Neuroinform
DOI 10.1007/s12021-013-9178-1

ORIGINAL ARTICLE

PRoNTo: Pattern Recognition for Neuroimaging Toolbox

J. Schrouff · M. J. Rosa · J. M. Rondina · A. F. Marquand · C. Chu · J. Ashburner · C. Phillips · J. Richiardi · J. Mourão-Miranda

Abstract In the past years, mass univariate statistical analyses of neuroimaging data have been complemented by the use of multivariate pattern analyses, especially based on machine learning models. While these allow an increased sensitivity for the detection of spatially distributed effects compared to univariate techniques, they lack an established and accessible software framework. The goal of this work was to build a toolbox comprising all the necessary functionalities for multivariate analyses of neuroimaging data, based on machine learning models. The "Pattern Recognition for Neuroimaging Toolbox" (PRoNTo) is open-source, cross-platform, MATLAB-based and SPM compatible, therefore being suitable for both cognitive and clinical neuroscience research. In addition, it is designed to facilitate novel contributions from developers, aiming to improve the interaction between the neuroimaging and machine learning communities. Here, we introduce PRoNTo by presenting examples of possible research questions that can be addressed with the machine learning framework implemented in PRoNTo, and cannot be easily investigated with mass univariate statistical analysis.

Keywords Neuroimaging software · Pattern recognition · Machine learning · Image analysis · MVPA

J. Schrouff and M.J. Rosa contributed equally to this work

J. Schrouff · C. Phillips
Cyclotron Research Centre, University of Liège, Liège, Belgium

M. J. Rosa (✉) · J. M. Rondina · J. Mourão-Miranda
Department of Computer Science, Centre for Computational Statistics and Machine Learning, University College London, Gower Street,
WC1E 6BT London, UK
e-mail: m.rosa@ucl.ac.uk

A. F. Marquand · J. Mourão-Miranda
Department of Neuroimaging, Centre for Neuroimaging Sciences, Institute of Psychiatry, King's College London, London, UK

C. Chu
Section on Functional Imaging Methods, Laboratory of Brain and Cognition, NIMH, NIH, Bethesda, USA

J. Ashburner
Wellcome Trust Centre for NeuroImaging, University College London, London, UK

C. Phillips
Department of Electrical Engineering and Computer Science, University of Liège, Liège, Belgium

J. Richiardi
Functional Imaging in Neuropsychiatric Disorders Lab, Department of Neurology and Neurological Sciences, Stanford University, Stanford, USA

J. Richiardi
Laboratory for Neurology & Imaging of Cognition, Departments of Neurosciences and Clinical Neurology, University of Geneva, Geneva, Switzerland

J. M. Rondina
Neuroimaging Laboratory, Department and Institute of Psychiatry, Faculty of Medicine, University of São Paulo, São Paulo, Brazil

Introduction

Two of the most fundamental questions in the field of neurosciences are how information is represented in the different brain structures, and how this information evolves with time. Various imaging modalities, such as functional Magnetic Resonance Imaging (fMRI) and Positron Emission Tomography (PET), have been developed to record brain activity and therefore allow the investigation of these questions. Until recently, methods used to analyze such data focused mainly on characterizing the relationship between a mental state and each image voxel time-series, i.e. following mass univariate statistical approaches such as the General Linear Model (GLM) in Statistical Parametric Mapping (SPM, (Friston et al. 2007)). In addition to functional modalities, other neuroimaging techniques exist, such as structural Magnetic Resonance Imaging (sMRI), which allows one to investigate

Published online: 16 February 2013

brain anatomy. For these data, one of the most commonly used analysis approach is Voxel-Based Morphometry, (VBM, (Ashburner & Friston, 2000)), which investigates focal differences in grey matter density between groups of subjects, again using a mass univariate approach. Although univariate analyses have proven powerful for making regionally specific inferences on brain function and structure, there are limitations to the type of research questions that they can address.

More recently, these mass univariate analyses have been complemented by the use of multivariate pattern analyses (MVPA), in particular using machine learning based predictive models (Pereira et al. 2009). These analyses focus on predicting a variable of interest (e.g. mental state 1 vs. mental state 2, or patients vs. controls) from the pattern of brain activation/anatomy over a set of voxels. Due to their multivariate properties, these methods can achieve relatively greater sensitivity and are therefore able to detect subtle, spatially distributed activations and patterns of brain anatomy.

Multivariate analyses range from 'mind reading' studies (e.g. Haynes & Rees 2006) to clinical applications (e.g. Borroni et al. 2006), and while most studies have focused on fMRI and sMRI data, applications extend to other modalities (Phillips et al. 2011). As examples, multivariate methods applied to fMRI have made it possible to decode the category of an object (Spiridon and Kanwisher 2002; Cox and Savoy 2003; Shinkareva et al. 2008) and orientation of a striped pattern (Haynes and Rees 2005; Kamitani and Tong 2005) visually presented to the subject, solely from the image patterns. The prediction of mental states related to memory retrieval (Polyn et al. 2005; Chadwick et al. 2010), or the patterns of hidden intentions (Haynes et al. 2007) can also be achieved. More recently, these methods have been applied to datasets involving subtler and higher-level cognitive tasks, including the prediction of subjective pain intensity (Marquand et al. 2010), as well as the content of semi-constrained brain activity (Schrouff et al. 2012a, b). Furthermore, multivariate pattern recognition methods can be extremely useful in distinguishing between groups of subjects (e.g. healthy versus patient), and can potentially be used as a diagnostic tool in a clinical setting (Kloppel et al. 2011). For example, the classification of subjects into healthy and patients with Alzheimer's Disease (AD), using structural MRI, has achieved accuracies between 86 % (Vemuri et al. 2008) and 96 % (Klöppel et al. 2008), depending on the sample size, information used and reliability of the diagnostic labels. Similarly, the authors of Phillips et al. (2011) were able to assimilate Locked-In Syndrome (LIS) patients as conscious, based on FDG-PET images, and therefore discriminate them from Vegetative State patients, and in Richiardi et al. 2012, minimally disabled multiple sclerosis patients were discriminated from controls. Not only diagnosis but also prognosis can be performed, as in Mourão-Miranda et al. (2012b), in which the authors predict the risk of mood disorders in healthy adolescents. The major assets of machine learning based predictive models for clinical applications are the objectivity and automaticity of these techniques, especially when the diagnosis of the considered illness remains uncertain using clinical neuroimaging examination and/or neuropsychological tests.

While multivariate pattern analyses can help to investigate brain function and potentially be used as a diagnostic approach for neurologic or psychiatric disorders, many existing implementations consist of small code snippets, or sets of packages, and lack a dedicated single, integrated, and flexible software framework. In addition, the use of existing packages often requires high-level programming skills.

To the extent of our knowledge, the five freely available packages for machine learning modeling of neuroimaging data are the *3dsvm* plugin for AFNI[1] (LaConte et al. 2005), the MATLAB MVPA toolbox,[2] PROBID,[3] PyMVPA (Hanke et al. 2009a, b) and Sci-kit Learn[4] (Pedregosa et al. 2011). The first two are small command line toolboxes and were specifically designed for the classification of fMRI images, therefore being limited to this type of data.

PyMVPA and Sci-kit Learn are sophisticated and flexible software packages primarily written in Python (a free and cross-platform programming language[5]). Being part of the larger Python environment, allows these toolboxes to easily access a range of other neuroimaging and machine learning packages, which renders them very general and able to support different types of neuroimaging data from Magneto/Electroencephalography (M/EEG) to s/fMRI (Hanke et al. 2009a). However, they are also command line-based and therefore do not provide user-friendly graphical interfaces (including a pre-defined interface for displaying results). In addition, they are not directly integrated (through a user-interface) with (MATLAB based) SPM software, which is widely used by the neuroscience community.

PROBID provides easy to use graphical interfaces but is optimized for groups' classification (i.e. classifying patients vs. healthy controls) and does not easily enable single subject analysis or a flexible cross-validation framework. It also does not provide multi-class classification.

