NeuroImage 96 (2014) 117–132

Contents lists available at ScienceDirect

NeuroImage

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

Pattern classification of fMRI data: Applications for analysis of spatially distributed cortical networks

Grigori Yourganov [a,*], Tanya Schmah [b], Nathan W. Churchill [b], Marc G. Berman [a], Cheryl L. Grady [b,c], Stephen C. Strother [b,d,e]

[a] Department of Psychology, University of South Carolina, Columbia, SC, USA
[b] Rotman Research Institute, Baycrest Centre, University of Toronto, Toronto, ON, Canada
[c] Department of Psychology, University of Toronto, Toronto, ON, Canada
[d] Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada
[e] Institute of Medical Science, University of Toronto, Toronto, ON, Canada

ARTICLE INFO

Article history:
Accepted 27 March 2014
Available online 4 April 2014

Keywords:
Multivariate pattern analysis
Classification
Regularization
Spatial maps
Principal component analysis (PCA)

ABSTRACT

The field of fMRI data analysis is rapidly growing in sophistication, particularly in the domain of multivariate pattern classification. However, the interaction between the properties of the analytical model and the parameters of the BOLD signal (e.g. signal magnitude, temporal variance and functional connectivity) is still an open problem. We addressed this problem by evaluating a set of pattern classification algorithms on simulated and experimental block-design fMRI data. The set of classifiers consisted of linear and quadratic discriminants, linear support vector machine, and linear and nonlinear Gaussian naive Bayes classifiers. For linear discriminant, we used two methods of regularization: principal component analysis, and ridge regularization. The classifiers were used (1) to classify the volumes according to the behavioral task that was performed by the subject, and (2) to construct spatial maps that indicated the relative contribution of each voxel to classification. Our evaluation metrics were: (1) accuracy of out-of-sample classification and (2) reproducibility of spatial maps. In simulated data sets, we performed an additional evaluation of spatial maps with ROC analysis. We varied the magnitude, temporal variance and connectivity of simulated fMRI signal and identified the optimal classifier for each simulated environment. Overall, the best performers were linear and quadratic discriminants (operating on principal components of the data matrix) and, in some rare situations, a nonlinear Gaussian naïve Bayes classifier. The results from the simulated data were supported by within-subject analysis of experimental fMRI data, collected in a study of aging. This is the first study that systematically characterizes interactions between analysis model and signal parameters (such as magnitude, variance and correlation) on the performance of pattern classifiers for fMRI.

Introduction

Analysis of fMRI data is a challenging task, because the signal of interest is typically weak, distributed among various spatial locations, and confounded by complex spatially correlated high-variance noise. Selecting the best method for data analysis is still an open question. A popular analysis approach is that of pattern classification, where brain patterns are examined to predict the behavioral task being performed when a given brain volume was acquired (see, for example, Haxby et al., 2001; Kamitani and Tong, 2005; Kjems et al., 2002; Mitchell et al., 2004; Morch et al., 1997; Strother et al., 2002; and review papers by Haynes and Rees, 2006; Norman et al., 2006; Pereira et al., 2009). This is frequently posed as a binary classification problem, where two classes of brain volumes correspond to two behavioral task conditions, and the classifier models attempt to predict the task conditions under which individual brain volumes were acquired. Classification of brain volumes is one way to investigate the differences in cortical recruitment between the two conditions.

A researcher who wants to perform classification of fMRI data faces the challenge of selecting a classification algorithm among a large group of methods, with varying properties and underlying assumptions. Some of these methods are probabilistic, i.e., they construct a probability distribution for each class, compute the likelihood of an fMRI volume belonging to each class, and assign the volume to the most probable class. For example, the multivariate Gaussian distribution is used to model data distributions for the linear discriminant (LD) and quadratic discriminant (QD) classification algorithms (see Seber, 2004, for a general introduction to linear and quadratic discriminants). Another group of classifiers is non-probabilistic: fMRI volumes are assigned to classes without constructing a probabilistic model. A popular example of non-probabilistic classifiers is support vector machines (SVMs; see Vapnik, 1995).

* Corresponding author at: Department of Psychology, University of South Carolina, 1512 Pendleton Street, Columbia, SC 29208, USA. Fax: +1 803 777 9558.
E-mail address: yourgano@mailbox.sc.edu (G. Yourganov).

http://dx.doi.org/10.1016/j.neuroimage.2014.03.074
1053-8119/

In addition to probabilistic and non-probabilistic classifiers, there is another distinguishing characteristic: whether the classifier is univariate or multivariate. Univariate classifiers treat fMRI voxels as mutually independent, whereas multivariate classifiers quantify the interactions between voxels. Given that the brain is a network of interacting cortical areas, the assumption of voxel independence does not hold well for fMRI data. Nevertheless, univariate classifiers, such as Gaussian Naïve Bayes method (GNB; see Mitchell et al., 2004; Schmah et al., 2010), are more constrained than multivariate ones such as LD and QD, and therefore are not as prone to overfitting.

Yet another distinction between classifiers is whether they are linear or non-linear. Linear classifiers separate the classes with a linear plane, and nonlinear classifiers define a more complex surface as the boundary separating the classes. Nonlinear classifiers (such as QD, or SVM with nonlinear kernels) are more flexible and potentially less biased than linear classifiers (such as LD or linear-kernel SVM), but they require larger training sets for robust classification, due to the larger number of degrees of freedom (Morch et al., 1997).

Several studies have compared the performance of classifiers on fMRI data (see, for example, Ku et al., 2008; Misaki et al., 2010; Schmah et al., 2010). Each of these studies concluded that linear multivariate methods such as LD and linear-kernel SVM, are more accurate than simpler methods such as GNB and K-nearest-neighbors. Two studies (LaConte et al., 2005; Misaki et al., 2010) have demonstrated that nonlinear kernels do not have any significant advantage over linear kernels when used in SVM classification; however, Schmah et al. (2010) demonstrate a case when SVMs with nonlinear kernels (radial basis functions and second-degree polynomial) outperform linear-kernel SVM. In general, the interaction of the classifier properties (e.g. linear/nonlinear, univariate/multivariate) with the properties of the BOLD signal (such as connectivity and contrast-to-noise ratio) is still not well understood. In the current paper, we investigate this problem by evaluating the performance of several classifiers on simulated fMRI data sets for varying signal parameters.

The studies described above evaluated the performance of classifiers on the accuracy of out-of-sample classification. However, accurate classification of data is usually not the sole purpose of an analysis; it is equally important to identify the cortical areas where the difference between the classes is localized, in order to interpret the neuronal basis underlying the cognitive contrasts. In other words, it is important to obtain a spatial map of classification, where voxels are weighted according to their contribution to classification. Kjems et al. (2002) have proposed a method of constructing "sensitivity maps" of voxels' contribution to classification. These maps have been developed for SVMs with linear (LaConte et al., 2005) and nonlinear (Rasmussen et al., 2011) kernels, as well as kernel logistic regression and kernel LD (Rasmussen et al., 2011). We have demonstrated a way to construct sensitivity maps for QD, and compared them with LD sensitivity maps (Yourganov et al., 2010). In the present paper, we propose a modification to the methodology of sensitivity map construction (discussed in "Constructing spatial maps for classifiers") and show how to construct spatial maps for linear and non-linear GNB.

Construction of spatial maps allows one to evaluate a classifier with a metric that is complementary to classification accuracy: the reproducibility of maps. This metric has been used in several neuroimaging studies (e.g., LaConte et al., 2003; Raemaekers et al., 2007; Strother et al., 1997, 2002; Tegeler et al., 1999). It has also been established that reproducibility is monotonically increasing with signal-to-noise ratio (LaConte et al., 2003), but the influence of functional connectivity on reproducibility of different classification models has not been studied systematically; we investigate this problem with simulation studies presented in the current paper. We measure within-subject reproducibility by computing the correlation between two spatial maps constructed from two independent samples drawn from the subject's data (Strother et al., 1997). Thresholding of maps is not required. This measure evaluates the stability of spatial locations where the neural effect of interest is expressed. Taken together, prediction and reproducibility are complementary metrics that reflect bias-variance tradeoffs in classifier model parameterization. Complex, flexible models tend to be more predictive but less reproducible (i.e. exhibiting greater variance of model parameters), whereas simpler models tend to be more reproducible but less predictive (i.e. more biased).

In the current paper, we compare a set of classification algorithms: LD, QD, linear SVM, and GNB (linear and non-linear), on both simulated and experimental fMRI data. We consider only within-subject classification tasks. We use the NPAIRS framework (Strother et al., 2002, 2010) to evaluate each classification algorithm using two metrics: out-of-sample classification accuracy, and reproducibility of spatial maps. Metrics are computed by splitting the data set into two independent subsets of approximately same size. The first subset is used to train the classifier, and the second subset serves as an independent test set to evaluate classification accuracy; then, these roles are switched. Reproducibility of spatial maps is computed by constructing the spatial maps (which indicate the relative contribution of each spatial location to classification) on the two subsets, and calculating the Pearson's correlation coefficient between the two spatial maps. Several such splits are performed, and the mean value of each of the two metrics is taken across the splits.

Previously, the NPAIRS framework was applied to evaluate the efficacy of pre-processing techniques (Churchill et al., 2012a, 2012b; LaConte et al., 2005; Strother et al., 2004). We show that this framework is also useful for evaluating and comparing different classifier models. Our paper presents two such evaluations: on a group of simulated fMRI datasets, and on experimental fMRI data collected in a study of age-related changes in cognitive ability. The main advantage of simulations is the knowledge of "ground truth", i.e. underlying parameters of active signal. We have focused on three parameters: mean magnitude and temporal variance of the signal, and connectivity across active areas; multiple datasets were constructed for different levels of these parameters, in order to systematically investigate their influence on performance of the classifiers. In experimental data, we do not have such knowledge of the "ground truth"; to partially compensate for this, we performed our evaluation on task contrasts of varying strength (a "strong" contrast is defined by a pair of dissimilar behavioral tasks that are known to recruit different cortical networks) as well on data coming from subjects belonging to different age groups.

The current study evaluates the classifiers on 6300 simulated datasets, in addition to 47 experimental fMRI datasets. Due the high computational burden of analyzing these large datasets, this paper focuses on classifier models that require a maximum of one tuning parameter. Alternative models with multiple parameters include SVMs with radial basis function (RBF) kernels, which have two parameters and therefore require grid optimization. SVMs with RBF kernels have been shown to accurately classify fMRI data in some, but not all, situations (Misaki et al., 2010; Schmah et al., 2010). Construction of spatial maps for RBF kernels has been described by Rasmussen et al (2011); evaluating them against spatial maps created for probabilistic classifiers is an interesting direction for future research. Another limitation in our analysis is our focus on binary classification problems (for evaluation of classifiers on multi-class problems, see Hassabis et al., 2009; Schrouff et al., 2012), an on within-subject classification.

Materials and methods

We compared classifiers on both simulated and experimental fMRI datasets; the experimental data were obtained from a battery of cognitive tasks, assessed in different age groups (Grady et al., 2010). The simulations were used to characterize classifiers as a function of different signal properties, including mean signal change, variance and network correlation. Similarly, we performed classification of experimental data for a range of within-subject binary problems, which implicate a variety of different brain networks, at different contrast strengths. For all analyses, we compared prediction and spatial reproducibility of the

different classifiers. In addition, we performed ROC analyses on the maps made for the simulated data, in order to compare the signal detection (measured with the area under the ROC curve) against the data-driven NPAIRS metrics. To account for high across-subject heterogeneity in the experimental data set, we performed some additional evaluation of classifiers: nonparametric testing of ranking for our metrics of prediction and reproducibility, and consensus analysis of the brain maps with DISTATIS (a multidimensional scaling procedure; Abdi et al., 2009).

Data

Simulated data

We utilized computer-generated data to simulate a blocked-design experiment with two conditions: *activation* and *baseline*. Full description of the simulation is given in earlier papers (the simulation algorithm was developed by Lukic et al., 2002; our modifications of the algorithm are described in Yourganov et al., 2011). Each simulated run consisted of 10 activation epochs and 10 baseline epochs, in an alternating order. All volumes consisted of a single slice (60 × 60 pixels). The volumes in the baseline condition were created by adding white Gaussian noise to the simplified "brain-like" background structure (the "background signal"); noise was spatially smoothed using a Gaussian filter with full width at half-maximum (FWHM) of 2 pixels. After smoothing, the standard deviation of the noise was 5% of the background signal. Images in the activation condition contained 16 Gaussian-shaped signal blobs distributed over the image and added to the smoothed noisy background image. The FWHM of the activation blobs (that we will call "active areas") varied between 2 and 4 pixels. Fig. 1 shows the background structure with and without activation, and Fig. 2 shows the composition of the timecourse of an active area (where the baseline epochs consist of noise, and the active epochs contain a mixture of active signal and noise). To simulate the hemodynamic response, each pixel's time course was convolved with a model hemodynamic response function (HRF) defined by the sum of two Gamma functions (Glover, 1999). Parameters of the HRF model have been taken from Worsley (2001), with TR (time to acquire the full brain volume) set to 2 s.

