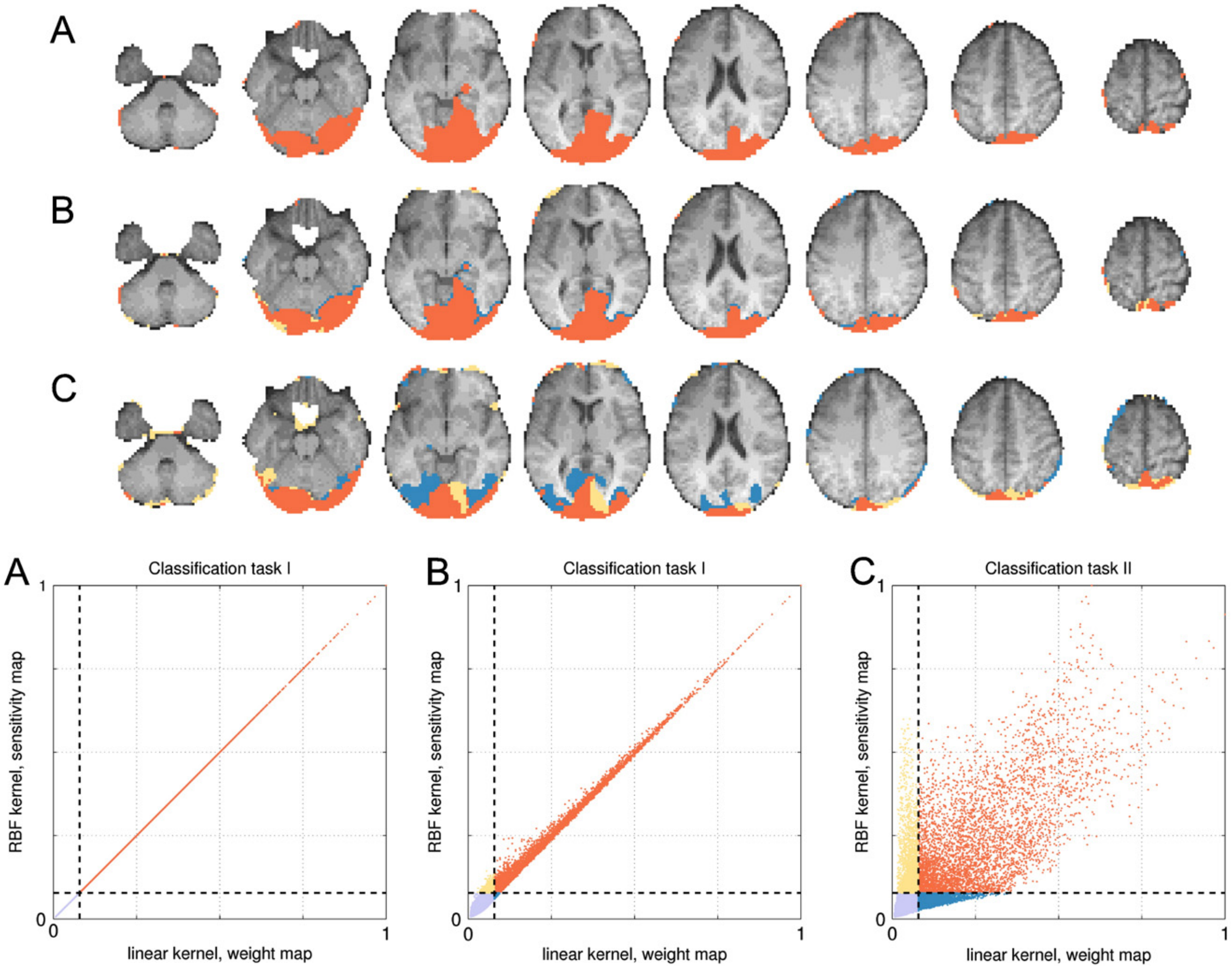


Fig. 4. fMRI experiment — comparison of SVM with linear and RBF kernels. The sensitivity map served as a visualization of SVM with RBF kernels, while the SVM with linear kernels were visualized via the weight map. In (A) and (B) the classifiers were trained to discriminate (LEFT) vs. (RIGHT) (classification task I), while in (C) the classification task was (NO) and (BOTH) vs. (LEFT) and (RIGHT). For (A) the width of the RBF kernel was set to a large value ($q = 2^{15}$), while kernel parameters in (B) and (C) were selected according to pr-maximization (see Materials and methods). For each model an average reproducible SPI was based on all 10 NPAIRS resampling splits. Voxels in the two maps were transformed to the same distribution via consensus analysis (see Materials and methods section). In the scatter plot each point corresponds to a specific voxel. The lines correspond to levels, where the upper 10 percentile in each image was retained. Color coding; orange — agreement in both visualization, blue — specific to linear kernel, yellow — specific to RBF kernel. The summary maps show the distribution of voxels projected onto the average anatomical scan of the six subjects included in the analysis.