Table 1 provides a summary of the available packages including some of the characteristics of each package (standard distribution, i.e. without additional toolboxes), as well as advantages and limitations. The goal of the Pattern Recognition for Neuroimaging Toolbox (PRoNTo) project was therefore to develop a user-friendly and open-source toolbox that could make machine learning modeling available to every neuroscientist.

PRoNTo is a MATLAB toolbox based on pattern recognition techniques for the analysis of neuroimaging data.

[1] http://afni.nimh.nih.gov/
[2] http://www.csbmb.princeton.edu/mvpa/
[3] http://www.kcl.ac.uk/iop/depts/neuroimaging/research/imaginganalysis/Software/PROBID.aspx
[4] http://scikit-learn.org/
[5] http://python.org/

Table 1 Comparison of the main features of the available software packages. Beta represents the coefficients resulting from a General Linear Model (GLM) univariate analysis (as performed in SPM), ASL stands for Arterial Spin Labelling, SVM for Support Vector Machines, GP for Gaussian Processes, (k)NN for (k-) Nearest Neighbors, SMLR for Sparse Multinomial Logistic Regression, LARS for Least Angle Regression, (K)RR for (Kernel) Ridge Regression, RVR for Relevance Vector Regression, PLR for Penalized Logistic Regression, NB for Naïve Bayes, DT for Decision Trees, RFE for Recursive Feature Elimination, L/QDA for Linear/Quadratic Discriminant Analysis, LASSO for least absolute shrinkage and selection operator and GUI for Graphical User Interface. AFNI (which stands for Analysis of Functional NeuroImages) is a set of freely available software packages. The definition of the different feature selection approaches is the following: "wrapper" methods rely on the predictive modeling framework (e.g. using the classification accuracy) to evaluate and select subsets of features; "filters" use other measures (e.g. mutual information) to score and select features; "embedded" methods perform feature selection as part of the model construction process (e.g. LASSO). Stars * mark work under progress/development

Feature	3dsvm	Matlab MVPA	PROBID*	PyMVPA*	Scikit-learn	PRoNTo*
Inputs	AFNI images: fMRI data	AFNI images: fMRI data	NIfTI images: s/fMRI, beta and ASL+ text	NumPy arrays, text, NIfTI images (e.g. s/fMRI and beta), EEP binary file	NumPy arrays + metadata	NIfTI images: s/fMRI and beta
Primary language	C	Matlab	Matlab	Python	Python	Matlab
Multiclass	Yes	Yes	No	Yes	Yes	Yes
Classifiers	SVM	Back propagation (Neural Network Toolbox)	Binary SVM and GP	Includes: kNN, SVM, SMLR	Includes: SVM, NN, GP, DT, L/QDA	Binary SVM, binary and multiclass GP
Regression (machines)	No	No	No	Yes (Includes: GP, LARS, PLR, RR, SMLR)	Yes (Includes: GP, LASSO, Elastic Net, RR, SVM	Yes (group level) (GP, RVR, KRR)
Interfaces	Basic GUI + Command line	Command line	GUI(s) + Command Line	Command Line	Command Line	GUI(s) + Matlab batch + Command Line
GUI for displaying results	No	No	Yes	No	No	Yes
Compatible with	AFNI	Import from/export to AFNI and Brain Voyager*	-	Includes: Import from FSL, Scikit-Learn	Includes: PyMVPA	Import from SPM + SPM Matlab batch
Feature selection	Masking	Masking	Masking	Masking, Filters, Wrappers	Filters, Wrappers, Embedded	Masking
Flexibility of the analysis	Low (fMRI design)	Low (fMRI design)	Low (mainly across groups)	High	High	High

Statistical pattern recognition is a field within the area of machine learning, which is concerned with automatic discovery of regularities in data through the use of computer algorithms, and with the use of these regularities to take actions such as classifying the data into different categories (Bishop, 2006). In PRoNTo, brain scans are treated as spatial patterns and statistical learning models are used to identify statistical properties of the data that can be used to discriminate between experimental conditions or groups of subjects (classification models) or to predict a continuous measure (regression models). In terms of neuroimaging modalities, PRoNTo accepts NIfTI files[6] and can therefore be used to analyze sMRI and fMRI, PET, SPM contrast images or beta maps (obtained from a previous GLM analysis) and potentially any other modality in NIfTI file format.[7] Kernel based[8] classification and/or regression can be performed within or between subject(s), from the same or different group(s), in one or multiple sessions/runs. Binary and multiclass designs are both supported. Its framework allows fully flexible machine learning based analyses and, while its use requires no programming skills, advanced users can easily access technical details and expand the toolbox with their own developed methods. Each step of the analysis can also be reviewed via user-friendly displays. Figure 1 provides an overview of the whole framework.

This paper is structured as follows. In the 'Methods' section we present a brief summary of pattern recognition for neuroimaging data. In 'Results', we present the framework of ProNTo via the analysis of three datasets: a single subject fMRI dataset with multiple runs, an event-related single subject fMRI dataset and a multiple subject sMRI dataset. This section particularly shows how PRoNTo can answer different questions which might be of interest for neuroscientists and describes PRoNTo's functionalities and related issues. Finally, the last sections discuss the limitations of our toolbox and future developments, while summarizing the main advantages and caveats of pattern recognition for neuroimaging.

[6] http://nifti.nimh.nih.gov/nifti-1/

[7] The current version of PRoNTo has however only been tested using the following modalities: fMRI, sMRI, PET, SPM contrast and Fractional Anisotropy (FA) images.

[8] Non-kernel based methods will be included in ensuing versions of the toolbox.

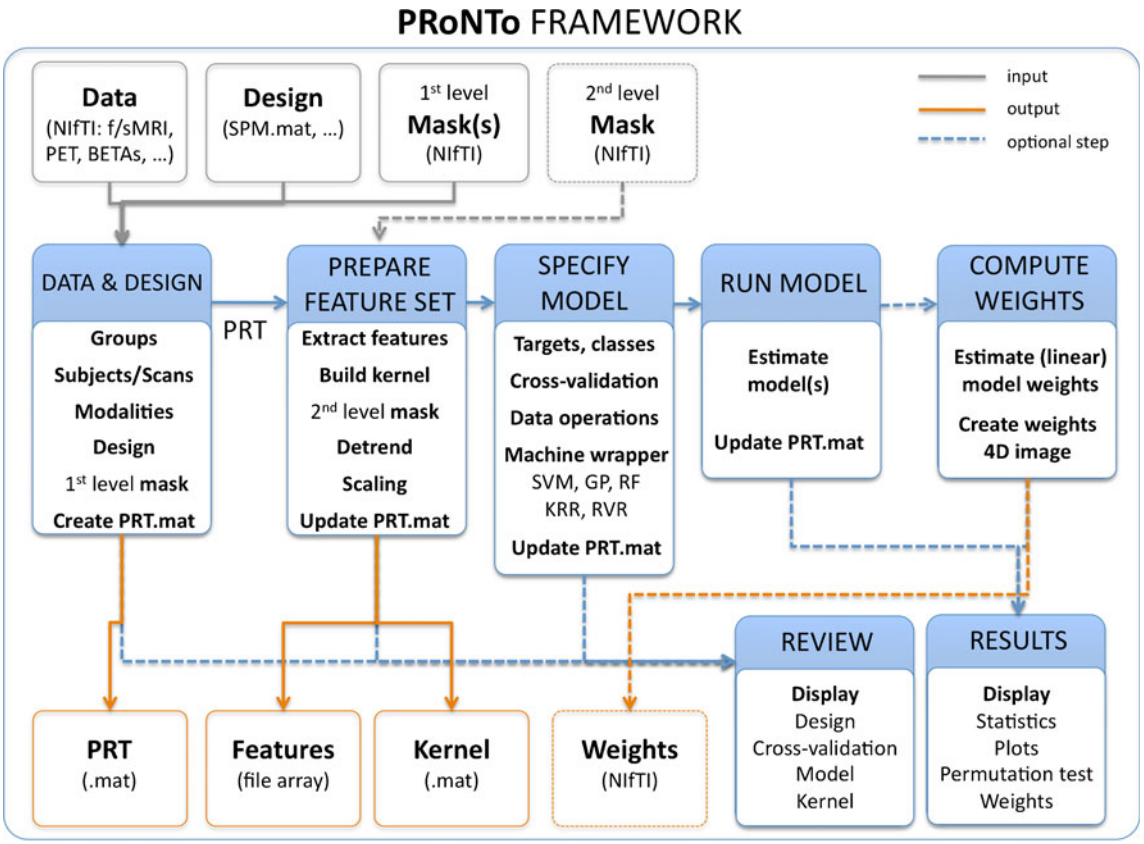

Fig. 1 PRoNTo framework. PRoNTo has five main analysis modules (*blue boxes* in the centre): dataset specification, feature set selection, model specification, model estimation and weights computation. In addition, it provides two main reviewing and displaying facilities (model, kernel and cross-validation displays, as well as, results display). PRoNTo receives as input any NIfTI images (comprising the data and a first-level mask, while an optional second-level mask can also be entered). In addition, when the dataset being analyzed comprises an experimental design, PRoNTo provides more than one way of specifying the design parameters, including loading an SPM.mat file. The outputs of PRoNTo include: a data structure called PRT.mat, a data matrix (with all features), one or more kernels, and (optionally) images with the classifier weights