Amplitudes of the active areas were sampled from a multivariate Gaussian distribution. The mean amplitude of each activation was set proportional to the local value of the background signal:

$$E[a_k] = Mb_k; \tag{1}$$

where a_k is the amplitude of kth activation, $E[a_k]$ is its expected value, b_k is the value of noise-free baseline image at the center of the kth area, and M is the proportionality constant. To study the effect of M on performance of the algorithms, M was set to different levels (0, 0.02 and 0.03) in different realizations of our simulated experiment.

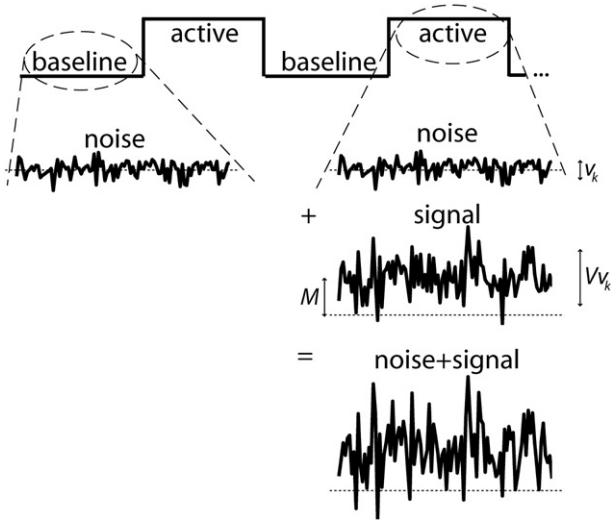

Fig. 2. Timecourse of the kth active area. The "baseline" epoch is composed of zero-mean Gaussian noise (with standard deviation v_k), and the "active" epoch contains a sum of baseline noise and Gaussian active signal. Mean and variance of the active signal are controlled by the parameters M and V, respectively.

The variance of the amplitude of the Gaussian activation signal in our multivariate Gaussian distribution, denoted by ${\sigma_k}^2$, was defined proportionally to the variance of the independent background Gaussian noise added to each pixel, ${v_k}^2$:

$${\sigma_k}^2 = V{v_k}^2, \tag{2}$$

where the proportionality constant V was varied from 0.1 to 1.6 (in increments of 0.25), in different realizations of the experiment, whereas v_k was kept fixed at 5% of the background signal. In this paper, we refer to V as the relative signal variance, which may be thought of as a form of physiological variation of the activation signal. The third parameter of our multivariate Gaussian model was the correlation coefficient, ρ, which defined the covariance between Gaussian activation signal amplitudes at the kth and lth locations ($k \neq l$):

$$\mathrm{cov}(a_k, a_l) = \rho\sigma_k\sigma_l. \tag{3}$$

The value of ρ was set to 0, and 0.5 and 0.99 to define a simple distributed spatial network (Lukic et al., 2002). This variable served to manipulate the connectivity between the areas of activation.

The amplitudes of the multivariate Gaussian signal in the "active" state are defined by the three parameters: M, V and ρ, which are the same for all active areas within a simulated data set. Of these parameters,

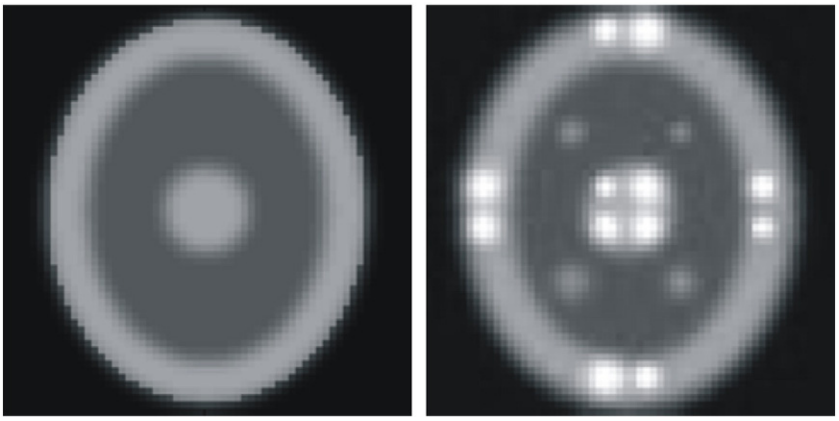

Fig. 1. Examples of simulated single-slice volumes in the baseline (left) and activation (right) conditions. Additive noise is not displayed.

M regulates the contrast-to-noise ratio, V regulates the dynamic range of the signal, and ρ defines connectivity. The details are described in Appendix A. For each active locus, M and V define the mean and the variance, respectively, of the Gaussian signal. For each setting of (M, V, ρ), we generated 100 data sets. Also, for ROC analysis, we generated an additional series of 100 "null" data sets that consisted entirely of baseline volumes.

When analyzing simulated data, the first 2 volumes of each epoch were discarded because of the delayed temporal response of HRF. This reduced the size of the data sets to 160 volumes. Voxels outside the "simulated brain" were discarded, leaving 2072 voxels for further analysis.

Experimental data

We also analyzed a set of experimental fMRI data that was collected by Grady et al. (2010). This study examined the impact of aging on cognitive abilities. Participants from two age groups (young, 20–31 years, 19 subjects; older, 56–85 years, 28 subjects) were scanned on a 3 T scanner (TR/TE = 2000/30 ms, flip angle = 70°, 28 axial slices, slice thickness 5 mm), during performance of several behavioral tasks, for four separate scanning runs. All four runs were acquired during the same session. We analyzed the epochs acquired during the execution of the following tasks:

1. *fixation* to a dot presented in the middle of the screen ("FIX");
2. *reaction task*: detection of a visual stimulus and reporting its position on the screen ("RT");
3. *perceptual matching*, where the participant had to match the "target" sample presented in the upper portion of the screen to one of the three stimuli presented in the lower portion ("PM");
4. *delayed matching* test of working memory, where the target stimulus was presented and then removed from the screen, followed by a 2.5 s blank-screen delay. After this, three stimuli were presented and the participant had to match them to the target ("DM").

During each experimental run, the fixation condition was presented in eight 20-second blocks. The other conditions were presented in blocks that were interleaved with the fixation blocks, each block lasting approximately 40 s (the duration varied slightly because the stimuli were generated at the time of scanning). There were 2 blocks per run for each of these three conditions, giving (8 blocks) × (20 scans) = 160 scans total per condition, on average.

Preprocessing was carried out in several steps. First, the transformation of aligning functional images to a common atlas was computed (details are described in Grady et al., 2010) but not immediately applied. Then the unaligned images underwent slice time correction (with AFNI package; Cox, 1996) and motion correction (with AIR package, Woods et al., 1998). Afterwards, the corrected images were transformed into a common anatomical space. Then the images were smoothed with a Gaussian kernel (FWHM = 7 mm), and artifact-carrying components were removed by using independent component analysis and performing manual identification of ICs with probable motion and physiological artifact (using MELODIC package; Beckmann and Smith, 2004). The white-matter signal (measured near the corpus callosum) was regressed from the time course of each voxel. The same was done for the cerebro-spinal fluid (CSF) signal, measured at the fourth ventricle. Finally, linear trends were removed. The voxels outside of the brain were masked out; additional voxels on the top of the brain were discarded because they showed high susceptibility to motion in several subjects. 21,401 voxels from the whole brain were retained for analysis.

This data set was used to evaluate the performance of several classifiers in a series of binary classification problems. Contrasts for each problem were defined by a pair of behavioral tasks. The volumes acquired during the corresponding blocks were classified according to the task performed. We performed our analysis on 4 contrasts: RT/FIX, DM/FIX, DM/RT and DM/PM. For each contrast, we ensured that the number of volumes was the same in both classes. After subsampling from the larger class, the number of volumes per class varied between 144 and 179 across subjects.

Performance metrics

Out-of-sample classification accuracy

A natural way to evaluate the performance of a classifier is to estimate how often it correctly classifies out-of-sample data, i.e. the data that are not included in the training set. We follow a cross-validation approach (see e.g. Efron and Tibshirani, 1993), where the data are repeatedly split into training and test sets. For each split, we record the proportion of test set vectors that were classified correctly, and then compute the mean value across splits; this is our measure of *classification accuracy*. We use classification accuracy as our prediction metric, as it allows us to compare both probabilistic (e.g. LD) and non-probabilistic (e.g. SVM) classifiers in the same framework.

Our resampling scheme is based on the NPAIRS framework proposed by Strother and colleagues (see Strother et al., 2002, 2004, 2010), where the data set is repeatedly split into two sets of approximately equal size. For each split, a classifier is trained on each half separately, the other half serving as a test set to evaluate classification accuracy. Then, these roles are switched; and finally we take the mean of the two classification accuracies. Our metric of classification accuracy is the mean frequency, across splits, of correct assignments of test data.

To avoid a possible bias in our evaluation, we must ensure the independence of the training and test set (the two "split-halves"). We do this by designing our splitting algorithms to ensure a sufficient time separation between any two volumes from different split-halves. In simulated data, a minimum separation of 40 s is used, which guarantees the independence of the two split-halves, since the timecourse of the HRF kernel in our simulations is 20 s long. We create 20 training-test splits for each simulated data set. In the experimental data, there are 4 runs for each subject, with the time between runs being much longer than typical estimates for the latency of the BOLD signal, so it is sufficient to ensure that, for each split, all volumes from a given run are assigned to the same split-half. This strategy gives 3 possible training-test splits (2 runs in each half-split).

Reproducibility of spatial maps

Following Strother et al. (2002), we also evaluate our classifiers using a reproducibility metric for spatial maps. A spatial map can be expressed as a vector of voxel weights, of the same size as our data vectors, where each weight reflects the contribution of the corresponding voxel in classifying the data. The process of creating spatial maps for each classifier is described in "Constructing spatial maps for classifiers".

We use the split-half resampling framework and compute a spatial map for each half. We then compute the Pearson's correlation coefficient on the paired voxel values of the two maps. The mean value of correlations (across splits) is our measure of *spatial map reproducibility*. This metric is complementary to classification accuracy, as it measures the robustness of classifier's spatial maps; this is a property not captured by classification accuracy.

Area under ROC curve

When evaluating a classifier on our simulated data, we use another well-known performance metric: area under receiver-operating-characteristic (ROC) curve. Full details of our ROC analysis are described in an earlier paper (Yourganov et al., 2011); here we give a brief description. To build a ROC curve for a setting of (M, V, ρ), we construct 100 simulated data sets with two conditions ("activation" and "baseline"), which we call H_1 sets, and also 100 simulated data sets of the same length consisting of only the "baseline" volumes, which we call H_0 sets. In the H_0 sets, there is no active signal that is present exclusively in one class but not the other. The H_0 and H_1 sets are analyzed by a classifier algorithm, creating a spatial map for each set. In this map, a voxel

weight that is above a particular threshold indicates an "active" voxel. For the H_1 set, an "active" voxel in one of the 16 active areas is a true positive, as it is known to contain a mixture of active signal and noise. For H_0, an "active" voxel is a false positive, as these data contain only baseline noise. In each simulated active brain area, we use the 100 H_1 and H_0 datasets to measure the fraction of true positives and false positives (respectively). A ROC curve is a plot of false-positives frequency versus true-positives frequency, for all possible thresholds.

A ROC curve is constructed for the peak voxels at the centers of the 16 active areas in H_1. We vary the threshold from most conservative (when no voxels pass the threshold) to most liberal (when all voxels pass). For each threshold, we compute the frequency at which the voxel (active in H_1) surpasses this threshold. A range of thresholds gives us a set of (true positive frequency, false positive frequency) pairs. We use the LABROC software (Metz et al., 1998) to generate smooth ROC curves from a set of discrete (FPF, TPF) pairs. Area under this curve corresponds to the probability that the classifier will assign a higher value to a voxel from H_1 than to a voxel from H_0; it is proportional to Mann–Whitney *U* test (Mason and Graham, 2002), which is a non-parametric equivalent to Student's unpaired *t* test (Conover, 1999). Rather than examining the area under the whole curve, we examine the partial area that corresponds to false positive frequencies between 0 and 0.1; this is equivalent to setting the critical significance level (α) equal to 0.1 (Skudlarski et al., 1999).

An alternative approach to constructing ROC curves (e.g., Skudlarski et al., 1999) does not involve H_0 sets. The frequency of false and true positives is computed from spatial maps generated for simulated H_1 sets. Since we know which voxels are active in H_1 sets, the true positive frequency is computed from a ratio of active voxels where the voxel weight surpasses a given threshold. False negative frequency is the ratio of above-threshold voxels that are known *not* to be active. This technique works well for univariate analysis, because the voxels that are used to compute true positive frequency are independent from voxels that are used to compute false positive frequency. However, if a spatial map is created with a multivariate method, the weights in different spatial locations are not independent. Therefore, ROC analysis of spatial maps created by multivariate classifiers requires an independent group of H_0 sets to compute the false positive frequency.

As a performance metric, the area under the ROC curve is similar to reproducibility because both these metrics evaluate spatial maps. The critical difference is that reproducibility evaluates the consistency of spatial maps, whereas the ROC metric compares these maps against a "gold standard", utilizing the ground-truth knowledge of which voxels are truly active and which are not. Both metrics could be applied to simulated data, but application of ROC methodology to spatial maps obtained from experimental data is infeasible, due to the lack of this ground-truth knowledge. However, a combination of classification accuracy and reproducibility can, to a point, serve as an alternative to ROC methodology (LaConte et al., 2003; Strother et al., 2002); classification accuracy uses a different kind of "ground truth" (i.e., the class membership of volumes).