type and classification task there is a large degree of consensus across classifier. In classification task I there is also a strong consensus between the pr-optimized linear and nonlinear models. For classification task II there is less consensus between the linear classifiers and the classifiers of classification task I. This is in contrast to the nonlinear models in classification task II, that show larger similarities with models in classification task I.

Discussion and conclusion

Performance of linear and nonlinear models

In the current paper we have investigated the probabilistic sensitivity map for visualization for nonlinear kernel models. We have illustrated the performance of the sensitivity map in a simple classification task in fMRI data. In classification task I that was fairly easy for the linear classifiers to solve, we have shown that the sensitivity map is an appropriate visualization of a corresponding nonlinear model as the sensitivity map indeed identified regions where discriminative information resides. Furthermore, there was a large agreement between the conventional weight map visualization for linear kernel methods and the proposed sensitivity map visualization for nonlinear kernel methods. In classification task II with a more complex object categorization, the nonlinear models outperformed linear models, both in terms of prediction accuracy and pattern reproducibility. Additionally, the nonlinear models identified voxels that could be expected to contribute with relevant information.

Further intuition about the improved performance of the nonlinear models can be gained from the unsupervised analysis in Fig. 2. While the classes in classification task II are not separable along any of the projections, it appears fairly easy for a nonlinear model to perform such a separation within the combination of PC1, PC2 and PC3. Note we cannot in general draw conclusions about the separability of the classes by inspecting a low dimensional representation. Thus, there may be a subspace of higher dimensionality that allows for class separation by linear models. Conceptually, in classification task I the class label can be inferred solely from the presence of activation in the left hemisphere visual cortex (or in the right visual cortex). In classification task II more information is required to support a successful classification. That is, besides knowing if the left visual cortex was activated we need to know if stimuli also evoked an activation in the right visual cortex (or vice versa).

Our label assignment in classification II may seem quite imprudent, however we advocate for our choices by the following. i) In

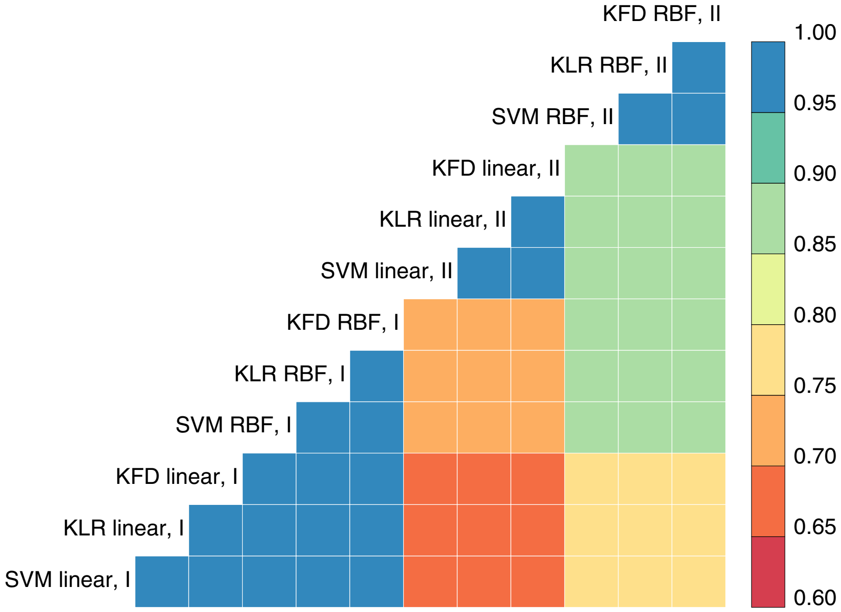


Fig. 5. Across classifier consensus analysis. For each classifier we obtained an average rSPI based on 10 NPAIRS splits. The plot shows the correlation of these brain maps across classification tasks and classification models. Models with code I and II are build on classification task I and II respectively.