Methods: Pattern Recognition Analyses

In this section, we present a brief overview of pattern recognition analysis and introduce some basic concepts that will be used in the next sections, however a more complete introduction to machine learning classifiers in the context of neuroimaging can be found elsewhere (e.g. Pereira et al. 2009 and Lemm et al. 2010).

In brief, given a dataset $D=\{\mathbf{x}_i, y_i\}$, $i=1\ldots N$, consisting of pairs of *samples* (or *feature vectors*) $\mathbf{x}_i \in R^d$ and *labels* y_i, the objective in supervised pattern recognition analysis is to learn a function from the data that can accurately predict the labels, i.e. $f(\mathbf{x}_i)=y_i$, of unseen or new patterns. The learned function is called *classifier model* if the labels are discrete values and *regression model* if the labels are continuous values. The dataset is usually partitioned into disjoint sets ('training' and 'test') and analysis proceeds in two phases. During the training phase, an algorithm learns some mapping between patterns and the labels on the training set and during the test phase, the learned function is applied to predict the labels from the unseen samples in the test set. For example, in the linear case, the learned function relies on a linear combination of the feature vectors $\mathbf{x}_i$, i.e. $f(\mathbf{x}_i)= w_0+\mathbf{w}^T\mathbf{x}_i$. The weights $\mathbf{w} \in R^d$ are the model parameters learned in the training phase and represent the relative contribution of each feature to the predictive task.

The core objective of a machine learning model is to generalize from its experience or training (Bishop, 2006). Therefore, the performance of the model is, in general, related to its ability to predict the labels for unseen patterns, which is also defined as the generalization ability of the model. To this end, the predicted labels are compared to the true labels, using the test dataset, from which a measure of accuracy (classifiers) or goodness of fit (regression) is derived. If the performance measure significantly exceeds the level that would be expected by randomly guessing the labels, the researcher can conclude the algorithm has learned some property of the data, and can therefore reject the null hypothesis that there is no information in the data about the label being predicted.

In the context of neuroimaging, the patterns consist of brain scans (voxels values) or measures derived from them (e.g. summarization of regions of interest, cortical thickness). The labels correspond to different mental states (e.g. looking at houses vs. looking at faces) or types of subjects (e.g. patients vs. healthy controls) in classification models or any continuous measure related to the brain scans in case of regression models (e.g. age, performance to task, or degree of illness as measured by a clinical scale). Due to the high dimensionality (d) of the pattern vectors in neuroimaging (~2–500,000 voxels in each image) when compared to the number of examples or scans (N) typically available (usually not exceeding a few hundreds or thousands) most classification/regression problems are ill-conditioned (i.e. d>>N). For this reason, machine learning approaches for neuroimaging have been increasingly relying on kernel methods (LaConte et al. 2005). Kernel methods consist of a collection of algorithms based on pair-wise similarity measures between all examples or patterns, summarized in a *kernel matrix* (N×N dimensions, instead of N×d). Kernel methods are extremely useful, and allow one to perform the learning using the kernel matrix instead of the data matrix which is computationally more efficient if d>N. In addition to the computational advantages, using the kernel formulation together with proper regularization enables the solution of ill-conditioned problems and therefore avoids overfitting (Shawe-Taylor and Cristianini 2004). Another important property of this approach is the fact that non-linear kernel functions can implicitly map the input data space into a higher dimensional feature space, such that the mapping does not have to be explicitly calculated, and one needs only to work with valid kernel matrices. This property is particularly beneficial for problems where N>d, as a dataset requiring non-linear separation in the original space may become linearly separated in the higher dimensional feature space. The application of kernel methods for neuroimaging problems has been growing, see for example (Mourao-Miranda et al. 2011) and (Chu et al. 2011). Further information about kernel algorithms, as well as kernel construction and properties, can be found in (Schölkopf and Smola 2002), (Shawe-Taylor and Cristianini 2004) and (Rasmussen and Williams 2006).

When comparing (multivariate) pattern recognition methods to univariate models, such as those based on the GLM, the major asset of the former lies in the fact that it takes the joint information of all features, as opposed to considering the features as independent from one another. One can also argue that this way of analyzing brain imaging data is closer to how the brain actually functions and to its structural organization. Furthermore, these methods can compute predictions on new samples, which make them suitable as diagnosis or prognosis tools. However, the relevance of each feature/voxel to the multivariate model cannot be interpreted in the same way we interpret univariate statistical tests. When using univariate techniques, one can perform statistical tests on the coefficients attributed to each voxel and therefore threshold the voxel-wise maps to find brain regions, which are significantly (in a statistical sense) linked to the task. When using multivariate techniques, in particular multivariate linear models, it is the combination of all weights that defines the model. The weights at each voxel are thus dependent on one another and no direct localization inferences or voxel-wise statistical test assuming independence can be performed on them.

Results and PRoNTo Framework

In this section, we introduce PRoNTo by presenting five examples of possible research questions that can be fully addressed within the software framework (Fig. 1). The first one consists of a general neuroscience question: "does the pattern of activation in brain regions A, B and C encode information about a variable of interest?" This variable of interest may be the type of stimulus, disorder or mental process, and can be studied with a variety of experimental designs and datasets. Here we use a single subject fMRI dataset from a block design visual experiment. The second question addresses a fundamental issue when analyzing fMRI data: "how do we account for the hemodynamic response function (HRF) and how much does correcting for the HRF affect the classification results?" To answer this question we again use a single subject fMRI dataset, this time acquired using an event-related design containing events that are confounded by the hemodynamics. We also explore the alternative approach of performing classification using the coefficients from a previous GLM analysis (instead of the preprocessed BOLD signal) as a way of avoiding the pitfalls of correcting for the HRF. The third question consists of a clinical application and might also be applied to different patient populations and image modalities: "Which features lead to the best discrimination between the considered groups?" This question was answered using the preprocessed sMRI images from a multi-subject dataset and classifying young versus old healthy subjects. The fourth question relates to finding the best strategy to deal with continuous measurements instead of categorical information (class 1, class 2, etc.) and was investigated through two possible sub-questions: "Can we predict age from brain scans?" and "Are the classifier's predictions for old subjects correlated with their age?" Finally, the fifth question regards the issue of confounds in the data by taking the example of multi-center acquisitions: "How different are the images acquired in different centers and can we predict where they were acquired?"

PRoNTo can be used in three ways: through a graphical user-interface requiring no programming skills, using the

MATLAB-batch system[9] currently embedded in the SPM framework, or by scripting function calls. It is also important to note that PRoNTo assumes the data have been previously pre-processed using SPM or any similar software generating NIfTI format output.

The following paragraphs describe the detailed analysis of these datasets. We hope the investigation of the questions mentioned above provides a basic knowledge about the PRoNTo framework and inspires neuroscientists to further explore their data using machine learning based predictive models.

Question 1. Does the pattern of activation in brain regions A, B and C encode information about a variable of interest?