Classification algorithms

We use our evaluation framework to assess the performance of several classifiers that varied on the following aspects: probabilistic versus non-probabilistic, linear versus non-linear, and univariate versus multivariate. Our pool of classifiers is applied to binary classification problems (e.g. comparing tasks 1 and 2), where the class assignment of an fMRI volume $\mathbf{x}$ is determined by the sign of a *decision function* $D(\mathbf{x})$: if it is positive, the volume is assigned to task 1, and if it is negative, to task 2. Zeros of $D(\mathbf{x})$ correspond to ambiguous cases that are equally likely to be acquired during task 1 and 2. A surface where $D(\mathbf{x}) = 0$ is called the *decision boundary*. The goal of every classifier is to build $D(\mathbf{x})$, under a specific set of model assumptions.

Most classifiers in our pool are *probabilistic*, i.e. they assign the volume to the most probable class (e.g. experimental task condition). In this case, a decision function is typically defined as a logarithm of the ratio of posterior probabilities:

$$D(\mathbf{x}) = \log\frac{P(\mathbf{x}\in class1|\mathbf{x})}{P(\mathbf{x}\in class2|\mathbf{x})}. \tag{4}$$

If the prior probabilities of the two classes are equal ($P(\mathbf{x}\in class1) = P(\mathbf{x}\in class2)$), we can use Bayes' theorem to re-express the decision function as a logarithm of the likelihood ratio:

$$D(\mathbf{x}) = \log\frac{P(\mathbf{x}|\mathbf{x}\in class1)}{P(\mathbf{x}|\mathbf{x}\in class2)}. \tag{5}$$

Probabilistic classifiers construct a probability distribution to model each class, so the likelihood functions can be computed directly. We have also included one non-probabilistic classifier into our pool, support vector machine (SVM) with a linear kernel, which computes the decision function without building a probabilistic model for the classes.

Probabilistic Gaussian classification

The probabilistic classifiers in our pool use the multivariate Gaussian distribution to compute the likelihood function for both classes. *Quadratic discriminant* (QD) is the most general method. For two classes, the decision function is

$$D_{QD}(\mathbf{x}) = \frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T{\Sigma_1}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T{\Sigma_2}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2), \tag{6}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the mean vector and the covariance matrix of the *i*th class (throughout the paper, we use boldface lowercase, boldface uppercase, and italic letters to denote vectors, matrices and scalars, respectively). The QD decision function is a quadratic form in $\mathbf{x}$ (producing a quadric decision boundary).

Linear discriminant (LD) is a more constrained classifier: the distributions for the two classes are assumed to be homoscedastic, i.e. to share the same covariance matrix $\boldsymbol{\Sigma}$. The decision function of Eq. (6) simplifies to a linear form:

$$D_{\mathrm{LD}}(\mathbf{x}) = -\left(\mathbf{x}-\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1). \tag{7}$$

Gaussian Naïve Bayes (GNB) classifiers constrain the covariance matrices to be diagonal. This is equivalent to assuming that the dimensions of our data (that is, voxels) are independent of each other; therefore, GNB is a univariate classifier, whereas QD and LD are multivariate. We evaluate two versions of GNB: *linear GNB* (GNB-L) adds a further constraint of homoscedasticity for classes 1 and 2, whereas *nonlinear GNB* (GNB-N) allows the covariance matrices to differ across classes. Decision functions for linear and nonlinear GNB are given by:

$$D_{\text{GNB-L}}(\mathbf{x}) = \sum_{j=1}^{p}\frac{\left(x_j-\mu_{2j}\right)^2-\left(x_j-\mu_{1j}\right)^2}{2{\sigma_j}^2}; \tag{8}$$

$$D_{\text{GNB-N}}(\mathbf{x}) = \sum_{j=1}^{p}\log\frac{\sigma_{2j}}{\sigma_{1j}} - \sum_{j=1}^{p}\frac{\left(x_j-\mu_{1j}\right)^2}{2{\sigma_{1j}}^2} + \sum_{j=1}^{p}\frac{\left(x_j-\mu_{2j}\right)^2}{2{\sigma_{2j}}^2}. \tag{9}$$

Here, μ_{1j} and μ_{2j} denote the *j*th element of the class-specific mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$; σ_{1j} and σ_{2j} denote the *j*th diagonal element of the (diagonal) within-class covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$; *p* denotes the total number of voxels in the analysis and, finally, σ_j denotes the *j*th diagonal element of the common covariance matrix $\boldsymbol{\Sigma}$. It should be noted that GNB-L is very similar to the univariate general linear model (GLM).

Both methods use the assumptions of voxel independence, normality and homoscedasticity, to different ends: GNB-L predicts the class membership of a test volume, whereas GLM estimates the statistical significance of each voxel in training volumes. An important difference between these models is that GNB-L is performed directly on the binary task design (where transitional volumes at the beginning of each epoch are discarded), whereas GLM is typically performed after convolving the task design with an HRF.

In all the algorithms described above, we estimate population parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ with the unbiased maximum-likelihood estimators $\mathbf{m}_i$ and $\mathbf{S}_i$, computed on the training set:

$$\mathbf{m}_i = \frac{1}{N}\sum_{k=1}^{N_i}\mathbf{x}_k; \tag{10}$$

$$\mathbf{S}_i = \frac{1}{N-1}\sum_{k=1}^{N_i}(\mathbf{x}_k-\mathbf{m}_i)(\mathbf{x}_k-\mathbf{m}_i)^{\mathrm{T}}. \tag{11}$$

Here, N_i is the number of training volumes for the ith class. For the LD and linear GNB, the pooled covariance matrix $\mathbf{S}$ is the average of $\mathbf{S}_1$ and $\mathbf{S}_2$. Probabilistic classifiers, as well as the framework for their evaluation, were implemented in MATLAB.

Non-probabilistic classification: support vector machines

Support vector machines (SVM) are a popular class of non-probabilistic classifiers. They do not build a probabilistic model for the classes, but create the decision function in a way that simultaneously maximizes the margin between the two classes and minimizes the misclassification rate (Cortes and Vapnik, 1995). We have tested the simplest version of SVM that uses a linear kernel. The decision function for a linear-kernel SVM is linear in $\mathbf{x}$:

$$D_{\mathrm{SVM}}(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b, \tag{12}$$

where $\mathbf{w}$ is the normal of the optimal discriminant hyperplane and b is the bias term. The vector $\mathbf{w}$ and the scalar b are found by minimizing the expression $\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N_{\mathrm{train}}}\xi_n{}^2$, subject to the constraints:

$$t_n(\mathbf{w}\cdot\mathbf{x}_n) \geq 1-\xi_n, \tag{13}$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are the training volumes and t_n is 1 for volumes in class 1 and -1 for volumes in class 2. The problem of finding optimal set of $(\mathbf{w}, b, \xi_1, \ldots, \xi_N)$ has a unique solution, which can be found by quadratic programming. The variables $\xi_1, \ldots, \xi_N$ are called slack variables; ξ_i measures the degree of misclassification for vector $\mathbf{x}_i$. The quantity $2/\|\mathbf{w}\|$ is called the margin. The tradeoff hyperparameter C specifies the importance of accuracy of classification relative to maximizing the margin; higher values of C force the slack variables to be smaller. We used a MATLAB library LIBSVM (Chang and Lin, 2011) to compute the SVM model for a given value of C.

Constructing spatial maps for classifiers

A spatial map, computed for a given classifier, indicates the relative importance of different spatial locations in building the decision boundary. We propose to construct spatial maps by taking a voxel-wise partial derivative of the decision function $D(\mathbf{x})$; this derivative demonstrates how much the decision boundary is dependent upon a value of a given voxel. This is similar to the technique of "sensitivity maps" proposed by Kjems et al. (2002). The value of the ith voxel of the spatial map is computed as

$$y_i = \frac{1}{N}\sum_{j=1}^{N}\frac{\partial}{\partial x_i}D\left(\mathbf{x}^{(j)}\right), \tag{14}$$

where $\mathbf{x}^{(j)}$ is the jth volume, and N is the number of volumes. Essentially, we take the decision function for each volume, compute its partial derivative at a specific voxel location, and average it across all volumes. Therefore, y_i indicates the average impact of the ith voxel on the decision function, and reflects the importance of this voxel in classification.

The sign of $\frac{\partial}{\partial x_i}D(\mathbf{x})$ encodes the class *preference* of the ith voxel: it indicates whether the signal in that voxel should be increased or decreased in order to increase $D(\mathbf{x})$ (Rasmussen et al., 2012A). For a two-class problem, we can say that a positive value of $\frac{\partial}{\partial x_i}D(\mathbf{x})$ corresponds to a preference of the ith voxel for task A, and a negative value corresponds to a preference for task B. Therefore, signed sensitivity maps can be interpreted analogously to statistical parametric maps (Worsley, 2001), where the sign of the voxel indicates whether the contrast is expressed positively or negatively in that voxel.

Let us consider the case when the derivative of $D(\mathbf{x})$ can be expressed analytically as a function of $\mathbf{x}$: $\mathbf{d}(\mathbf{x}) = \frac{\partial}{\partial x_i}D(\mathbf{x})$. In this case, the spatial map is the average of the values of $\mathbf{d}(\mathbf{x}_k)$ computed across all training vectors $\mathbf{x}_k$:

$$\mathbf{y} = \frac{1}{N_{\mathrm{train}}}\sum_{k=1}^{N_{\mathrm{train}}}\mathbf{d}(\mathbf{x}_k). \tag{15}$$

For QD, taking the derivative of Eq. (6) is a linear form in $\mathbf{x}$:

$$\mathbf{d}_{\mathrm{QD}}(\mathbf{x}) = -\mathbf{S}_1{}^{-1}(\mathbf{x}-\mathbf{m}_1) + \mathbf{S}_2{}^{-1}(\mathbf{x}-\mathbf{m}_2). \tag{16}$$

Here, we substitute the maximum likelihood estimators $\mathbf{m}_i$ and $\mathbf{S}_i$ (given in Eqs. (10) and (11)) for the population parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$. To compute the spatial map, $\mathbf{d}_{\mathrm{QD}}(\mathbf{x}_k)$ is averaged over all training vectors $\mathbf{x}_k$. The decision function of LD is linear, therefore its derivative is a constant independent of $\mathbf{x}$:

$$\mathbf{d}_{\mathrm{LD}}(\mathbf{x}) = \mathbf{S}^{-1}(\mathbf{m}_2-\mathbf{m}_1); \tag{17}$$

so the LD map is equal to $\mathbf{d}_{\mathrm{LD}}$ and no averaging of $\mathbf{d}_{\mathrm{LD}}$ across the training set is needed.

Although it is possible to express the derivative of $D_{\mathrm{GNB\text{-}L}}(\mathbf{x})$ and $D_{\mathrm{GNB\text{-}N}}(\mathbf{x})$ as a function of vector $\mathbf{x}$, it is simpler to express it in terms of the jth voxel (in this model, the voxels are independent of one another). The jth voxel of the derivative of the GNB-L decision function in Eq. (8) is

$$d_{\mathrm{GNB\text{-}L}\,j}(\mathbf{x}) = \frac{m_{1j}-m_{2j}}{s_j{}^2}; \tag{18}$$

note the similarity to T statistic that is used to construct spatial parametric maps in univariate general linear model (Worsley, 2001). For GNB-N, the voxelwise derivative is

$$d_{\mathrm{GNB\text{-}N}\,j}(\mathbf{x}) = -\frac{x_j-m_{1j}}{s_{1j}{}^2} + \frac{x_j-m_{2j}}{s_{2j}{}^2}. \tag{19}$$

Finally, the spatial map for linear-kernel SVM is a constant independent of $\mathbf{x}$, similar to LD. It is given by a weighted sum of the support vectors (see LaConte et al., 2005, where other methods of creating spatial maps for SVM are also discussed).

Regularization of multivariate classifiers

The decision functions for LD and QD, given by Eqs. (6) and (7), require estimation of the inverse of the population covariance matrix $\boldsymbol{\Sigma}$ (or $\boldsymbol{\Sigma}_i$). For fMRI data, this is problematic because the number of observations (fMRI volumes) is usually much smaller than the number of dimensions (voxels), making the sample covariance matrix $\mathbf{S}$ rank-deficient and thus impossible to invert. This problem is overcome by regularizing $\mathbf{S}$, that is, by approximating $\mathbf{S}$ with a lower-rank, invertible matrix. We tested two approaches to regularization. The first approach is to approximate $\mathbf{S}$ by a subset of its highest-ranking principal components (the size of this subset corresponds to the amount of regularization: a smaller subset is equivalent to more highly regularized data).