neuroimaging, it may be tempting to conclude that nonlinear models are unnecessary since typical neuroimaging data sets are characterized by few samples in a high dimensional feature space. Our simple example illustrates, that nonlinear models are indeed important if generalizability of the model is to be considered. ii) In more complex paradigms, the categorization of scans may not be obvious. Scans can also be characterized by multiple labels rather than a single label as in Schmah et al. (2010). In such scenarios there may be a considerable heterogeneity within classes, and nonlinear models may be more powerful. Our example indicates, that while linear models may be relative sensitive to the scan labeling the nonlinear methods are capable of implementing more complex decision rules. For these methods the probabilistic sensitivity map is a simple but yet powerful tool for interpretation of such nonlinear models.

Interpretation of the sensitivity map

The sensitivity map is defined so that the map reveals brain regions that are of relevance to the mental state classifier. The sensitivity map provides a global analysis, where the change in classifier output is measured when changing a single voxel while holding all other voxels constant. This measure is calculated throughout the entire brain. The sensitivity map for a linear kernel classifier is identical to the squared weight map, that is widely used for interpretation of linear models. According to the map definition, there is no “sign” information present in the map. We square the derivative to avoid cancelation of positive and negative terms that may occur due to the potential existence of local sensitivities with opposite direction. This is in contrast to a conventional GLM contrast map and the weight map visualization of linear models, where “negative” regions occur due to contrast or target label definitions respectively. Note, that in the general nonlinear case sign information is meaningless because regions can have arbitrary multiplicative relations with other regions, hence only the relative strength of involvement is relevant.

A similar approach for evaluation of the relative importance of voxels to the classifier is the noise perturbation sensitivity analysis, that has been applied to neural networks (Hanson et al., 2004). To determine the relative voxel importance to the generalization results, Gaussian noise is added to a single voxel at a time in this procedure, and it is observed how the classifier performance is affected. For each voxel the noise is sampled and added hundreds of times to achieve a stable estimate of the relative importance. This is repeated for all voxels leading to a computationally expensive procedure. To our knowledge this analysis has not been applied to nonlinear kernel models in neuroimaging. While the noise perturbation analysis relies on realizations of random noise, the sensitivity map we investigate here provides a deterministic visualization procedure. Additionally, the sensitivity map is computationally efficient to generate (see Appendix A).

Pre-processing and resampling choices

We applied a default pre-processing strategy in SPM8. As with all other models the appropriate pre-processing strategy for nonlinear models is indeed data set dependent, and we refer to Strother et al. (2002) for a discussion of this issue. Different pre-processing strategies in the context of classification models have been investigated in LaConte et al. (2005) and Chen et al. (2006). To investigate the potential of the sensitivity map as a visualization of nonlinear kernel methods, we have adopted a rather simple strategy with whole-brain and single block classification. The whole brain approach is attractive, since no a-priori hypothesis on the spatial location of relevant regions is required. In machine learning in general, feature selection has several benefits such as facilitation of data visualization/understanding, reduction in data storage, classifier training time and improvement of prediction performance (Guyon, 2003). In applications of kernel methods to fMRI data analysis feature selection has indeed proven to improve prediction performance and identification of voxels relevant to the classification task (Hanson and Halchenko, 2008; Martino et al., 2008). To this end, the sensitivity map provides a quantitative measure of the relative importance of input features to the classification. Hence, the sensitivity map is directly applicable in feature extraction schemes, where nonlinear kernel methods are to be considered. In our research, we performed recursive feature elimination (RFE), which increased the performance (in terms of classification accuracy) of the SVM with the RBF kernel in classification task II, while the performance of the SVM with a linear kernel only was degraded when features were eliminated. However, since the main focus of this study was to evaluate the performance of the sensitivity map as a tool for model visualization and not only to boost the classification accuracy, we have chosen to restrict the present paper to a simple classification framework with whole brain classification. In data sets with more subtle effects where only a few voxels contain discriminative information the whole-brain (and RFE) approach may fail. In such scenarios it may be useful to consider univariate feature selection procedures (Guyon, 2003; Martino et al., 2008) or to work on beta estimates or t-values from a conventional univariate correlation analysis (Misaki et al., 2010).