This is one of the most fundamental questions in neuroscience research and the answer to this question, in multiple contexts, has provided crucial knowledge of how information is processed in the brain (see for example Haxby et al. 2001 and Kamitani and Tong 2005). Here we describe a simple brain decoding analysis using a dataset that has been previously used in pattern recognition for neuroimaging studies (Haxby et al. 2001; Hanson et al. 2004; O'Toole et al. 2005) and for describing the functionalities of other software toolboxes (Hanke et al. 2009a, b).

The data consists in a block design fMRI experiment acquired using a visual paradigm, where the participants passively viewed greyscale images of eight categories: pictures of faces, cats, five categories of manmade objects (houses, chairs, scissors, shoes, and bottles), and control, non-sense images. We chose to analyze the data from a single subject (participant 1), consisting of 12 runs, each comprising eight blocks of 24 s showing one of the eight different object types and separated by periods of rest. Each image was shown for 500 msec followed by a 1500 msec inter-stimulus interval. Full-brain fMRI data were recorded with a volume repetition time of 2.5 s. Each category block therefore corresponds roughly to nine scans, separated by six scans of rest. For further information on the acquisition parameters, please consult the original reference (Haxby et al. 2001).[10]

The data were pre-processed using SPM8. We motion corrected, segmented and normalized the scans according to the MNI template. No smoothing was applied to the data. To test the information encoded in different brain regions we created several anatomical masks to constrain the analyses to: the whole-brain, visual cortex, fusiform gyrus, cerebellum, brainstem and one control region, a manually drawn 16 mm radius sphere outside the brain.[11] The anatomical masks were extracted using the LONI Probabilistic Brain Atlas (LPBA40)[12] maximum probability template and a high-resolution structural image provided with the same dataset (Haxby et al. 2001).

The rest of the analysis was fully performed using PRoNTo. The data were linearly detrended and mean centered across samples. For simplicity, we chose to classify only *faces* versus *houses*. We used a leave one-block (run) out cross validation approach (Fig. 2), resulting in a total of 12 folds and repeated the procedure using features from different brain regions as defined by the previously created masks. Irrespective of the cross-validation scheme it is important to ensure that the test and training data are independent for each fold.

PRoNTo allows the user to specify a first-level mask (common to all subsequent analyses, except for the control region analysis), which is used only for the purpose of making feature extraction and detrending (or scaling in the case of PET) more efficient, and a second-level mask to test specific anatomical hypotheses such as our Question 1.

In the current version of PRoNTo, two kernel classification algorithms are embedded in the framework: Support Vector Machines (SVM, Burges 1998, LIBSVM implementation[13])[14] and (binary and multiclass) Gaussian Process classification (GPC, Rasmussen and Williams, 2006, GPML toolbox[15]). All algorithms are wrapped into what is called a 'machine', which allows easy integration of new machine learning algorithms, enhancing the exchange of newly developed methods within the community. For this reason, we envisage that the list of algorithms available in PRoNTo will grow in the near future. For the list of regression algorithms, please see below (Question 4a).

In what follows we provide results for SVM and GPC when classifying faces and houses using the different anatomical masks. Table 2 provides the balanced accuracies,

$$Acc = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right), \quad (1)$$

where F/TP and F/TN are the false/true positives and negatives, respectively (please note that True Positives correspond to the percentage of examples of class 1 correctly classified and True Negatives correspond to the percentage of examples of class 2 correctly classified), as well as the corresponding p-values (obtained using a permutation test with 100 repetitions) for each classifier and mask. Permutation tests are preferable when the assumption of

[9] Developed at http://fbi.uniklinik-freiburg.de/.

[10] This dataset is freely available to download from the PyMVPA data archive: http://www.pymvpa.org/datadb.html.

[11] The control region was created using MRICRON (http://www.nitrc.org/projects/mricron).

[12] http://www.loni.ucla.edu/Atlases/LPBA40.

[13] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[14] The SVM C parameter that accounts for the trade-off between the width of the SVM margin and the number of support vectors is set to its default value of 1. The next releases of PRoNTo will allow the user to optimize this parameter using nested cross-validation.

[15] http://www.gaussianprocess.org/gpml/code/matlab/doc/

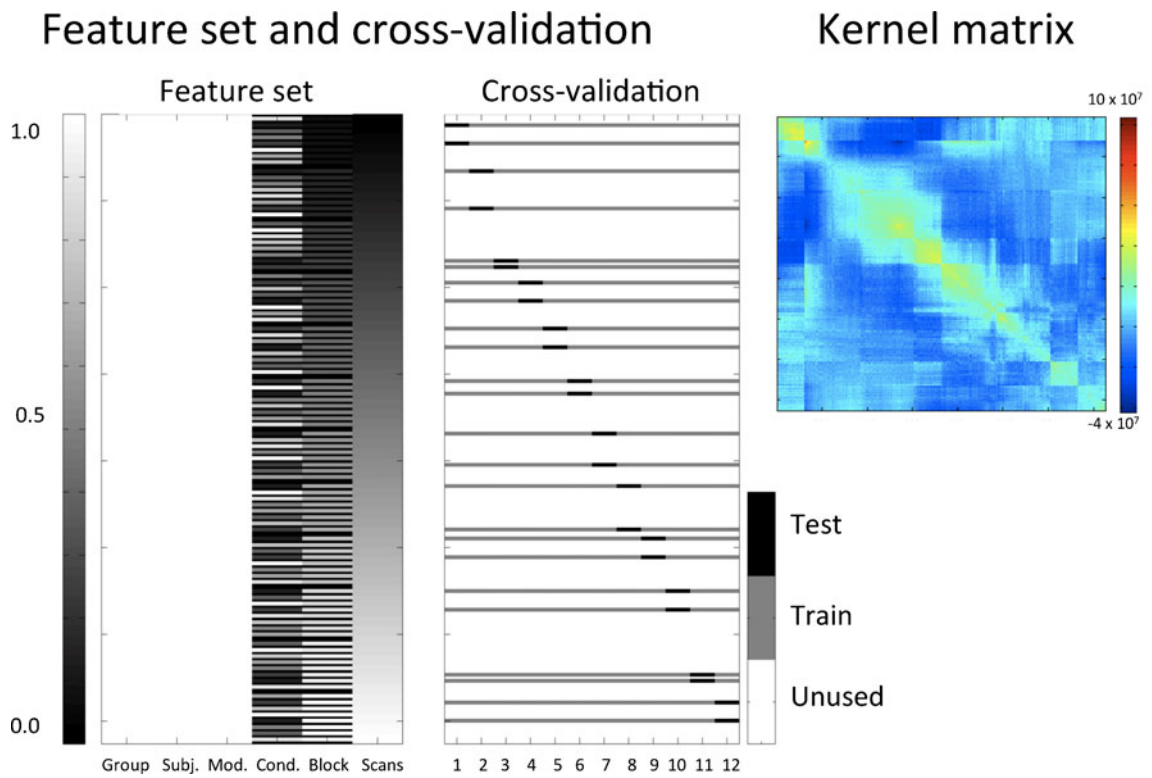

Fig. 2 Feature set (*left*), cross-validation scheme (*middle*) and kernel (*right*) used in Question 1. The feature set consisted of a single data modality from a single subject in a single group but with 9 conditions (8 objects + rest) randomly repeated 12 times. Scan and block index, as well as stimulus type are colorcoded in the left plot. Here the blocks correspond to the chunks into which the data were split in order to build the data matrix file (please consult PRoNTo's manual for more information) and avoid memory problems. To classify the visual stimuli presented to the subject (faces or houses) we used a leave one run out cross-validation scheme. As can be see, this results in 12 folds (corresponding to the twelve experimental runs). Since only the 'faces' and 'houses' are used in the classification most of the images are not used (in *white*) in the training (*grey*) and testing (*black*). The kernel was constructed using all scans (1452 in total), and therefore is a 1452×1452 matrix

independence between test examples does not hold (Golland and Fischl, 2003). In this case, even though the training and test data are independent, the test samples in each block are highly correlated. Another option available in PRoNTo to deal with this issue is to compress the samples within each block (temporal compression) and test only on one image per block.