For this purpose, we use a set consisting of (first, second, …, Kth) PCs, where K is less than the number of volumes in the training set. In an earlier paper (Yourganov et al., 2011), we examined different approaches for selecting K, and came to a conclusion that K should be selected in a cross-validation framework. The value of K that optimizes reproducibility of spatial maps computed using split-half resampling scheme is very close to the value that optimizes the area under the ROC curve, which could be viewed as the "true dimensionality" of the simulated data. Optimization of prediction accuracy produced less robust estimates of K but nevertheless was still sensitive to the underlying structure of simulated networks. Here, we select K by optimizing *both* reproducibility and classification accuracy. Following Zhang et al. (2008) and Rasmussen et al. (2012B), we use the NPAIRS framework to compute classification accuracy and reproducibility for a range of values of K, and compute the mean values of these two metrics across splits. We then select the value of K that minimizes the Euclidean distance from a perfect classifier performance of (reproducibility = 1, prediction = 1); this distance is given by

$$\Delta = \sqrt{(1-\textit{mean reproducibility})^2 + (1-\textit{mean class} \cdot \textit{accuracy})^2}. \quad (20)$$

A second approach to regularization uses ridge regression, as suggested by Kustra and Strother (2001). Here, instead of $\mathbf{S}^{-1}$ we compute $(\mathbf{S}+\lambda\mathbf{I})^{-1}$, where $\mathbf{I}$ is the identity matrix and λ is the hyper-parameter. An increased λ tends to produce an increasingly regularized covariance matrix. We select λ by optimizing the Δ metric, as described above. Two approaches to regularization give us two implementations of LD: LD-PC and LD-RR, which use PC and ridge regularization, respectively. For QD, only the first approach (PC regularization) was applied. The regularization hyper-parameter of support vector machines is the "slack variable" coefficient C. We select this parameter by minimization of the Δ metric. Finally, GNB in our implementation does not require regularization and therefore has no hyper-parameters.

Evaluation of performance on experimental data

To evaluate our classifiers on simulated data, we create 100 simulated "subjects" by sampling from a population defined by particular levels of M, V and ρ. Since all simulated "subjects" are sampled from the same population, they form a relatively homogeneous set. In contrast, we expect much less homogeneity across the subjects that have participated in the aging study. When evaluating performance on real experimental data, we have performed some additional analysis to account for inter-subject variability.

Evaluating performance of within-subject classification

We used the data from the aging study to perform within-subject classification of fMRI volumes, for different behavioral contrasts. We tested 4 binary contrasts: RT/FIX, DM/FIX, DM/RT, and DM/PM. For each classifier, the distribution of prediction and reproducibility values across subjects is visualized with box-and-whisker plots. Outliers are not shown in the plots.

We perform additional statistical testing to answer the question whether the ranking of classifiers is consistent across subjects. For this purpose, we employ a non-parametric statistical test described in Conover (1999; pages 369–373), and also in Demsar (2006) and in our earlier paper (Schmah et al., 2010). First, Friedman test is applied to test the null hypothesis that all methods perform equally well. If null hypothesis is rejected, we proceed to post-hoc testing, where we test for a significant difference in ranking between all pairs of classifiers. For each contrast separately, we compute the average rank of each classifier across subjects, as well as the critical distance (the minimum difference between average ranks of classifiers that are considered significantly different). Results of this evaluation are presented as a plot of average classifier rankings, on a scale of 1 to 6; a rank of 1 denotes the best performance and a rank of 6 is the worst. A bar on a ranking plot represents a particular contrast. The classifiers that are not found to be significantly different are linked together with a horizontal line under the bar. If the null hypothesis is rejected, the critical distance (CD) is shown beside the bar.

Normalizing individual spatial maps

In order to compare the spatial maps created by different classifiers for the same subject and contrast, the maps have to be normalized so the distribution of noise is matched across methods. This normalization is described in the NPAIRS literature (LaConte et al., 2003; Strother et al., 2002), where the resulting normalized maps are called "reproducible statistical parametric maps, Z-scored" (rSPM{Z}). The distribution of signal and noise is computed from the scatter plots of pairwise voxel values, from two split-half maps (which are divided by their respective standard deviations in order to bring them to the same scale). The scatter plot has a major axis along the line of identity and a minor axis perpendicular to it. Variation along the minor axis is due to the noise uncorrelated with signal of interest, and variation along the major axis contains a mixture of signal and noise. For two split-half maps $\mathbf{z}_1$ and $\mathbf{z}_2$, the projection of scatter-plot points onto the major and minor axes is $(\mathbf{z}_1+\mathbf{z}_2)/2$ and $(\mathbf{z}_1-\mathbf{z}_2)/2$, respectively (Strother et al., 2002). To obtain rSPM{Z}, we divide the projection onto the major axis by the standard deviation of the projection onto the minor axis. We repeat this procedure for all splits; each split gives us a scatter plot, and rSPM{Z} patterns are averaged across splits. In this way, we compute a mean rSPM{Z} for every individual subject analysis.

Consensus of individual maps across classifiers

We also investigate the question of similarity of spatial maps created by different classifiers. For this purpose, we use DISTATIS, a variant of multidimensional scaling which identifies the pattern of correlation between classifier maps that is most consistent across subjects. Full treatment of DISTATIS can be found in publications by Abdi et al. (2005, 2009).

The goal of multidimensional scaling is to find a low-dimensional representation of high-dimensional data, in such a way that the distances between low-dimensional representations of any two data points are good approximations to the distances between these points in the original high-dimensional space. In our study, the high-dimensional data are the spatial maps, and we define the distance between ith and jth map as $1-\rho_{ij}$, where ρ_{ij} is Pearson's correlation coefficient of ith and jth maps; thus we create the distance matrix between all possible pairs of datapoints. Multidimensional scaling finds a low-dimensional representation from the eigendecomposition of the distance matrix. DISTATIS is a generalization of this method for a set of distance matrices. This method combines the distance matrices into a single *compromise* matrix, and projects the original distance matrices onto the compromise matrix.

In order to apply DISTATIS, we compute within-subject distance matrices. Our pool of classifiers consists of six methods, so we compute a 6×6 distance matrix for each of the 47 subjects. We double-center these matrices; let $\mathbf{S}_i$ denote the doubly-centered distance matrix for the ith subject. We then compute the similarities between distance matrices for each pair of subjects. The similarities are computed with an R_V coefficient (Robert and Escoufier, 1976), which indicates how much information is shared between two matrices (Abdi et al., 2009). We form the 47×47 matrix of R_V coefficients, and compute its first eigenvector $\mathbf{p}_1$. The ith coordinate of this eigenvector indicates how similar the ith subject's distance matrix is to all other subjects' distance matrices. Then, the compromise matrix $\mathbf{S}_+$ is formed as a weighted sum of doubly-centered distance matrices:

$$\mathbf{S}_+ = \sum_{i=1}^{\#\text{subjects}} \alpha_i \mathbf{S}_i \quad (21)$$

where the weight α_i is the ith coordinate of $\mathbf{p}_1$, divided by the sum of all coordinates of $\mathbf{p}_1$. The compromise matrix $\mathbf{S}_+$ is the best way (in the least-squares sense) to represent all 47 distance matrices with a single matrix.

A low-dimensional representation of $\mathbf{S}_+$ is computed from its eigendecomposition. For easy visualization, it is common to use 2-dimensional representation. We project the centroids of the spatial maps created by each of the six methods onto the space defined by the first two principal components of $\mathbf{S}_+$. In order to compute the confidence intervals around the centroids, we have drawn 1000 bootstrap samples from our set of 47 subjects. On the DISTATIS plots, the projections of centroids are marked with a +, and the confidence intervals are shown as ellipses around the centroids.

Results

We have evaluated the following pool of classifiers:

1. QD: quadratic discriminant with PC regularization of covariance matrices (the same number of principal components is used to regularize both classes);
2. LD-PC: linear discriminant with PC regularization of the pooled covariance matrix;
3. LD-RR: linear discriminant with ridge regularization of the pooled covariance matrix;
4. SVM: support vector machine with a linear kernel;
5. GNB-N: nonlinear Gaussian Naïve Bayes classifier, where the (diagonal) covariance matrices are allowed to differ across classes and the decision function is nonlinear;
6. GNB-L: linear GNB classifier, where the covariance matrix is pooled across classes and the decision function is linear.

Performance on simulated data

Our pool of classifiers (QD, LD-PC, LD-RR, SVM, GNB-L, GNB-N) was applied to analyze simulated fMRI data. The performance of each classifier was evaluated with three metrics: accuracy in classifying fMRI volumes according to the task, reproducibility of the classifier maps, and the area under the ROC curve. The results of this evaluation are presented in Figs. 3 and 4. Fig. 3 displays the performance metrics of classification accuracy (top) and map reproducibility (bottom), and Fig. 4 displays the performance as measured by the partial area under the ROC curve. The lines in Fig. 3 show mean performance, taken across 100 data sets generated for each setting of simulation parameters (M, V, ρ); the error bars show standard deviation of performance. For the ROC metric shown in Fig. 4, we computed a curve for each of the 16 active areas; the lines show the mean value of partial ROC area across the 16 areas, and error bars show its standard deviation across these areas. The black dashed line in Fig. 4 indicates the partial ROC area of 0.05, which corresponds to random guessing.

The columns of Figs. 3 and 4 correspond to the levels of mean signal change M (0, 0.02 and 0.03). The three levels of network correlation ρ (0, 0.5 and 0.99) are shown as three sub-panels of each level of M. Each of these sub-panels consists of a performance plot, where the horizontal axis represents signal variance V (relative to the noise variance), going from 0.1 to 1.6. The vertical axis of the plot is the mean magnitude of the performance metric.

Classification accuracy

In terms of classification accuracy (top row of Fig. 3), there are strong similarities between LD-PC, LD-RR, SVM and GNB-L. These methods can all be referred to as "linear classifiers", because they all use a linear decision function to compute class membership, and the decision boundary that separates the two classes is a hyperplane. Their accuracy is lowest when $M = 0$, and tends to increase as mean signal strength M grows. For $M = 0.03$, we see a negative effect of increasing V, which is modulated by ρ (it is weakest when $\rho = 0$, and strongest when $\rho = 0.99$). Since V is the variance of the active signal, its growth causes an increase in overlap between the "active" and the "baseline" classes, making discrimination between the two classes more difficult.

The remaining classifiers in our pool, QD and GNB-N, show quite different trends in performance for $M = 0$ (different from the linear classifiers as well as from each other); although the overall positive influence of M is also observed, the effect of V and ρ is more complex. These methods can use the difference in the covariance matrices to their advantage, which is particularly helpful when the separation between the class centroids is small (that is, M is low). For $M = 0$, both nonlinear methods get better as V increases. When $\rho = 0$, the functional nodes of the active network are independent, and GNB-N is the best model for our data, achieving 60% mean accuracy at the largest setting of V. On the other hand, when the functional nodes are strongly correlated ($\rho = 0.99$), QD is the most accurate classifier, peaking at 66% when V reaches 1.6. For the intermediate level of correlation ($\rho = 0.5$), QD and GNB-N perform at the similar level.

When $M > 0$, the two nonlinear classifiers perform similarly to the linear classifiers. When M is 0.03, performance is high (greater than 70%), but the influence of V is detrimental to performance, due to growing overlap between the classes. When $M = 0.02$, V does not have a strong impact on linear as well as nonlinear classifiers.

Reproducibility of spatial maps

The bottom row of Fig. 3 shows the reproducibility of spatial maps produced by our pool of classifiers. If we compare it with the classification accuracy plot, we see that the classifiers here can be grouped in a different way:

1. univariate methods: two versions of GNB
2. multivariate methods that use PC regularization: LD-PC and QD
3. other multivariate methods: SVM and LD-RR.

Inside each group, reproducibility is quite similar, but the groups are clearly distinct in most cases. We see the same pattern for all levels of M:

- As expected, network coupling (ρ) has no effect on univariate methods. There is a slight detrimental effect of V, which is noticeable at high levels of M. In most cases, univariate maps are less reproducible than multivariate maps. It is interesting to note here that pooling of variance across classes has no effect on reproducibility, because performance of GNB-N and GNB-L is the same.
- PC-based methods (LD-PC and QD) get a tremendous boost from increasing V, when the active areas are coupled ($\rho > 0$). The reproducibility of PC-based methods increases approximately linearly with V; for sufficiently large levels of V, reproducibility of these two methods greatly surpasses reproducibility of all other methods in our pool.
- Other multivariate methods (SVM and LD-RR) are not influenced by V and ρ, and have effectively identical performance. In terms of relative ranking, they perform better than univariate methods, but, in most cases, worse than PC-based multivariate methods.

This ranking of the methods is largely consistent across M. Overall, M has a positive effect on reproducibility: for all methods, spatial maps become more reproducible as M grows.

Partial area under the ROC curve

Fig. 4 shows the performance of our pool of algorithms measured with partial area under the ROC curve. We see that the classifiers group in the same fashion with respect to ROC area as they do with respect to spatial map reproducibility:

1. univariate methods;
2. PC-based multivariate methods;
3. other multivariate methods (SVM and LD-RR).