Note that model hyperparameter selection was performed on the same data set as performance in terms of prediction accuracy and pattern reproducibility was evaluated on. Hence, the relative performances of the classifiers cannot be considered to be unbiased. If the objective of the analysis was to maximize e.g. prediction accuracy the standard procedure would be to select hyperparameters based on a validation set. Since we were interested in both classification accuracy and pattern reproducibility, and due to the limited amount of data in the particular data set, we choose the split half resampling strategy.

Linear vs. nonlinear classifiers

In the present study we have shown that the performance of existing classification procedures within the field of neuroimaging is highly dependent on label assignment. We advocate that nonlinear models should indeed be considered in future classification experiments. Since at large kernel widths, the RBF kernel will approach linear kernels, there is no obvious reason, from a classification accuracy perspective, to restrict the investigation to linear kernels. We can think of two motivations for application of nonlinear modeling in neuroimaging. First, nonlinear modeling allow for signal detection, where changes in interregional interactions are related to the experimental variable. Such interaction effects cannot be detected by a linear method by definition. Second, in classification tasks where the exact labeling of scans are unknown or where each class consists of several sub-classes (Schmah et al., 2010), nonlinear models may be more powerful in comparison to their linear counterparts. An important feature of many kernel methods is that optimum model parameters can be found as a solution to a convex optimization problem. In such a case there exists a single and global optimum. This is in contrast to other flexible learning methods e.g. neural networks, where one may get stuck in a local minimum during model optimization.

When using linear kernel methods, it is relatively simple to identify voxels that drive the classifier. In the present study we have investigated one strategy for generation of activation maps for nonlinear kernel methods as proposed by Kjems et al. (2002) and LaConte et al. (2005). Now, it is relevant to address the issue when nonlinear models should be used in neuroimaging. As stated by O'Toole et al. (2007), it is difficult to assess, a priori, if a linear or a nonlinear classifier should be used for a specific data set. A sensible strategy is to consider both models and use prediction error levels and robustness of visualizations as a performance metric to choose one if necessary (Kjems et al., 2002; Strother et al., 2002). The relative performance of the classifiers will depend on the particular properties of brain response patterns and on the amount of data available as discussed by Mørch et al., (1997) and Misaki et al. (2010).

In summary, the present study has provided evidence that the probabilistic sensitivity map is a viable tool for visualization of nonlinear kernel methods in neuroimaging. We hope that the proposed scheme for visualization of nonlinear kernel methods will increase the use of nonlinear methods and thus potentially increase the sensitivity to functional events that may otherwise be too nonlinear to be detected by linear methods.

Supplementary data to this article can be found online at doi:10.1016/j.neuroimage.2010.12.035.

Acknowledgments

This work is partly supported by the Danish Lundbeck Foundation through the program www.cimbi.org. We thank Stephen C. Strother, The Rotman Research Institute, Baycrest and Department of Medical Biophysics, University of Toronto, Canada and Tanya Schmah, Department of Computer Science, University of Toronto, Canada, for discussion and access to their preprint (Schmah et al., 2010); and the reviewers for many suggestions that improved the paper.

Appendix A

The following piece of code shows how to calculate the sensitivity map for a classifier with an RBF kernel. Let K be the $(N\times N)$ training kernel matrix, alpha a $(1\times N)$ vector with model coefficients, and X be a $(P\times N)$ matrix with training examples in columns. In “Matlab notation” the sensitivity map is calculated as

```
map=X*diag(alpha)*K-X*diag(alpha*K);
s=sum(map.*map,2)/numel(alpha);
```

where s is a $(P\times 1)$ vector with estimates of voxel sensitivities as elements. Note that we have omitted the kernel width here.

Appendix B

In the following we give a brief description of the classifiers used in the present study.