Table 2 Balanced accuracies and corresponding p-values (obtained using a permutation test with 100 repetitions) for two classifiers (SVM and GPC) and different masks. These results were obtained using the Haxby et al. (2001) dataset when trying to classify houses versus faces using one subject and a leave one block out cross validation scheme. The asterisk indicates a p-value<0.05

Classifier	SVM (%)	GPC (%)
Masks	Acc (p-value)	Acc (p-value)
Whole-brain	94.00 (0.01)*	88.40 (0.01)*
Visual cortex	99.50 (0.01)*	97.70 (0.01)*
Fusiform g.	99.50 (0.01)*	99.10 (0.01)*
Cerebellum	69.90 (0.02)*	66.20 (0.07)
Brainstem	63.90 (0.04)*	65.70 (0.09)
Control	60.20 (0.06)	60.20 (0.14)

The results found are similar to the ones reported originally in Haxby et al. (2001), where the classification was performed by comparing the within-category and between-category pattern correlations on even and odd numbered runs, using object-selective voxels in ventral temporal cortex. As can be seen, both SVM and GPC classifiers achieve very high accuracies when using the visual cortex mask. As expected, this part of the cortex processes visual stimuli and therefore contains useful information to distinguish between different types of visual input (in this case, faces and houses). The increase in accuracy achieved when using this mask compared to the whole brain suggests, as expected, that we are discarding a lot of irrelevant information related to other processes other than distinguishing between visual objects. In addition, limiting the features to the fusiform gyrus, which activates maximally with faces (as found in previous univariate analysis, such as Henson et al. (2002)) did not decrease the accuracy. This result indicates that the face-specific information processed in this part of the cortex is sufficient for significant discrimination between faces and houses.

Using the cerebellum and brainstem masks substantially decreased the accuracy results, as expected, confirming that

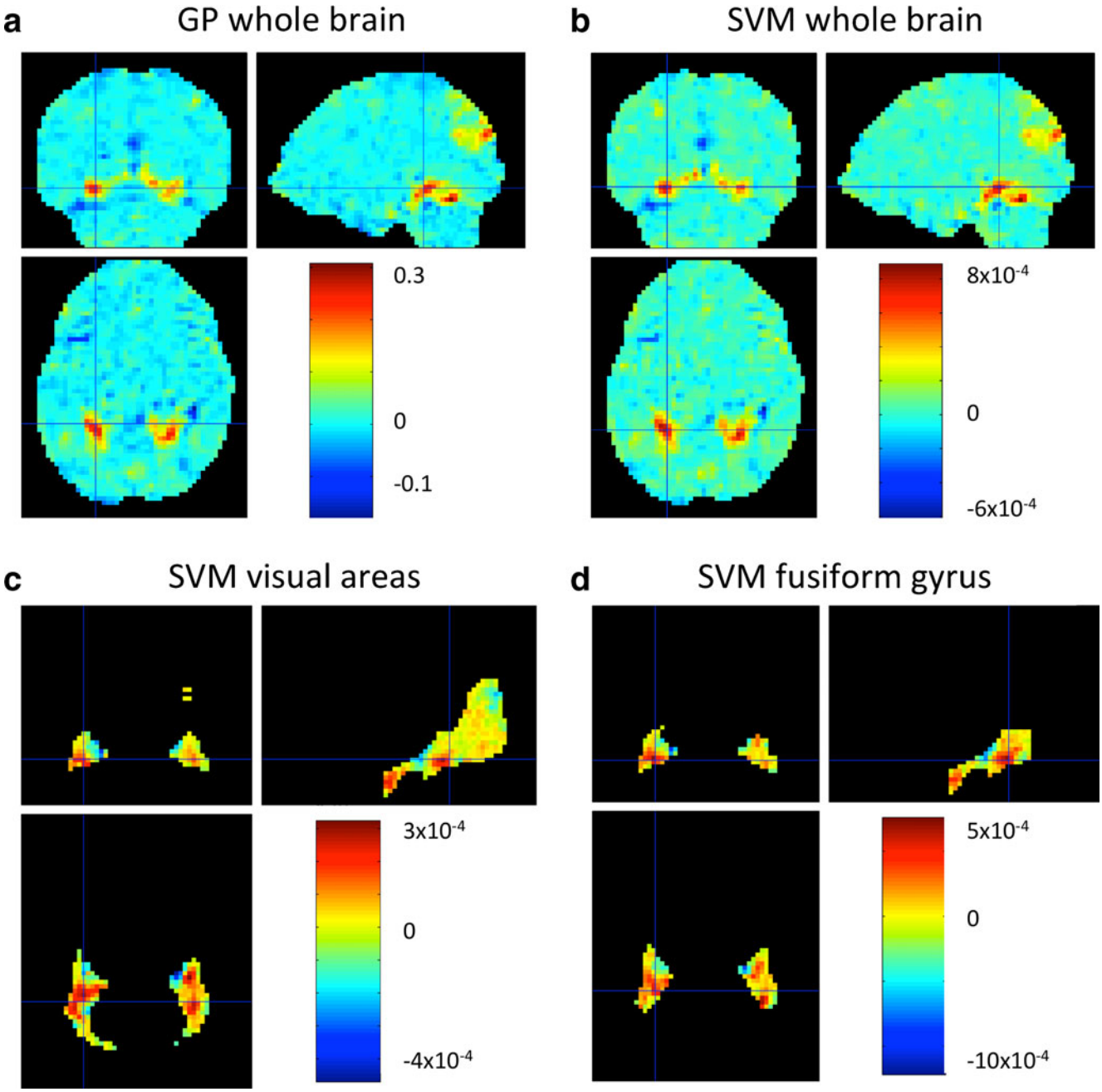

Fig. 3 Model weights obtained with GP using the whole-brain mask (**a**) and model weights obtained with SVM using the whole brain (**b**), visual areas (**c**) and fusiform gyrus (**d**) masks. These weights are for the block design fMRI single-subject dataset. The discrimination task involved classifying the category (faces versus houses) of the object seen by the subject

the patterns of activation in these areas contain less information encoding differences between visual stimuli. In addition, as a sanity check, we verified that when using the control region as features, none of the classifiers provided significant results (classification remained at chance level), since there is no information in the signal to discriminate between the tasks.

PRoNTo also allows the user to create images (NIfTI files) comprising the weight vectors output by linear classifiers. Figure 3 shows whole-brain, visual cortex and fusiform gyrus weight vectors created using SVM and GPC (averaged across all cross-validation folds). Please note that the weight maps are displayed without a threshold or statistical test. This results from the fact that due to the multivariate nature of the patterns, spatial inference on the weights cannot be performed using univariate statistics. The authors intend to develop an original and more easily interpretable way of inferring relevant brain areas from the model weights. However, one can still navigate these images to