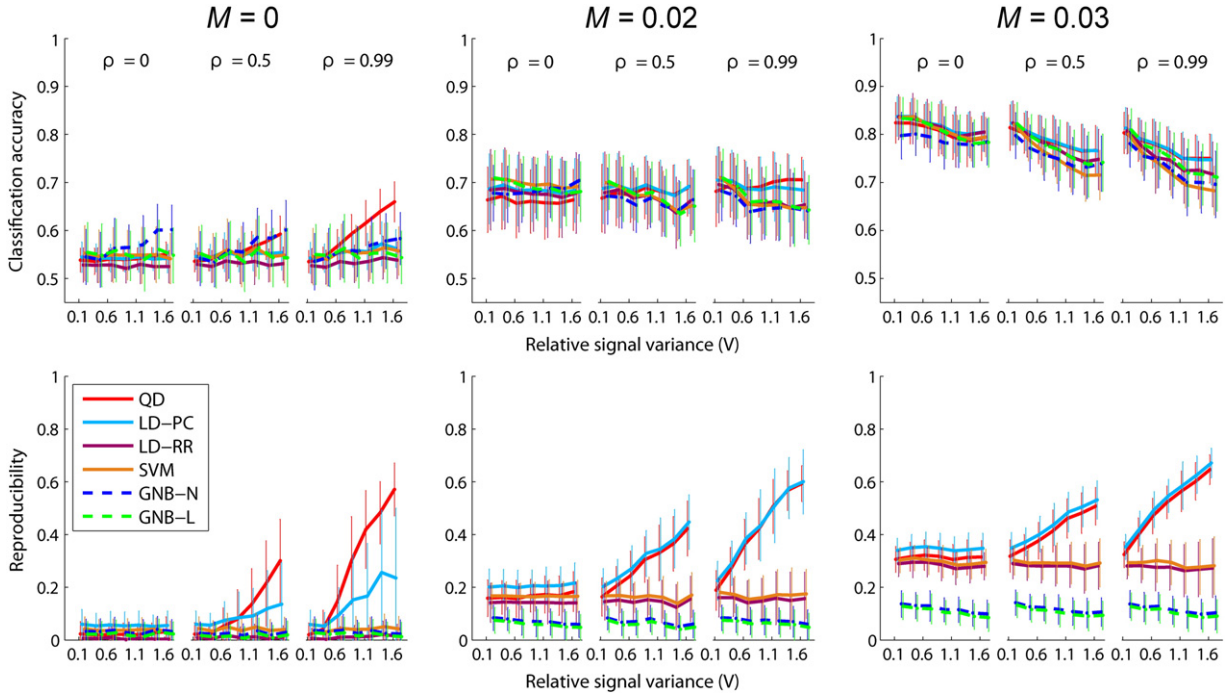

Fig. 3. Performance of the pool of six classifiers on simulated data sets. Performance is measured with classification accuracy (top row) and reproducibility of spatial maps (bottom row). The three columns correspond to three levels of mean signal magnitude M, and the three sub-columns correspond to three levels of spatial correlation ρ. Relative temporal variance of the active signal (V) is plotted on the x-axis. The colored lines correspond to the mean performance of six classifiers across 100 simulated datasets, and the error bars indicate the standard deviation. The classifiers are: quadratic discriminant with PC regularization (QD), linear discriminant with PCA regularization (LD-PC) and with ridge regularization (LD-RR), linear-kernel SVM, and linear and nonlinear Gaussian naive Bayes classifiers (GNB-L, GNB-N).

The second group is the best performer in terms of ROC area (as is true for reproducibility in Fig. 3). Both methods are sensitive to V, and their performance increases as V grows from 0 to 1. When mean signal is relatively strong ($M = 0.03$), both LD-PC and QD are near-perfect in their signal detection (partial ROC area approaches the theoretical maximum value of 0.1), for all levels of V and ρ.

Univariate methods are uniformly the worst performers. In the absence of mean signal, they never rise significantly above chance. When $M > 0$, they are much better than chance, but their performance drops as V grows. This decline is less severe when M is large. As expected, ρ has no effect on performance of univariate detectors. The third group of algorithms (SVM and LD-RR) is intermediate in terms of performance: better than univariate methods, but never as good as PC-based multivariate methods.

The black curve in Fig. 4 shows the ROC performance of a univariate GLM. The GLM maps (Worsley, 2001) were created by performing voxelwise T tests on the beta weights for the two epochs (active and baseline); the beta weights were computed with the same HRF kernel that had been used to create the simulated data. This figure shows that GLM is a more powerful signal detector than GNB, which is not surprising, because GLM utilizes the "true" HRF to model the transitions between the epochs, whereas GNB (like all other classifiers in our pool) discards the transition scans. With respect to the influence of M and V, we observe the same trends in GLM as in GNB: the effect of M is beneficial, and the effect of V is detrimental (ρ has no effect on univariate GLM). These trends are expected, because the magnitude of the T statistic is proportional to the contrast magnitude (controlled by M) and inversely proportional to the sample variance (controlled by V).

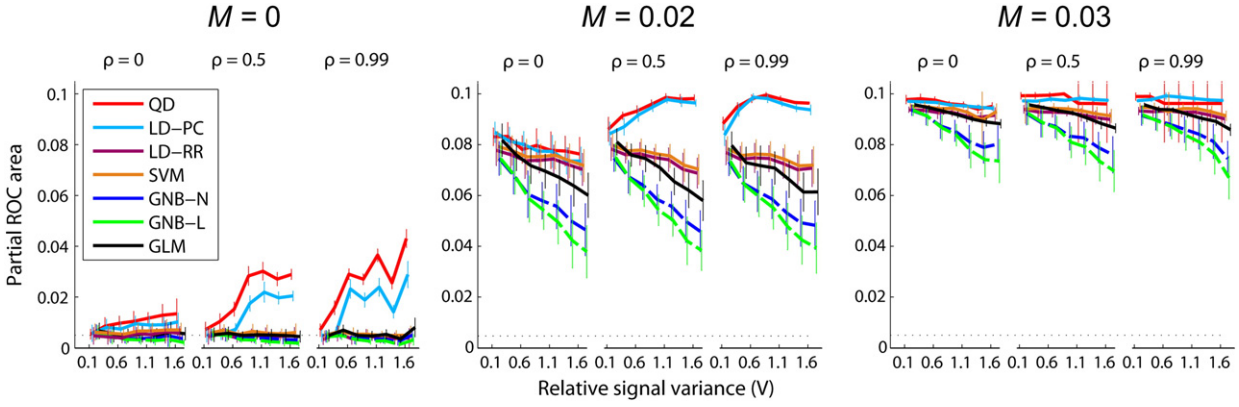

Fig. 4. Performance of the pool of six classifiers on simulated data sets, measured by partial area under the ROC curve corresponding to false positive frequency from 0 to 0.1. The three columns correspond to three levels of mean signal magnitude M, and the three sub-columns correspond to three levels of spatial correlation ρ. Relative temporal variance of the active signal (V) is plotted on the x-axis. ROC curves are measured at the centers of each of the 16 active loci. Colored lines show, for each classifier, the mean partial ROC area across the 16 loci, and the error bars show the corresponding standard deviation. Dashed black lines indicate the partial ROC area for random guessing.

Performance on experimental data

Within-subject reproducibility and classification accuracy

Fig. 5 displays the within-subject classification accuracy and reproducibility on a box-and-whisker plot; two age groups are plotted separately (left and right columns correspond to young and older subjects). In terms of classification accuracy, the difference between the classifiers is not significant for a given contrast, which is also observed in simulations for $M > 0$. However, there is a difference between contrasts, with DM/FIX and RT/FIX being the easiest to classify, and DM/PM the hardest. We will therefore refer to the DM/FIX and RT/FIX as the "strong contrasts", and the DM/RT and DM/PM as the "weak contrasts". With respect to reproducibility, the grouping of classifiers is similar to the grouping we see in the simulations: (1) QD and LD-PC, (2) SVM and LD-RR, and (3) GNB-L and GNB-N. The second and third groups are quite similar to each other, particularly in the older group.

Comparing the performance of our classifier pool across the age groups, we can see that the classifiers frequently perform better on the older subjects. We have evaluated the significance of the age-related difference in accuracy with a Mann–Whitney *U* test, for all task contrasts and classifier models. To account for multiple comparisons, we have used false discovery rate correction with significance level $\alpha = 0.05$ (Genovese et al., 2002). We observed significantly increased classification accuracy for the older group in the DM/FIX contrast, for LD-PC, LD-RR and SVM; and in the RT/FIX contrast, for LD-PC, LD-RR and QD. Reproducibility of maps was not significantly different between the two age groups.

In order to test whether classifiers' performance were statistically distinguishable, we examined the ranking of classifiers, and performed post-hoc Friedman tests; results are shown in Figs. 6 (for the young group) and 7 (for the older group). In the younger group, there is a large overlap in classifier accuracy, particularly for the strong contrasts; in the two weaker contrasts, the univariate classifiers have the lowest rank. In the older group, the multivariate linear classifiers (LD-RR, LD-PC and SVM) tend to be the highest-ranking with respect to their accuracy, although they are not always significantly different in ranking from the other classifiers. With respect to reproducibility of maps, the rank of algorithms displayed in Figs. 6 and 7 corresponds to the trend observed in simulations. QD and LD-PC have the highest rank. They are followed by LD-RR and SVM (SVM tending to rank higher than LD-RR). Finally, both versions of GNB receive the lowest ranking.

It is important to note that high within-subject reproducibility, as displayed in Fig. 5, does not translate into high across-subject reproducibility. If the heterogeneity across subjects is large, it is possible that the maps generalize poorly across subjects, despite being highly reproducible within a subject. This is the case of the LD-PC and QD maps for the two weaker contrasts, DM/RT and DM/PM. Supplementary Fig. 1 shows the across-subject reproducibility of maps created by the six classifiers; it was computed as the Pearson's correlation across all possible pairings of subjects within an age group. For the two weak contrasts, across-subject reproducibility is lower than for the two strong contrasts, and for the weakest contrast (DM/PM) it is near zero on average.

Consensus of spatial maps across classifiers

We next examined the question of similarity of subjects' spatial maps *across methods*. For each contrast and method, we have computed the across-subject average rSPM{Z} maps, and correlated them across classifiers; the correlations are given in Supplementary Tables 1–4. The across-method correlation between the average maps was at least 0.63 for the weakest contrast (DM/PM), and at least 0.83 for the other three contrasts. However, these tables do not account for across-subject variability in spatial brain map patterns; a more thorough analysis of consensus across classifiers was performed using DISTATIS, a multidimensional scaling procedure with bootstrapped confidence estimates.

Fig. 8 plots the 2-dimensional representation of similarity between classifiers' brain maps, for each of the four contrasts. The first and second principal components are represented by the horizontal and vertical axes, respectively; on each axis, we specify the amount of variance explained by each component. In this coordinate space, we represent the centroids of the spatial maps created by each of the six methods as '+' symbol. In order to compute the confidence intervals around the centroids, we have drawn 1000 bootstrap samples from our set of 47 subjects. The 95% bootstrapped confidence intervals are shown as ellipses around the centroids. Classifiers with overlapped confidence ellipses are not significantly distinguishable in the DISTATIS space.

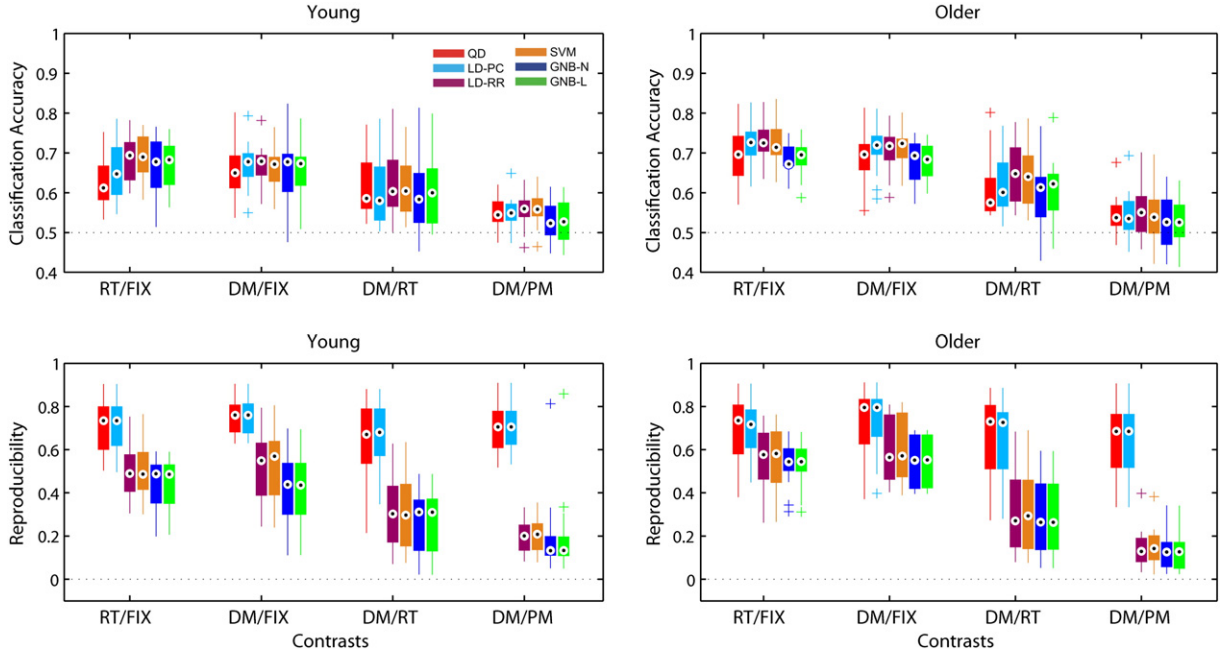

Fig. 5. Performance of the pool of classifiers on the dataset from the aging study, based on metrics of classification accuracy (top) and spatial map reproducibility (bottom). Left and right columns correspond to subjects in the young and the older age groups, respectively. Results are shown for 4 binary task contrasts, ordered from strong to weak within each panel. The tasks are: reaction task (RT), delayed matching (DM), perceptual matching (PM), and fixation (FIX).