B.1. Kernel logistic regression

In binary classification logistic regression models the probability distribution of the labels given a feature vector as

$$p(t|\mathbf{x},\theta) = \sigma\left(t\left(\mathbf{w}^{\top}\phi(\mathbf{x}) + b\right)\right) \tag{10}$$

where $\sigma(a) = 1/(1+\exp(-a))$ is the sigmoid function. A common strategy is to find model parameters that minimize the regularized log-likelihood function (Cessie and Houwelingen, 1992)

$$L(\mathbf{w},\mathbf{b}) = \sum_{n=1}^{N}\log\left\{1+\exp\left(-\mathbf{t}_{\mathbf{n}}\left(\mathbf{w}^{\top}\phi(\mathbf{x}_{\mathbf{n}})+\mathbf{b}\right)\right)\right\} + \frac{\lambda}{2}||\mathbf{w}||_2^2, \tag{11}$$

where the ℓ_2 penalty term controls for over-fitting to the training data. The regularization parameter λ controls the amount of shrinkage of the norm of $\mathbf{w}$. By the so-called Representer theorem (Kimeldorf and Wahba, 1971; Cawley and Talbot, 2004) the solution to the optimization problem (11) can be written as a linear combination of the training points $\mathbf{w} = \sum_{n=1}^{N}\alpha_n\phi(\mathbf{x}_n)$. Hence, the cost function may be expressed in terms of the kernel function by

$$L(\alpha,\mathbf{b}) = \sum_{n=1}^{N}\log\left\{1+\exp\left(-\mathbf{t}_{\mathbf{n}}\left(\alpha^{\top}\mathbf{k}_{\mathbf{n}}+\mathbf{b}\right)\right)\right\} + \frac{\lambda}{2}\alpha^{\top}\mathbf{K}\alpha, \tag{12}$$

where $\mathbf{k}_n$ is the n'th column of the so-called Gram matrix $\mathbf{K}$ with elements $K_{i,j} = k\left(\mathbf{x}_i,\mathbf{x}_j\right)$. This is a convex optimization problem, and we fit the model parameters via the Newton–Raphson algorithm.

B.2. Support vector machine

The SVM classification scheme (Vapnik et al., 1995) seeks to establish an optimal separating hyperplane — optimal in the sense that the distance from the closest members of either classes to the plane is maximized (maximization of the margin). Model estimation involves minimizing the following cost function

$$L(\mathbf{w}, b) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n, \tag{13}$$

subject to the constraints

$$t_n\left(\mathbf{w}^\top\phi(\mathbf{x_n}) + \mathbf{b}\right) \geq 1 - \xi_\mathrm{n}, \quad \mathbf{n} = 1, \ldots, \mathbf{N}. \tag{14}$$

ξ_n are non-negative "slack variables" that allow points to be within the margin or on the wrong side of the decision boundary. The hyperparameter C controls the trade-off between margin maximization and slack variable penalty. The solution of Eq. (13) is unique and has the form $\mathbf{w} = \sum_{n=1}^{N}\beta_n t_n \phi(\mathbf{x}_n)$, where $\beta \geq 0$. Typically, only a subset of the training data $\mathbf{x}_n$ will have $\beta_n > 0$. Such points are called support vectors. Hence, the SVM solution will be sparse in the scan dimension if the solution only depends on a subset of the training examples. The decision function is expressed in terms of the kernel function as in Eq. (3) with $\alpha_n = \beta_n t_n$. For further introduction to the SVM we refer to Burges (1998). To fit the model parameters we used the LIBSVM software package (Chang and Lin, 2001).

B.3. Kernel Fisher discriminant

Kernel Fisher discriminant (KFD) analysis is a supervised dimensionality reduction technique for two-class classification problems (Mika et al., 2000). KFD seeks to find an optimal projection direction along which the ratio of the between-class scatter to the within-class scatter is maximized. The Fisher discriminant is given by the vector $\mathbf{w}$ that optimizes the objective function

$$L(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{S}_\mathrm{B}\mathbf{w}}{\mathbf{w}^\top(\mathbf{S}_\mathrm{W} + \lambda\mathbf{I})\mathbf{w}} \tag{15}$$

where $\mathbf{S}_\mathrm{B} = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^\top$ is the between-class scatter matrix, $\mathbf{S}_\mathrm{W} = \sum_\mathrm{k}\sum_{\mathbf{x}_i \in C_\mathrm{k}}(\mathbf{x}_i - \mathbf{m}_\mathrm{k})(\mathbf{x}_i - \mathbf{m}_\mathrm{k})^\top$ is the within-class scatter matrix, and λ is a regularization parameter. Here $\mathbf{m}_k$ are class means. Since the solution of Eq. (15) lies in the span of the training examples we can re-express the objective function in terms of the kernel function by (Mika et al., 2000)

$$L(\alpha) = \frac{\alpha^\top \mathbf{M}\alpha}{\alpha^\top(\mathbf{N} + \lambda\mathbf{K})\alpha} \tag{16}$$