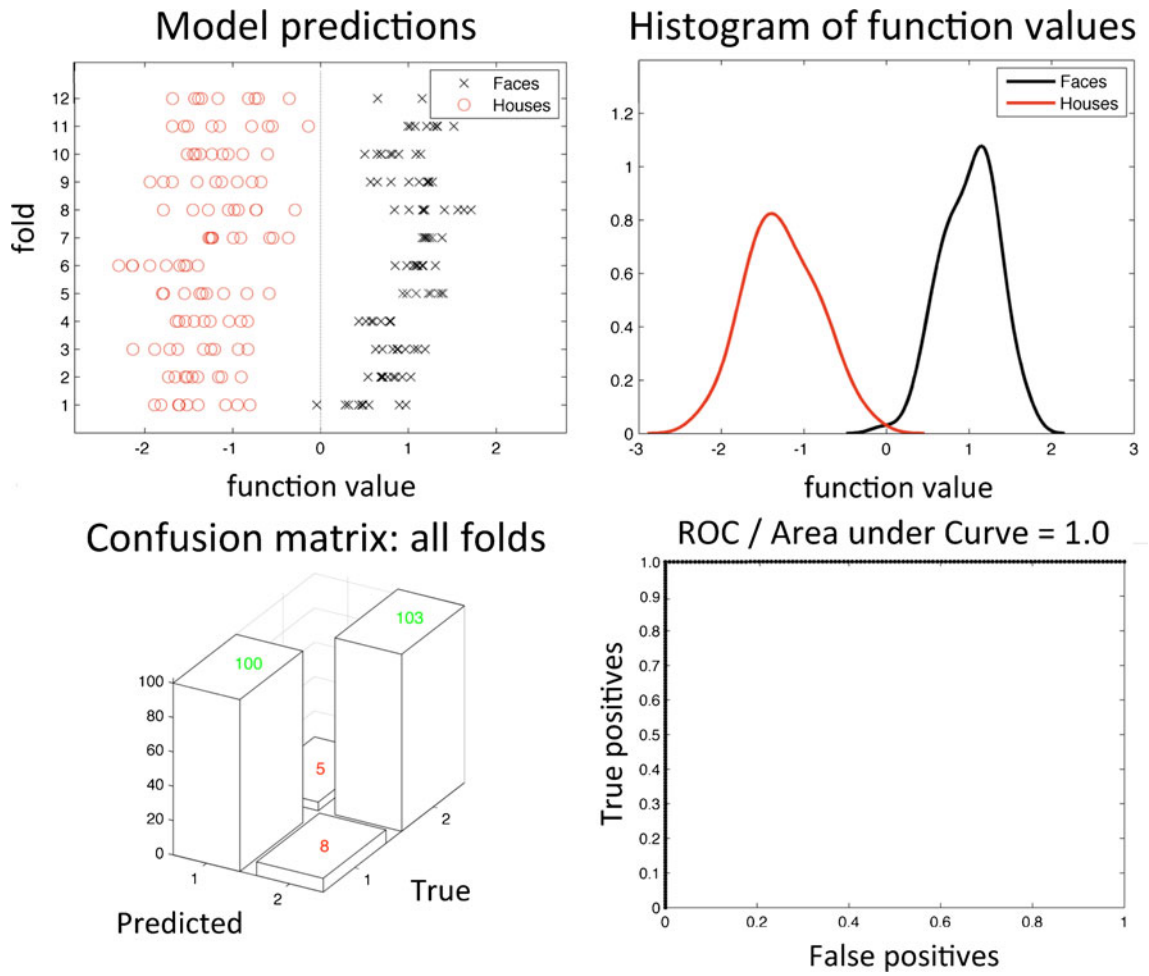

Fig. 4 Plot types provided in PRoNTo for classification approaches. Together with the model weights, PRoNTo allows the user to plot: the prediction values (per fold), histograms of the function values for each class, Receiver-Operating Characteristic Curves (ROC) and 3D confusion matrices. All of these plots are available for each model and cross-validation fold (including average of all folds), and were here plotted for SVM using the visual cortex mask

identify the most discriminative voxels for each mask. For example, the most discriminative voxels from the whole brain analysis lie in visual cortex areas, including the fusiform gyrus (Fig. 3). The patterns are also consistent across classifiers.

In addition to the weights, one can also use PRoNTo to display other information about the models. Figure 4 shows the prediction values, histogram of function values, confusion matrix and Receiver Operating Characteristic (ROC) curve output by SVM using the visual cortex mask. The results correspond to the average of all folds.

To conclude, we have shown that with a single-subject fMRI dataset and different anatomical masks it is possible to show that some brain regions encode information related to a particular stimulus. In the example dataset we used, we verified that the pattern of activation in the visual cortex and face-processing fusiform gyrus contained enough information to distinguish between at least two types of visual stimuli (faces and houses). On the contrary, using the control region did not provide significant discrimination results, as expected, since the signals from outside the brain should not encode information about visual stimuli.

Question 2. How do we account for the HRF and how much does correcting for the HRF affect the classification results?

When working with BOLD (blood-oxygen-level-dependent contrast) fMRI time-series, especially in highly overlapping event-related designs, there is an important issue that needs to be carefully addressed before further analyses. As is well known, the HRF is a delayed and dispersed version of the underlying neuronal response to an experimental event (Fig. 5a). Depending on the TR (repetition time), the effect of the HRF can be felt over multiple scans, which leads to temporal contamination across samples. This can confound subsequent machine learning based analyses and needs to be accounted for. In PRoNTo, the user can control for two parameters, which determine the shape of the HRF: the HRF

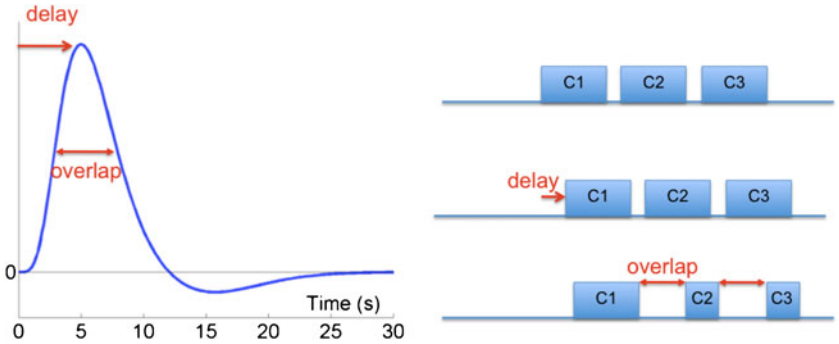

Fig. 5 HRF correction. On the left is the standard HRF response. On the right is the effect of the delay and overlap on the number of independent scans (C1, C2 and C3 correspond to three different experimental conditions and the blue boxes correspond to various scans acquired during each condition). In fMRI datasets, the nature of the HRF (i.e. being a delayed and dispersed version of the neuronal response to an experimental event) might lead to less independent scans/events than the ones originally acquired. In PRoNTo, this issue is accounted for by discarding overlapping scans in terms of BOLD signal

delay (time it takes for the hemodynamic response to peak after the stimulus), which will shift the onsets in time, and the HRF overlap (i.e. the dispersion of the HRF). Given this overlap, scans in which the BOLD signal corresponds to more than one condition, are discarded and not included in further analyses (Fig. 5b).

Here we use a single subject event-related fMRI dataset to show how much the delay and overlap parameters accounting for the HRF affect the classification results. The dataset is freely available from the SPM website[16] and comprises a repetition priming experiment, where two sets of 26 familiar (famous) and unfamiliar (non-famous) faces were presented against a checkerboard baseline. A random sequence of two presentations of each face was created from each set. The faces were presented for 500 ms with a stochastic distribution of stimulus onset asynchrony (SOA) determined by a minimal SOA of 4.5 s and 52 randomly interspersed null events. The subject was asked to make fame judgments by making key presses. Whole brain fMRI data were recorded with a volume repetition time of 2 s. For further information on the acquisition parameters, please consult the original work (Henson et al. 2002). The data were pre-processed using SPM8. This included motion correction, segmentation and normalization to the MNI template. No smoothing was applied.

After preprocessing we ran all analyses using PRoNTo. We used a whole-brain mask and SVM to classify between famous and non-famous faces using the second presentation of each stimulus in a leave one-block-out cross validation scheme. In this case each block corresponds to an event (face presentation). Figure 6 shows the accuracy of the classifier as a function of the hemodynamic delay and overlap parameters. The analysis was repeated for each set of these parameters, which varied from 0 to 15 s in 0.5 s intervals. As can be seen in Fig. 6, the accuracy changes substantially (from a minimum of 40 % to a maximum of 83 %) with different parameter sets, and therefore correcting for the HRF should always be carefully considered. Ideally, these parameters should be estimated from the data, and the authors are currently working on an HRF optimization approach for machine learning based analyses.

As an alternative to correcting for the HRF using a set of arbitrary parameters, one can first run a univariate GLM analysis and use the images of the estimated coefficients (beta images) as the inputs to the classifier. This way the HRF is accounted for in the GLM by convolving the stimulus time-series, or regressors, with a canonical hemodynamic function. Here we used the same repetition priming dataset to classify famous versus non-famous faces but first we fitted a GLM in all voxels within the brain, using SPM8. The design matrix comprised as many columns as events (all famous and non-famous faces presented, in order to obtain a beta image per event) plus the movement parameters and the mean regressor. The betas corresponding to the second repetition of famous and non-famous faces were used for classification using SVM and a leave one-subject (sample) per group out cross-validation scheme. PRoNTo is a highly flexible framework that accommodates multiple types of experimental designs. For instance, here we used the between group classification (and a leave one subject per group out cross-validation) to discriminate between the betas from famous and non-famous faces (each group comprised one type of beta images, e.g. famous faces, and one subject here corresponds to one beta image). Different groups can also correspond to different populations of subjects (such as patients and healthy controls).