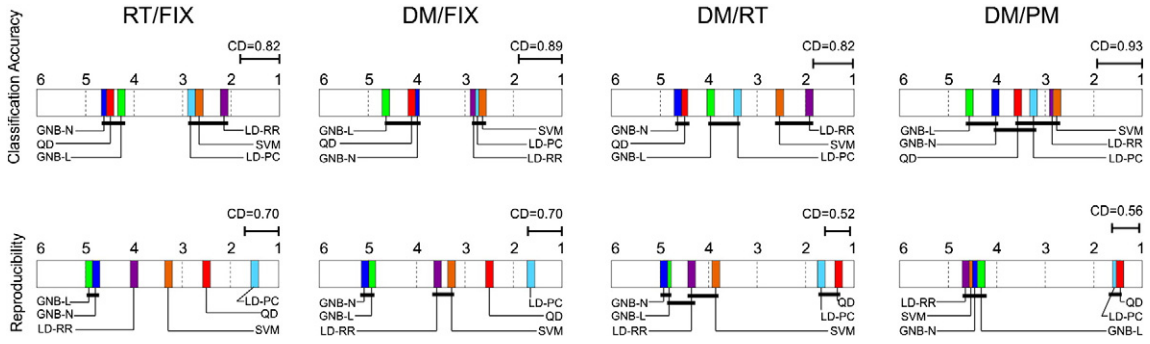

Fig. 6. Ranking of the 6 classifiers in the young age group, based on metrics of classification accuracy (top) and reproducibility (bottom). The mean (across subjects) rank of each classifiers is represented with a colored box. Classifiers that are linked with a black horizontal bar are not significantly different in their ranking. Results are shown for 4 binary task contrasts. The critical distance (CD) is displayed for the contrasts where a significant difference between the classifiers' performance has been found; the performance of the two classifiers is considered significantly different if their rank difference is greater than CD.

We can observe a familiar grouping of methods in Fig. 8, where pairs of classifiers have overlapping confidence ellipses: (QD and LD-PC), (SVM and LD-RR), and (GNB-L and GNB-N). This pairing is observed in simulated data when we evaluate the reproducibility and ROC properties of the algorithms (Fig. 3, bottom row; Fig. 4); it is also observed in evaluation of within-subject reproducibility in the aging study (bottom rows of Figs. 5, 6 and 7). Within each pair of methods, there is a similarity in computational models: both QD and LD-PC use PCA-based regularization, both GNB-L and GNB-N are univariate Gaussian Naive Bayes classifiers, and both SVM and LD-RR use a L2 penalty for regularization. This pairwise similarity between methods appears strongest in the DM/FIX and RT/FIX contrasts, as the corresponding ellipses overlap almost completely. In two weak contrasts (DM/RT and DM/PM), the (QD and LD-PC) and (SVM and LD-RR) pairs of ellipses are less overlapped but still not significantly different while the GNB ellipses remain identical. More striking is the fact that LD-RR and SVM, while having unique pattern features as shown by their negative weighting on PC_2 (vertical) axis, are most like LD-PC and QD for the strongest contrasts and DM/RT with negative PC_1 weights, but for the weakest contrast DM/PM they are like the GNB patterns with positive PC_1 weights. Therefore, the strength of the contrast interacts with the pattern similarities between the three pairs of methods with LD-RR and SVM appearing most sensitive to contrast effects such that their patterns may reflect either multivariate or univariate features.

Group-average classifier maps

Figs. 9 and 10 show the spatial maps, created by our pool of classifiers, normalized into rSPM{Z} patterns and averaged across all subjects within the age group. After the averaging, the group-average maps are corrected for multiple comparisons using false discovery rate (FDR) correction (Genovese et al., 2002) and thresholded at $FDR \leq 0.1$. Figs. 9A and B display the maps created for the RT/FIX contrast for the young and the older groups, respectively. Figs. 10A and B are the DM/FIX maps for the young and the older group. The group-average GNB-N and GNB-L maps for these two contrasts were identical; only the GNB-L maps are shown. The group-average maps created for the weaker contrasts (DM/RT and DM/PM) are not shown, because they do not contain any significant voxels at $FDR \leq 0.1$ threshold (this holds for all classifiers from our pool). Supplementary Table 5 shows Pearson's correlation coefficient between the average map computed for the young group with the average map computed for the older group (both maps unthresholded); this correlation is shown for all classifiers and contrasts.

Comparing Figs. 9A and B, we can see a set of similar peak activations between the young and the old groups. All multivariate classifiers identify the posterior cingulate cortex as having a significant preference to FIX condition. In the young group, this preference is also found in the left angular gyrus and lingual gyrus, in the QD map. Univariate classifiers do not find any brain areas with a significant preference for FIX in either group. Also, multivariate classifiers find a larger set of areas with significant preference for RT than univariate classifiers do. In the young group, these areas are: bilateral cerebellum, contralateral intraparietal lobule (IPL), middle cingulate/supplementary motor area (SMA), and primary motor cortex (contralateral stronger than ipsilateral). The LD-PC and QD maps show a larger set of voxels than LD-RR and SVM, and GNB does not identify activation in the contralateral primary motor cortex/SMA, contralateral IPL, and middle cingulate/SMA. In the older group, preference for RT is found in the same areas as in the young group, but the spatial extent is larger, whereas FIX activations

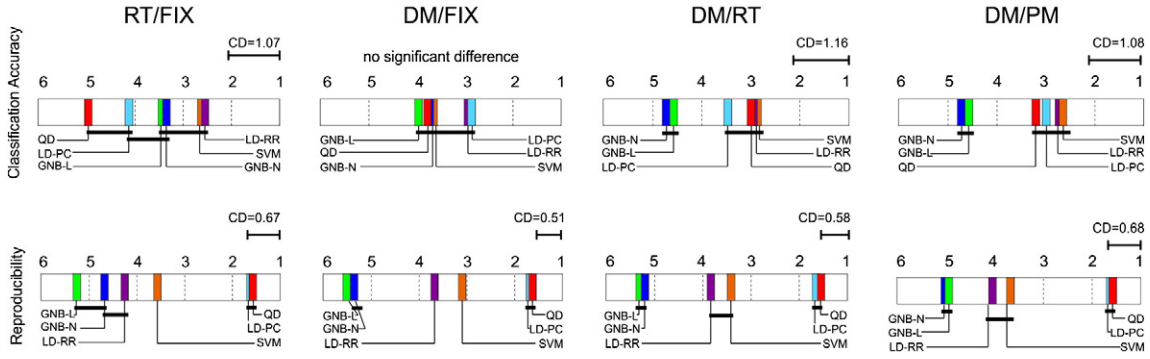

Fig. 7. Ranking of the 6 classifiers in the older age group, based on metrics of classification accuracy (top) and reproducibility (bottom). The mean (across subjects) rank of each classifiers is represented with a colored box. Classifiers that are linked with a black horizontal bar are not significantly different in their ranking. Results are shown for 4 binary task contrasts. The critical distance (CD) is displayed for the contrasts where a significant difference between the classifiers' performance has been found; the performance of the two classifiers is considered significantly different if their rank difference is greater than CD.

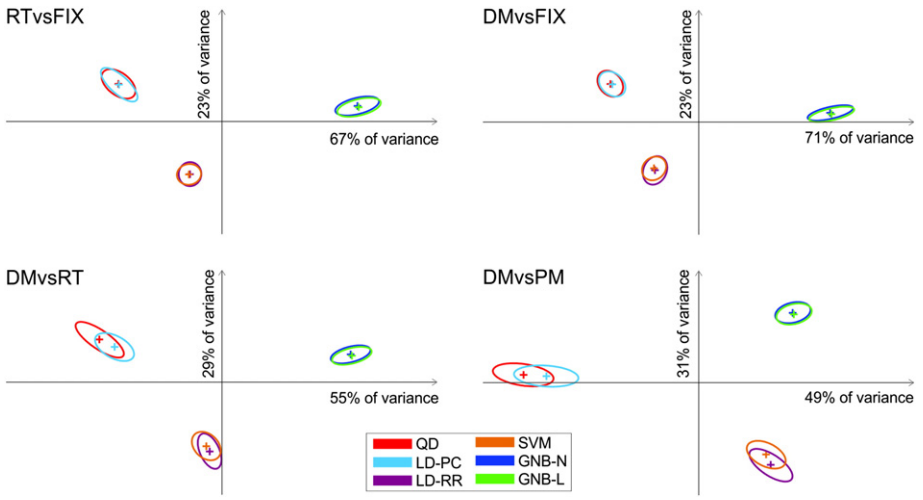

Fig. 8. DISTATIS plots of similarity of within-subject maps created with different classifiers. Each classifier model is represented as a centroid '+' with 95% confidence ellipse (based on bootstrap resampling, 1000 iterations). Models that are closer in DISTATIS space produce more similar spatial patterns, and are not significantly different if their confidence ellipses overlap.

are not detected by many classifiers. In addition, LD-PC and QD find significant preference in anterior insula/temporal poles. In the older group, significant preference for RT is found bilaterally in the IPL and primary motor cortex (although there is still greater spatial extent in contralateral activation); this is different from the young group, where this preference tends to be predominantly on the contralateral side.

The group-average maps for the DM/FIX contrast, displayed in Fig. 10, are similar to the RT/FIX maps. Compared with RT/FIX maps,

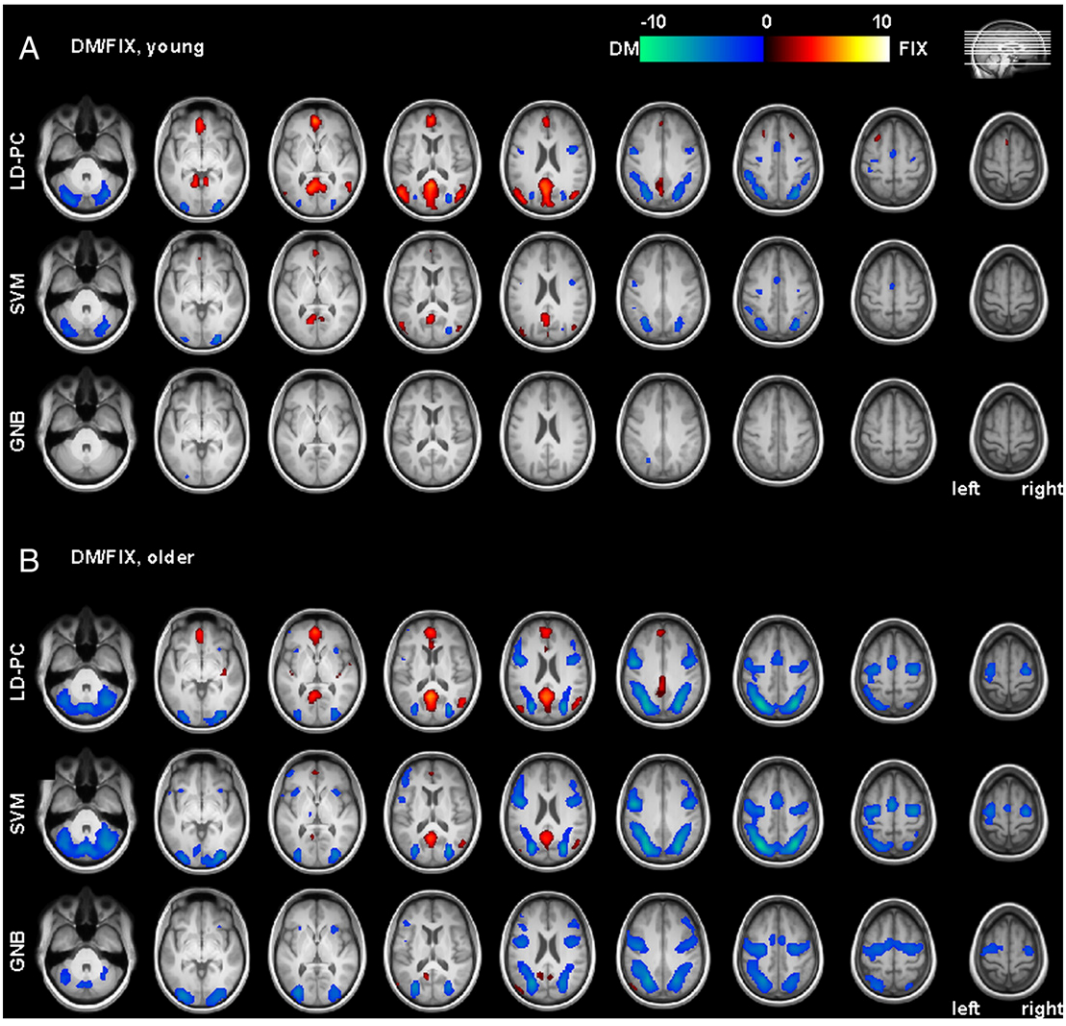

Fig. 9. Group-average *Z*-scored classifier maps for the RT/FIX contrast for the young (A) and older (B) groups. Maps are thresholded at false-discovery rate 0.1. Color bar represents average *Z* scores.