with $\mathbf{M} = (\mathrm{u}_1 - \mathrm{u}_2)(\mathrm{u}_1 - \mathrm{u}_2)^\top$, $\boldsymbol{\mu}_k^{(i)} = \frac{1}{N_k}\sum_{j \in C_k} k\left(\mathbf{x}_i, \mathbf{x}_j\right)$, and $\mathbf{N} = \mathbf{K}\mathbf{K}^\top - \sum_\mathrm{k}\mathrm{N}_\mathrm{k}\mathrm{u}_\mathrm{k}\mathrm{u}_\mathrm{k}^\top$. Here $\boldsymbol{\mu}^{(i)}$ denote the i'th element in $\boldsymbol{\mu}$, and N_k denote the number of members in class k. The optimization problem can be solved as a generalized eigenvalue problem (Bie et al., 2005). We used the algorithm developed by Zhang et al. (2009). For further implementation details we refer to this paper.

References

Baehrens, D., Schroeter, T., Harmeling, S., Kawanab, M., Hansen, K., Müller, K.-R., 2010. How to explain individual classification decisions. J. Mach. Learn. Res. 11, 1803–1831.

Bie, T.D., Cristianini, N., Rosipal, R., 2005. Eigenproblems in pattern recognition. Handbook of Geometric Computing: Applications in Pattern Recognition, Computer Vision, Neuralcomputing, and Robotics, pp. 129–170.

Burges, C., 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. 2 (2), 121–167.

Cawley, G., Talbot, N., 2004. Efficient model selection for kernel logistic regression. Pattern Recognit. 2, 439–442.

Cessie, S.L., Houwelingen, J.V., 1992. Ridge estimators in logistic regression. Appl. Stat. 41 (1), 191–201.

Chang, C.-C., Lin, C.-J., 2001. LIBSVM: a Library for Support Vector Machines. http://www.csie.ntu.edu.tw/cjlin/libsvm.

Chen, X., Pereira, F., Lee, W., Strother, S., Mitchell, T., 2006. Exploring predictive and reproducible modeling with the single-subject FIAC dataset. Hum. Brain Mapp. 27 (5), 452–461.

Cox, D.D., 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, 261–270.

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

Formisano, E., De Martino, F., Bonte, M., Goebel, R., 2008a. "Who" is saying "What"? Brain-based decoding of human voice and speech. Science 322, 970–973.

Formisano, E., Martino, F.D., Valente, G., 2008b. Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning. Magn. Reson. Imaging 26, 921–934.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.P., Frith, C.D., Frackowiak, R.S.J., 1995. Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp. 2, 189–210.

Golland, P., Grimson, W.E.L., Shenton, M.E., Kikinis, R., 2005. Detection and analysis of statistical differences in anatomical shape. Med. Image Anal. 9, 69–86.

Grosenick, L., Greer, S., Knutson, B., 2008. Interpretable classifiers for FMRI improve prediction of purchases. IEEE Trans. Neural Syst. Rehabil. Eng. 16 (6), 539–548.

Guyon, I., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Mach. Learn. 46 (1), 389–422.

Hansen, L., Larsen, J., Nielsen, F., Strother, S.E., 1999. Generalizable patterns in neuroimaging: how many principal components? Neuroimage 9, 534–544.

Hansen, L., Nielsen, F., Strother, S., Lange, N., 2001. Consensus inference in neuroimaging. Neuroimage 13, 1212–1218.

Hansen, L.K., 2007. Multivariate strategies in functional magnetic resonance imaging. Brain Lang. 102 (2), 186–191.

Hanson, S.J., Halchenko, Y.O., 2008. Brain reading using full brain support vector machines for object recognition: there is no "face" identification area. Neural Comput. 20 (2), 486–503.

Hanson, S.J., Matsuka, T., Haxby, J.V., 2004. Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a "face" area? Neuroimage 23 (1), 156–166.

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

Kimeldorf, G., Wahba, G., 1971. Some results on Tchebycheffian spline functions. J. Math. Anal. Appl. 3, 82–95.

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

Klöppel, S., Stonnington, C., Chu, C., Draganski, B., Scahill, R., Rohrer, J., Fox, N., Jack, C., Ashburner, J., Frackowiak, R., 2008. Automatic classification of MR scans in Alzheimer's disease. Brain 131 (3), 681–689.