The accuracy obtained using the beta images instead of the preprocessed BOLD signal was found to be 73 %, $p=0.01$ (whole brain). This means that the patterns of GLM coefficients carry enough information to discriminate between the stimuli. The weight maps show that the most discriminative voxels were located in the hippocampus areas (Fig. 7), which is consistent with the fact that the

[16] http://www.fil.ion.ucl.ac.uk/spm/

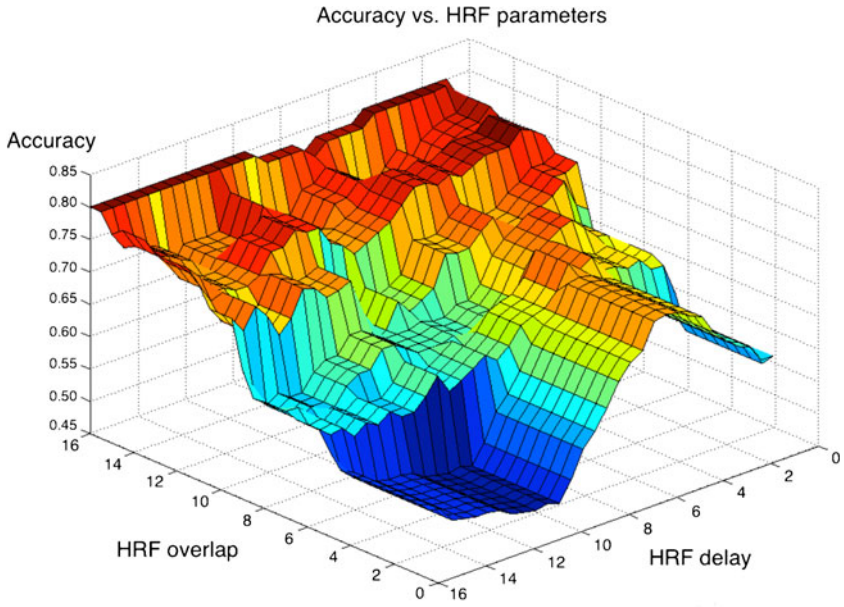

Fig. 6 Classifier accuracy as a function of the HRF parameters. We varied the HRF parameters (overlap and delay) between 0 and 15 s and plotted the accuracy of SVM in discriminating between famous and non-famous faces on a prime repetition event-related single subject fMRI dataset

hippocampus has been shown to activate when forming face memories (Kapur et al. 1995). The fact that the voxels in visual cortex do not seem to contribute as much as the hippocampus to the discrimination task suggests that by using the betas, instead of the preprocessed BOLD signal, we have accounted for some of the main effects of the visual stimulus and are now looking for subtler processes, such as memory, to discriminate between the two conditions.

Question 3. Which features lead to the best discrimination between the considered groups?

Due to the various modalities now available to acquire data of brain activity or structure, one might ask which one of these would yield the best accuracy to distinguish between conditions or groups of subjects. One example may concern whether PET or anatomical MRI scans provide a better imaging biomarker for Alzheimer's disease. Furthermore, the software packages for data preprocessing can produce different aspects of a same imaging modality (e.g. segmenting a sMRI image into grey and white matter density maps), which leads to the same question.

To demonstrate how to answer this sort of question with PRoNTo, we used the IXI dataset,[17] which consists of sMRI images (T1 and T2 sequences) of healthy subjects ranging from 20 to 90 years old and acquired in three centers. Images were segmented into different tissue types via the "new segmentation" algorithm (Ashburner and Friston, 2005) implemented for SPM8. Rigidly aligned grey and white matter maps, down-sampled to 1.5 mm isotropic resolution, were then used to diffeomorphically register all subjects to their common average, using a matching term that assumed a multinomial distribution (Ashburner and Friston, 2009). Registration involved estimating initial velocities, from which the deformations were computed by a geodesic shooting procedure (Ashburner and Friston, 2011). Please note that this registration might induce a slight bias in the accuracies for the current work, as it used a population average template (which incorporated the test as well as the training data). To ensure no bias, one should make sure that the training and test sets are separated at all steps of analysis, from the preprocessing to the modeling, especially if feature selection strategies are involved in the analysis.

Two sets of features were generated from the registration. The first of these was computed from the divergence of the

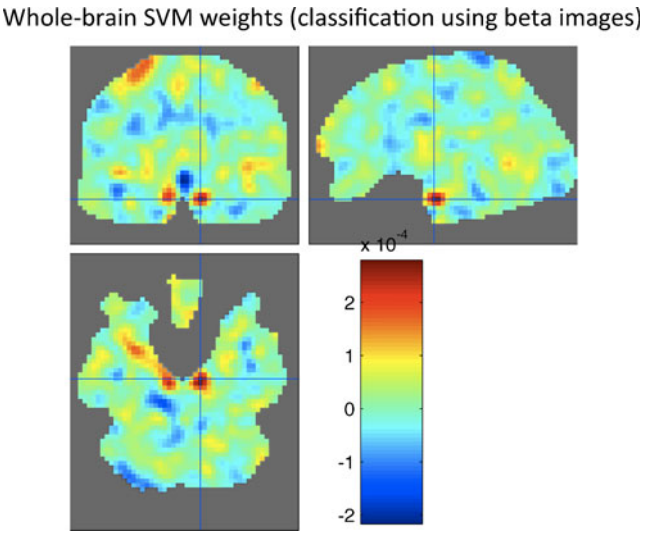

Fig. 7 Whole-brain model weights obtained with SVM using the beta images (GLM coefficients) instead of the preprocessed BOLD signal for the famous versus non-famous faces dataset

[17] IXI - Information eXtraction from Images, funded by EPSRC GR/S21533/02, http://www.brain-development.org/

initial velocity, and encodes the rate of volumetric expansion used by the registration. This measure is similar to the logarithms of the Jacobian determinants of the deformations. The second measure was the "scalar momentum" (Singh et al, 2010), which was spatially smoothed by convolving with a Gaussian of 10 mm full width at half maximum. These particular features were chosen because we had previously examined the effectiveness of a number of features (including Jacobian determinants, rigidly aligned GM, spatially normalised GM and Jacobian scaled spatially normalised GM) derived from the same IXI dataset (unpublished work). The divergences were found to perform better than other features when not using spatial smoothing. When smoothing is used, the scalar momentum smoothed by about 10 mm FWHM was found to outperform the other types of features. In this work, we compared the effectiveness of these feature sets for discriminating between 'young' (20 to 30 years old, 99 subjects) and 'old' (60 to 90 years old, 170 subjects) healthy volunteers (Fig. 8). Please note that the comparison is based on the "nature" of the features (i.e. divergence versus scalar momentum), rather than on subsets of relevant features (voxels) within the feature set, which is a question addressed by feature selection algorithms (not available in the current version of PRoNTo, see "future work" below).

The two feature sets were modeled independently using both an SVM and a GP classifier. The features were mean centered and cross-validation was performed on the basis of a Leave-One subject-Out scheme. The results are displayed in Table 3. These show a slightly better balanced accuracy for the scalar momentum based model. The Bayes' factor of 91.33 computed from the negative marginal likelihoods of the two GP models (Fig. 9) suggests that there is 'strong evidence' (according to Jeffreys' 'Theory of Probability' (Jeffreys 1961)) that the scalar momentum based model is more plausible than the divergence based model.