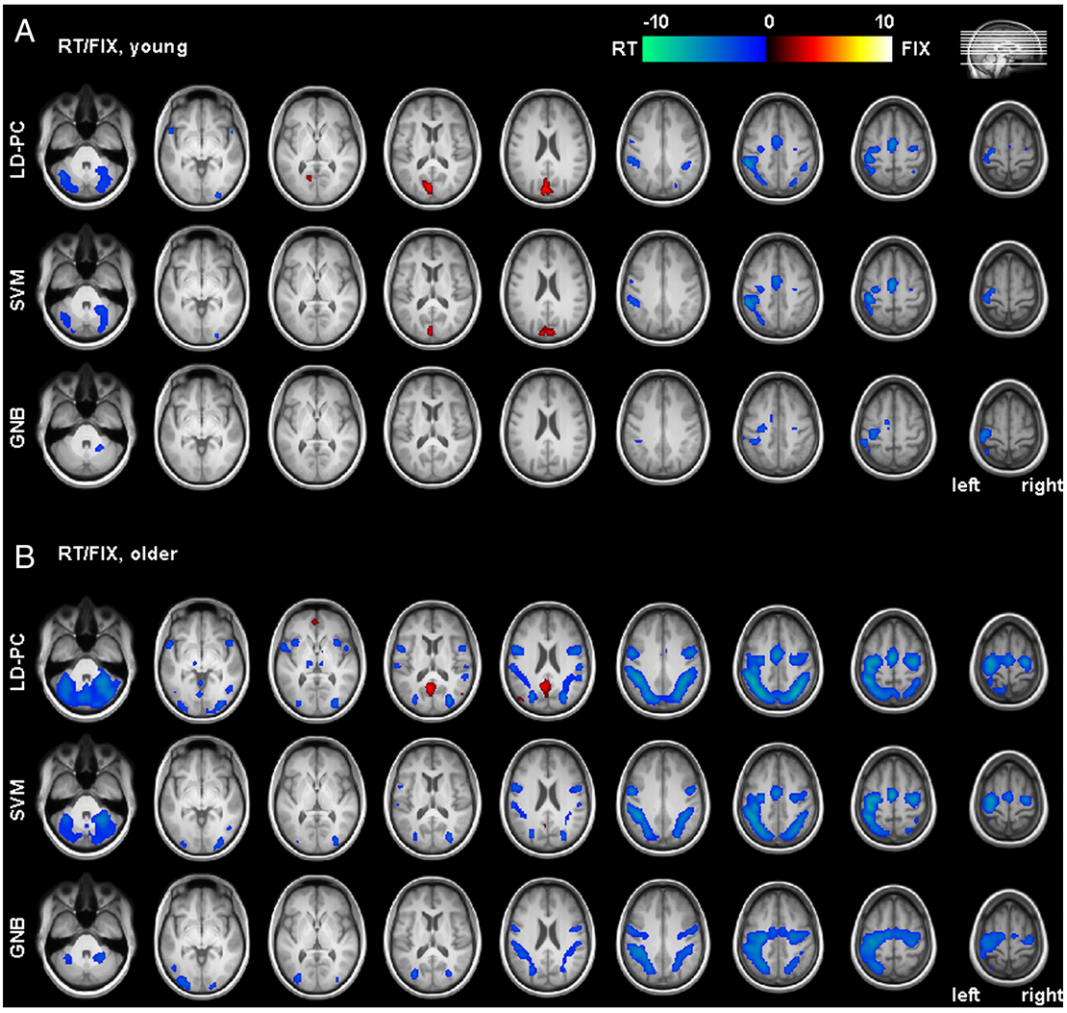

Fig. 10. Group-average *Z*-scored classifier maps for the DM/FIX contrast for the young (A) and older (B) groups. Maps are thresholded at false-discovery rate 0.1. Color bar represents average *Z* scores.

the preference for FIX is more extensive throughout the brain in the DM/FIX contrast, particularly in the young group. LD-PC and QD maps reveal this preference in posterior and anterior cingulate cortex, and in the inferior parietal lobe. These areas are known as parts of default mode network (Grady et al., 2010); the activity in these areas is known to increase when the participants perform a passive task (such as FIX) or attend to internally driven cognitive processes, and to decrease when performing an active task (such as DM or RT) requiring focusing on external stimuli. LD-RR and SVM find a subset of these areas. The only area with significant FIX preference that GNB finds in the older group is posterior cingulate, and no such areas are found by GNB in the young group. With respect to DM preference, LD-PC and QD find the largest number of brain areas, and GNB finds the smallest number (in the young group, the only such area found by GNB is contralateral IPL). The multivariate maps reveal this preference in cerebellum, bilateral premotor and primary motor cortex, IPL and SMA. In the older group, the spatial extent of the areas with significant DM preference is larger than in the young group (this is consistent with the RT/FIX contrast). In addition, this preference is found in some areas that are not observed n the young group: bilateral interior insula and in left dorsolateral prefrontal cortex.

Discussion

We evaluated the performance of six pattern classification algorithms, using the NPAIRS resampling framework (Strother et al., 2002). Our pool of classifiers includes representatives of a variety of important types of classification algorithms: linear vs. nonlinear, multivariate vs. univariate, and probabilistic vs. non-probabilistic. To determine how the performance of these classifiers is influenced by the magnitude, temporal variance and spatial correlation of the active signal, we applied them to a series of simulated fMRI sets.

The active signal is absent in the simulated "baseline" volumes; therefore, when we increase its magnitude M, the separation between the "active" and "baseline" classes grows, and the accuracy of classifiers grows accordingly. The accuracy of classifiers becomes more similar as M grows. When $M = 0.03$, the increase of temporal variance (V) of the active signal is detrimental to classification accuracy, because it increases the spread of the volumes in the "active" class, and decreases the separation between the classes.

When $M = 0$, the difference between the classes is in their covariance matrices rather than in their mean signal amplitudes. Linear classifiers ignore this difference, but nonlinear classifiers are able to use it to their advantage: nonlinear methods (QD and GNB-N) get more accurate as V increases. This beneficial effect of V is modulated by ρ. QD is the most accurate classifier when active networks are strongly coupled and $\rho = 0.99$. When $\rho = 0$, the functional nodes of the active network are independent, and GNB-N becomes the best model for our data. This suggests the utility of nonlinear classifiers in situations when the difference between classes is driven by variance/covariance rather than by magnitude of BOLD signal. This agrees with our previous finding

(Schmah et al., 2010), where we used several fMRI sessions of stroke patients and classified the volumes into "early" and "late" classes depending on the timing of the session relative to the stroke onset. We found that nonlinear classifiers (e.g., QD, SVM with quadratic and RBF kernels) were significantly more accurate than linear classifiers (e.g., LD and linear-kernel SVM), which could be explained by the change in brain connectivity (due to the process of post-stroke recovery of function) that was happening between the early and late sessions. We can hypothesize that nonlinear classifiers could be useful for across-subject studies of diseases that are related to connectivity deficits (e.g., schizophrenia, depression, and Alzheimer's disease; see Greicius, 2008, for a review of connectivity deficits in neuropsychiatric disorders).

In terms of reproducibility of spatial maps and partial ROC area, there is a difference between classifiers at all levels of M. The best reproducibility is achieved by multivariate methods that use PCA regularization. We have previously shown that the simulated active signal is efficiently captured by the first few principal components of the data matrix when $\rho > 0$ and V is sufficiently high (Yourganov et al., 2011). With growing V, there is an increase in the portion of total variance that is due to the active signal. In PCA, principal components are ranked by the amount of variance that they explain; as V increases, the variance explained by the first few principal components is increasingly due to correlated active signal. PCA is, in essence, a detector of correlated sources of variance; this helps the methods that use PCA for regularization to improve their signal detection (Fig. 4) and the reproducibility of their maps (Fig. 3, bottom row). Univariate methods, as well as multivariate methods that do not use PCA regularization, do not benefit from increasing V; univariate maps are the least-reproducible and also the worst in terms of ROC area.

Our results suggest that a researcher should consider classifying fMRI data with linear and quadratic discriminants using PC regularization, and select the better performer of the two. In terms of classification accuracy, the most accurate method from the (LD-PC, QD) pair often outperforms linear SVM and LD-RR, particularly given significant network structure (i.e., $V \geq 0.6$ and $\rho \geq 0.5$). In terms of reproducibility of maps and ROC area, LD-PC and QD are always the best performers among the tested methods. Running a combination of QD and LD-PC does not take more computational time than running either QD or LD-PC by itself, because the computation is dominated by PCA decomposition of the data (Schmah et al., 2010). When the difference between classes is driven by the difference in means, LD-PC is likely to be the best method of the two; when it is driven by difference in connectivity between brain areas with weak mean differences, QD is likely to be best. However, for the weak contrasts DM/RT and DM/PM of our experimental data in Fig. 8 this only appears as a nonsignificant mean shift of the QD ellipsoid relative to that of LD-PC.

In the within-subject analysis of experimental data, the difference between the classifiers' accuracy is more pronounced than in our simulations. SVM and LD-RR tend to be the best-ranking classifiers, although for most contrasts they are not significantly different from one or both of QD and LD-PC (this observation has also been reported by Churchill et al., in press). GNB-L and GNB-N tend to be the worst-ranking. With respect to reproducibility of spatial maps, the same grouping is observed in experimental data and in simulations: (best) LD-PC and QD, (intermediate) LD-RR and SVM, (worst) GNB-N and GNB-L. Nonparametric testing of ranking also shows this grouping (Figs. 6 and 7), and evaluation of consensus across classifiers (Fig. 8) reinforces this finding including the tendency for LD-RR and SVM to be more like GNB for weaker contrasts. It should be noted that both LD-RR and SVM use an L2-type penalty in regularization, which could account for their similarity (observed also by Rasmussen et al., 2012b, and by Churchill et al., in press).

All classifiers tend to be more accurate when the contrast is defined by a pair of tasks that recruit spatially different brain networks. DM/FIX and RT/FIX are examples of such "strong" contrasts: FIX condition corresponds to passive fixation and is expected to recruit areas from the "default mode network" (Toro et al., 2008), which is often anti-correlated with "task-positive" areas recruited by active visuomotor tasks such as DM and RT. The DM/PM contrast can be assumed to be the weakest, because of the similarity between the DM and PM tasks (because of this similarity, we did not study PM/FIX and PM/RT contrasts). Finally, the DM/RT contrast is intermediate. These four contrasts represent a range of "contrast strengths", which is somewhat analogous to varying M, V and ρ in our simulations. In the strong contrasts, the across-subject reproducibility of classifier maps is higher compared to the weak contrasts; also, the consensus between classifiers is higher in the strong contrasts.

Comparing the classifiers' performance on simulated (Fig. 3) and experimental (Fig. 5) data, we can see that the reproducibility of spatial maps tends to be higher in experimental data, particularly for PC-based methods (QD and LD-PC). This can be explained by the fact that experimental data consist of a much greater number of voxels, among which a large portion of voxels forms multiple, spatially extensive cortical networks such as the default-mode network and the task-positive network, compared to the simulated spatially sparse single-network brain. Principal component analysis reliably detects these networks, resulting in spatial maps that are highly reproducible even when they are not predictive (e.g., in the DM/PM contrast). It is also possible that the magnitude of the task-driven signal in experimental data is higher than in our simulations; unfortunately, this hypothesis is hard to test because of the difficulty of separating the signal from the noise in experimental data. This difficulty is discussed in a recent review paper (Welvaert and Rosseel, 2013). For a particular method, we can define the signal and noise axes (see Normalizing individual spatial maps) assuming that signal is reproducible within a spatial map and noise is not. This, however, makes the definition of contrast-to-noise ratio and dynamic range specific to our method of analysis, and therefore not equivalent to the true signal parameters M and V in our simulations.

The group-average classifier maps created for the two weaker contrasts do not contain any voxels that pass the relatively liberal threshold of significance of $FDR \leq 0.1$. The maps created for strong contrast reveal preference for FIX in the default-mode areas, and a preference for RT or DM in the "task-positive" areas such as motor and premotor cortices, supplementary motor area, intraparietal lobule and cerebellum. This is most evident from the LD-PC and QD maps; univariate (GNB) maps are the least informative, particularly in the young group. Comparing the group-average maps for the two age groups, we can see that the older group recruits the task-positive areas more extensively than the young group, and, conversely, the recruitment of default-mode areas is more extensive in the young group. This is consistent with the results obtained by Grady et al. (2010) on the same data set with a different method of analysis (partial least squares analysis, with pooling across subjects).

Analyzing fMRI data with LD and QD on a PC subspace has an additional advantage not discussed in this paper: the additional information obtained by determining the optimal PC subspace. Size of this subspace is the number of orthogonal dimensions in the model that captures the relevant information in the data and ignores the noise (Yourganov et al., 2011). This number has been shown to have a neurobiological significance: it relates to behavioral measures of post-stroke recovery of motor function (Yourganov et al., 2010), as well as the strength of self-control (Berman et al., 2013). The highest-ranking principal components capture the most important correlated sources of variance in the BOLD signal; therefore, we expect LD and QD on regularized a PC subspace to be efficient detectors of brain networks as we have demonstrated in our simulated and experimental data.