Koutsouleris, N., Meisenzahl, E.M., Davatzikos, C., Bottlender, R., Frodl, T., Scheuerecker, J., Schmitt, G., Zetzsche, T., Decker, P., Reiser, M., Möller, H.-J., Gaser, C., 2009. Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. Arch. Gen. Psychiatry 66 (7), 700–712.

Kwok, J., Tsang, I., 2004. The pre-image problem in kernel methods. IEEE Trans. Neural Netw. 15 (6), 1517–1525.

LaConte, S., Peltier, S., Hu, X., 2007. Real-time fMRI using brain-state classification. Hum. Brain Mapp. 28 (10), 1033–1044.

LaConte, S., Strother, S., Cherkassky, V., Anderson, J., 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., 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, pp. 137–148.

Martino, F.D., Valente, G., Staeren, N.L., Ashburner, J., Goebel, R., Formisano, E., 2008. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. Neuroimage 43, 44–58.

Mika, S., Rätsch, G., Schölkopf, B., Smola, A., Weston, J., Müller, K.-R., 2000. Invariant feature extraction and classification in kernel spaces. Adv. Neural Inf. Process. Syst. 12, 526–532.

Mika, S., Schölkopf, B., Smola, A., Müller, K., Scholz, M., Rätsch, G., 1999. Kernel PCA and de-noising in feature spaces. Advances in neural information processing systems 11 (1), 536–542.

Misaki, M., Kim, Y., Bandettini, P., Kriegeskorte, N., May 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., 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, pp. 259–270.

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

Mourão Miranda, J., Ecker, C., Sato, J.R., Brammer, M., 2008. Dynamic changes in the mental rotation network revealed by pattern recognition analysis of fMRI data. J. Cogn. Neurosci. 21 (5), 890–904.

Mourão Miranda, J., Friston, K., Brammer, M., 2007. Dynamic discrimination analysis: a spatial–temporal SVM. Neuroimage 36 (1), 88–99.

Mourão Miranda, J., Reynaud, E., McGlone, F., Calvert, G., Brammer, M., 2006. The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. Neuroimage 33 (4), 1055–1065.

Ni, Y., Chu, C., Saunders, C.J., Ashburner, J., 2008. Kernel methods for fMRI pattern prediction. Proc. IEEE International Joint Conference on Neural Networks (IJCNN), pp. 692–697.

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

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

Pereira, F., Botvinick, M., in press. Information mapping with pattern classifiers: A comparative study. NeuroImage. doi:10.1016/j.neuroimage.2010.05.026.

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

Rakotomamonjy, A., 2003. Variable selection using SVM based criteria. J. Mach. Learn. Res. 3, 1357–1370.

Sato, J., Fujita, A., Mourão Miranda, J., Thomaz, C., Martin, M., Brammer, M., Junior, E., 2009. Evaluating SVM and MLDA in the extraction of discriminant regions for mental state prediction. Neuroimage 46 (1), 105–114.

Schmah, T., Yourganov, G., Zemel, R., Hinton, G., Small, S., Strother, S., 2010. A comparison of classification methods for longitudinal fMRI studies. Neural Comput. 22, 2729–2762.

Shawe-Taylor, J., Cristianini, N., 2004. Kernel Methods for Pattern Analysis. Cambridge University Press.

Sigurdsson, S., Philipsen, P., Hansen, L., Larsen, J., Gniadecka, M., Wulf, H., 2004. Detection of skin cancer by classification of Raman spectra. IEEE Trans. Biomed. Eng. 51 (10), 1784–1793.

Stephan, K., Kasper, L., Harrison, L., Daunizeau, J., den Ouden, H., Breakspear, M., Friston, K., 2008. Nonlinear dynamic causal models for fMRI. Neuroimage 42 (2), 649–662.

Strother, S., Anderson, J., Hansen, L., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., Rottenberg, D., 2002. The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. Neuroimage 15 (4), 747–771.

Teixeira, A., Tomé, A., Stadlthanner, K., Lang, E., 2008. KPCA denoising and the pre-image problem revisited. Digital Signal Process. 18 (4), 568–580.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer Verlag, New York.

Wang, Z., Childress, A., Wang, J., Detre, J., 2007. Support vector machine learning-based fMRI data group analysis. Neuroimage 36 (4), 1139–1151.

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

Zurada, J., Malinowski, A., 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, vol. 6, pp. 447–450.

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