We suggest that the current focus in the literature on using SVM (or equivalently LD-RR) even with, but mostly without, careful regularization emphasizes prediction at the expense of stability of brain maps. Our previous results (Schmah et al., 2010) demonstrate the usefulness of PC-regularized LD and QD in classifying fMRI volumes. The current work suggests that the spatial maps created for these two classifiers

have the advantage of being highly reproducible (see also Churchill et al., in press) in addition to being highly accurate. Reproducibility is an important criterion in fMRI analyses, as reliable activation maps are required to interpret the brain regions that underlie task performance. Although we examined a range of simulated parameters and experimental task contrasts, it is important to test whether these effects generalize to other experimental datasets, as well as testing classifiers that have been omitted in the current study (in particular, SVMs with nonlinear kernels). Currently, our results indicate that a switch to LD and QD on a carefully regularized PC basis may lead to generally improved classification results and spatial activation maps.

Acknowledgments

The authors thank Natasa Kovacevic for preprocessing the fMRI data acquired in the study of aging. This work was supported by the Brain Network Recovery Group through a grant from the James S. McDonnell Foundation (no. 22002082). S.C.S. is partially supported by the Centre for Stroke Recovery of the Heart and Stroke Foundation of Canada.

Appendix A. Simulated data

The simulated volumes consist of three additive parts: (1) background structure, (2) baseline noise, and (3) active signal.

1) Background structure is based on a PET image of a phantom (Lukic et al., 2002). It is displayed in Fig. 1, left. It consists of simulated "gray matter" along the edge of the phantom and in its center, and of "white matter" elsewhere. Before spatial smoothing, the amplitude of "gray matter" is 4 units, and the amplitude of "white matter" is 1 unit. The background structure is spatially smoothed using a Gaussian kernel with FWHM (full width at half maximum) of 2 pixels.
2) Baseline noise is zero-mean white Gaussian, smoothed with the same Gaussian kernel. After smoothing, the standard deviation of baseline noise is 5% of the background signal.
3) Active signal is added at 16 specific locations (see Fig. 1, right), 12 of them in the "gray matter" and 4 in the "white matter". It is the sum of a constant term and a variable term. The constant term is the value of the background structure at this location, multiplied by M. It specifies the expected magnitude of the active signal. The variable term is a zero-mean random variable, which is created as described below.

To create correlations between our 16 active loci, we construct a 16×16 covariance matrix $\mathbf{S}$. The $(i, j)^{\text{th}}$ element of $\mathbf{S}$ is

- if $i \neq j$ and both i and j are in gray matter, $s_{ij} = 4^*4^*\rho V v_i^2$;
- if $i = j$ and i is in gray matter, $s_{ii} = 4^*4^*V v_i^2$;
- if $i \neq j$, and one of them is in gray matter and the other in white matter, $s_{ij} = 4^*\rho V v_i^2$;
- if $i \neq j$ and both i and j are in white matter, $s_{ij} = \rho V v_i^2$;
- if $i = j$ and i is in white matter, $s_{ii} = V v_i^2$.

Here, v_i is the standard deviation of the baseline noise at the ith active area; for all active areas, it is equal to 5% of the background signal. The parameter ρ controls the correlation between the active loci, and the parameter V controls the temporal variance of the active signal in a locus. Matrix $\mathbf{S}$ defines the covariance between the active loci. To create a variable term, we take a 16-dimensional vector sampled from a Gaussian distribution with zero mean and identity covariance matrix. Then, we multiply it by the Cholesky decomposition of the matrix $\mathbf{S}$. This gives us the magnitude of activations of the 16 centers of active loci. Then, it is spatially blurred using a Gaussian kernel with FWHM varying from 2 to 4 pixels for different loci.

Appendix B. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2014.03.074.

References

Abdi, H., O'Toole, A.J., Valentin, D., Edelman, B., 2005. DISTATIS: the analysis of multiple distance matrices. Proceedings of the IEEE Computer Society: International Conference on Computer Vision and Pattern Recognition, pp. 42–47.
Abdi, H., Dunlop, J.P., Williams, L.J., 2009. How to compute reliability estimates and display confidence and tolerance intervals for pattern classifiers using the Bootstrap and 3-way multidimensional scaling (DISTATIS). NeuroImage 45 (1), 89–95.
Beckmann, C.F., Smith, S.M., 2004. Probabilistic independent component analysis for functional magnetic resonance imaging. IEEE Trans. Med. Imaging 23, 137–152.
Berman, M., Yourganov, G., Askren, M.K., Ayduk, O., Casey, B.J., Gotlib, I.H., Kross, E., McIntosh, A.R., Strother, S.C., Wilson, N.L., Zayas, V., Mischel, W., Shoda, Y., Jonides, J., 2013. Dimensionality of brain networks linked to life-long individual differences in self-control. Nat. Commun http://dx.doi.org/10.1038/ncomms2374 (Article # 1373).
Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2 (3) (Article # 27).
Churchill, N.W., Oder, A., Abdi, H., Tam, F., Lee, W., Thomas, C., Ween, J.E., Graham, S.J., Strother, S.C., 2012a. Optimizing preprocessing and analysis pipelines for single-subject fMRI. 1. Standard temporal motion and physiological noise correction methods. Hum. Brain Mapp. 33 (3), 609–627.
Churchill, N.W., Yourganov, G., Oder, A., Tam, F., Graham, S.J., Strother, S.C., 2012b. Optimizing preprocessing and analysis pipelines for single-subject fMRI. 2. Interactions with ICA, PCA, task contrast and inter-subject heterogeneity. PLoS One 7 (2).
Churchill, N.W., Yourganov, G., Strother, S.C., 2014. Comparing within-subject classification and regularization methods in fMRI for large and small sample sizes. Hum. Brain Mapp. http://dx.doi.org/10.1002/hbm.22490 (in press).
Conover, W.J., 1999. Practical Nonparametric Statistics, 3rd ed. Wiley, New York, NY.
Cortes, C., Vapnik, V.N., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.
Cox, R.W., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29, 162–173.
Demsar, J., 2006. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30.
Efron, B., Tibshirani, R., 1993. Introduction to the Bootstrap. Academic Press, San Diego.
Genovese, C.R., Lazar, N.A., Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 15 (4), 870–878.
Glover, G.H., 1999. Deconvolution of impulse response in event-related BOLD fMRI. NeuroImage 9, 416–429.
Grady, C.L., Protzner, A.B., Kovacevic, N., Strother, S.C., Afshin-Pour, B., Wojtowicz, M., Andreson, J.A.E., Churchill, N., McIntosh, A.R., 2010. A multivariate analysis of age-related differences in default mode and task-positive networks across multiple cognitive domains. Cereb. Cortex 20, 1432–1447.
Greicius, M., 2008. Resting-state functional connectivity in neuropsychiatric disorders. Curr. Opin. Neurol. 21 (4), 424–430.
Hassabis, D., Chu, C., Rees, G., Weiskopf, N., Molyneux, P.D., Maguire, E.A., 2009. Decoding neuronal ensembles in the human hippocampus. Curr. Biol. 19 (7), 546–554.
Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430.
Haynes, J.D., Rees, G., 2006. Decoding mental states from brain activity in humans. Nat. Rev. Neurosci. 7, 523–534.
Kamitani, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. Nat. Neurosci. 8, 679–685.
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, 772–786.
Ku, S.-P., Gretton, A., Macke, J., Logothetis, N.K., 2008. Comparison of pattern recognition methods in classifying high-resolution BOLD signals obtained at high magnetic field in monkeys. Magn. Reson. Imaging 26, 1007–1014.
Kustra, R., Strother, S., 2001. Penalized discriminant analysis of [^{15}O]-water PET brain images with prediction error selection of smoothness and regularization hyperparameters. IEEE Trans. Med. Imaging 20, 376–387.
LaConte, S., Anderson, J., Muley, S., Ashe, J., Frutiger, S., Rehm, K., et al., 2003. The evaluation of preprocessing choices in single-subject BOLD fMRI using NPAIRS performance metrics. NeuroImage 18, 10–27.
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.
Lukic, A.S., Wernick, M.N., Strother, S.C., 2002. An evaluation of methods for detecting brain activations from functional neuroimages. Artif. Intell. Med. 25, 69–88.
Mason, S.J., Graham, N.E., 2002. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: statistical significance and interpretation. Q. J. R. Meteorol. Soc. 128, 2145–2166.
Metz, C.E., Herman, B.A., Shen, J.-H., 1998. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. Stat. Med. 17, 1033–1053.
Misaki, M., Kim, Y., Bandettini, P.A., Kriegeskorte, N., 2010. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. NeuroImage 53 (1), 103–118.

Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X., Just, M., Newman, S., 2004. Learning to decode cognitive states from brain images. Mach. Learn. 57 (1–2), 145–175.
Morch, N., Hansen, L.K., Strother, S.C., Svarer, C., Rottenberg, D.A., Lautrup, B., Savoy, R., Paulson, O., 1997. Nonlinear versus linear models in functional neuroimaging: learning curves and generalization crossover. Information Processing in Medical Imaging, vol. 1230. Springer-Verlag, New York, pp. 259–270.
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 (9).
Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. NeuroImage 45, S199–S209.
Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R.J., Kahn, R.S., Ramsey, N.F., 2007. Test–retest reliability of fMRI activation during prosaccades and antisaccades. NeuroImage 36, 532–542.
Rasmussen, P.M., Madsen, K.H., Lund, T.E., Hansen, L.K., 2011. Visualization of nonlinear kernel models in neuroimaging by sensitivity maps. NeuroImage 55 (3), 1120–1131.
Rasmussen, P.M., Schmah, T., Madsen, K.H., Lund, T.E., Yourganov, G., Strother, S., Hansen, L.K., 2012a. Visualization of nonlinear classification models in neuroimaging — signed sensitivity maps. Paper presented at the Biosignals 2012, International Conference on Bio-inspired Systems and Signal Processing.
Rasmussen, P.M., Hansen, L.K., Madsen, K.H., Churchill, N.W., Strother, S.C., 2012b. Model sparsity and brain pattern interpretation of classification models in neuroimaging. Pattern Recogn. 45 (6), 2085–2100.
Robert, P., Escoufier, Y., 1976. A unifying tool for linear multivariatestatistical methods: the RV-coefficient. Appl. Stat. 25, 257–265.
Schmah, T., Yourganov, G., Zemel, R.S., Hinton, G.E., Small, S.L., Strother, S.C., 2010. Comparing classification methods for longitudinal fMRI studies. Neural Comput. 22 (11), 2729–2762.
Schrouff, J., Kussé, C., Wehenkel, L., Maquet, P., Phillips, C., 2012. Decoding semi-constrained brain activity from fMRI using support vector machines and Gaussian processes. PLoS One 7 (4).
Seber, G.A.F., 2004. Multivariate Observations. Wiley-Interscience.
Skudlarski, P., Constable, T.R., Gore, J.C., 1999. ROC analysis of statistical methods used in functional MRI: individual subjects. NeuroImage 9 (3), 311–329.
Strother, S.C., Lange, N., Anderson, J.R., Schaper, K.A., Rehm, K., Hansen, L.K., et al., 1997. Activation pattern reproducibility: measuring the effects of group size and data analysis models. Hum. Brain Mapp. 5, 312–316.
Strother, S.C., Anderson, J., Hansen, L.K., Kjems, U., Kustra, R., Sidtis, J., et al., 2002. The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. NeuroImage 15, 747–771.
Strother, S., La Conte, S., Kai Hansen, L., Anderson, J., Zhang, J., Pulapura, S., et al., 2004. Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics: I. A preliminary group analysis. NeuroImage 23 (Suppl. 1), 196–207.
Strother, S.C., Oder, A., Spring, R., Grady, C., 2010. The NPAIRS computational statistics framework for data analysis in neuroimaging. Paper Presented at the 19th International Conference on Computational Statistics: Keynote, Invited and Contributed Papers.
Tegeler, C., Strother, S.C., Anderson, J.R., Kim, S.G., 1999. Reproducibility of BOLD-based functional MRI obtained at 4 T. Hum. Brain Mapp. 7, 267–283.
Toro, R., Fox, P.T., Paus, T., 2008. Functional coactivation map of the human brain. Cereb. Cortex 18 (11), 2553–2559.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-VerlagNew York, Inc., New York, NY, USA.
Welvaert, M., Rosseel, Y., 2013. On the definition of signal-to-noise ratio and contrast-to-noise ratio for fMRI data. PLoS One 8 (11), e77089.
Woods, R.P., Grafton, S.T., Holmes, C.J., Cherry, S.R., Mazziotta, J.C., 1998. Automated image registration: I. General methods and intrasubject, intramodality validation. J. Comput. Assist. Tomogr. 22, 139–152.
Worsley, K.J., 2001. Statistical analysis of activation images. Functional MRI: Introduction to Methods pp. 251–270.
Yourganov, G., Schmah, T., Small, S.L., Rasmussen, P.M., Strother, S.C., 2010. Functional connectivity metrics during stroke recovery. Arch. Ital. Biol. 148 (3), 259–270.
Yourganov, G., Xu, C., Lukic, A., Grady, C., Small, S., Wernick, M., Strother, S.C., 2011. Dimensionality estimation for optimal detection of functional networks in BOLD fMRI data. NeuroImage 56 (2), 531–543.
Zhang, J., Liang, L., Anderson, J.R., Gatewood, L., Rottenberg, D.A., Strother, S.C., 2008. A Java-based fMRI processing pipeline evaluation system for assessment of univariate general linear model and multivariate canonical variate analysis-based pipelines. Neuroinformatics 6, 123–134.