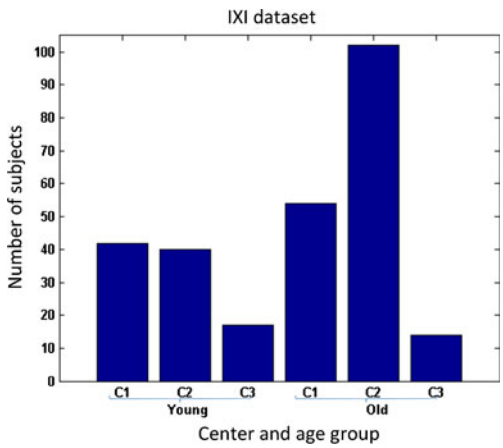

Fig. 8 Subset of the IXI dataset chosen for further analysis. It comprises data from young (20–30) and old (60–90) healthy subjects, which were acquired in three different centers (c1, c2 and c3)

Table 3 Balanced accuracies and p-values (SVM only, for computational reasons) for discriminating between 'young' (20 to 30 years old, 99 subjects) and 'old' (60 to 90 years old, 170 subjects) healthy volunteers using different types of features (scalar momentum and divergences) and different classifiers (SVM and GPC). The asterisk indicates a p-value<0.05

Classifier	SVM (%)	GPC (%)
Features	Acc (p-value)	Acc
Scalar momentum	99.00 % (0.002) *	99.00 %
Divergences	98.10 % (0.010) *	97.00 %

While knowing which feature set leads to the best modeling might be interesting, it would also be useful to consider both feature sets jointly, combining the information they contain. In the near future, the authors intend to provide a set of operations to perform multi-modal classification/regression.

Question 4. How to deal with continuous measurements?

This question can be addressed in two main ways: the first one is to correlate the predictions from the classifier with continuous data, while the second involves predicting the value of the continuous variable from the data using a regression model. The two options might be desirable, depending on the data and clinical measurements. One may indeed want to test two effects, such as the age or performance in a behavioral memory fMRI task for example. In this case one could use different modeling strategies. One possibility is to train a classifier to discriminate young versus old based on their patterns of brain activation during the task and then correlate

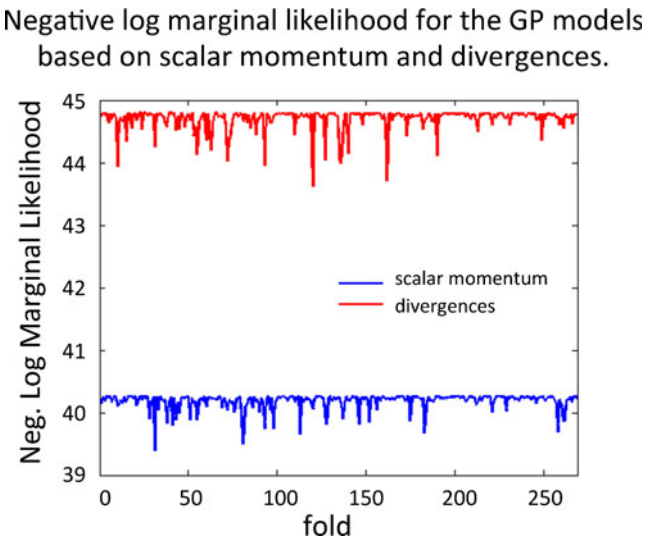

Fig. 9 Negative log marginal likelihood for the GP models based on scalar momentum and divergences. X-axis: folds. Y-axis: negative log marginal likelihood (NLML) of the GP model based on divergences (in *red*) and scalar momentum (in *blue*). For all folds, the NLML values are larger for divergences than for scalar momentum suggesting that the scalar momentum based GP model is more plausible than the divergences based GP model

the classifier's predictions with the subject's performance. Another possibility would be to learn directly the relationship between performance and brain activation by training a regression model to predict the subject's performance based on their patterns of brain activation during the task. Finally, one could also use the age as a covariate to account for age effect in the regression model (i.e. regress out the age effect), therefore discarding the 'interaction' effect.

The IXI dataset was used to test the two main ways of considering continuous inputs. Please note that considering a regression with a covariate is not possible in the first version of the software, but will be feasible within PRoNTo's framework: confound effects will be removed from the kernel using a residual forming matrix as in Chu et al. (2011).

- Question 4a: Can we predict age from brain scans?

To address this question, we predicted the age of the 'old' group (age range: 60.01 to 86.32, mean ± std: 68.02±5.88) using a regression model with the scalar momentum feature set. Similar work was previously done by Franke et al (2010). To finesse the answer to Question 3, this regression was also performed on the divergence features. No further operations were applied to the data and a "Leave One Subject Out" cross-validation scheme was used to compute a mean squared error (MSE), as well as the correlation between the predictions and the targets (i.e. the 'true' age).

In PRoNTo, regression can be performed using Kernel-Ridge Regression (KRR, Hastie et al. 2003), Relevance Vector Regression (RVR, Tipping, 2001) or Gaussian Processes Regression (GPR, Rasmussen and Williams, 2006, GPML toolbox). In its current form, the software allows regression only at the group level, i.e. when providing one image and one continuous measurement per subject. This limitation should be overcome in the next version in which multiple continuous measures per subject will be handled. The correlation and MSE for each feature set and regression algorithm are displayed in Table 4.

Please note that for KRR, the hyperparameter λ controlling for the regularization varied as 10^i with i=1…5, leading to the correlation (ρ) and MSE values in Table 5. The results presented in Table 4 are the (rounded) highest values obtained (in bold). This shows that optimizing such hyperparameter can lead to significant changes in correlation and MSE. Therefore nested cross-validation should be implemented in a next version to perform this hyperparameter optimization. This estimation is done automatically for GPR. Essentially, predicting the means using GPR is equivalent to KRR, although GPR usually includes the hyperparameter estimation, and predicts the variances.

Table 4 Correlation between the age of the 'old' group and the model predictions, as well as MSE, for each feature set and regression machine (KRR, RVR and GPR)

Regressor	KRR (%)	RVR (%)	GPR (%)
Features	Corr	Corr	Corr
	MSE	MSE	MSE
Scalar momentum	0.50	0.60	0.60
	28.98	23.90	22.20
Divergences	0.37	0.50	0.50
	31.06	24.90	26.00

Table 5 Effect of the parameter λ of Kernel Ridge Regression on the correlation and MSE values for the two different feature sets (scalar momentum and divergences). The results presented in Table 4 are the (rounded) highest values shown in this Table (in bold)

λ (10^)	Scalar momentum		Divergences	
	Corr	MSE	Corr	MSE
0	0.34	36.53	0.18	53.97
1	0.35	35.07	0.18	53.24
2	0.41	30.10	0.20	48.05
3	**0.50**	**28.98**	0.27	36.24
4	0.50	32.49	**0.37**	**31.06**
5	−0.16	34.42	0.32	32.15

The results show that GPR and RVR performed best for both datasets but the highest correlation and minimal mean squared error were obtained for the scalar momentum feature set. To further compare the two modalities, the absolute difference between the targets and the predictions were computed in terms of years (using the results from GPR). For each fold, it was then checked whether this difference was larger for the divergences or for the scalar momentum. The results showed that the difference is larger for the divergences 91 times out of 170 folds (53.53 %), which is a non-significant result (Friedman test on the absolute differences: p=0.36). Across folds, the RMS is 4.71 for the scalar momentum features while it is 5.10 for the divergence features.

This shows that although the difference between the accuracies (Question 3) and correlations obtained from models based on the two considered features is small (8 % difference), the scalar momentum might contain more information to predict or classify age.

Please note that when using the scalar momentum of all subjects available in the IXI dataset (i.e. age range=19–86), the correlation and mean absolute error (MAE) obtained after RV regression are similar to the results of Franke et al. (2010) when considering no dimensionality reduction and reach values of ρ=0.9 and MAE=5.74. However, the authors of Franke et al. (2010) show that feature selection (Principal Component Analysis (PCA) in the present case) improved the correlation and decreased the MAE. While such feature selection steps are not yet available in PRoNTo, the software was designed to allow the easy

addition of new modules and we therefore invite researchers from the machine learning community to share their work via an implementation in a future version of PRoNTo.

- Question 4b: Are the classifier' predictions for old subjects correlated with their age?

In this case, we considered the correlation between the predictive probabilities obtained from the previous binary GP classifier based on the scalar momentum, which discriminates between 'young' and 'old' and the age of the subjects in the 'old' group, which shows the largest age range (60–90). The GP predictive probability measures the classifier confidence about the class membership of the test example. When using the predictive probabilities, we expected an anti-correlation since the lower the probability the older the subject (model was young=+1 versus aged=-1). The obtained Spearman correlation coefficient (the distribution of the predictive probabilities being not normal according to a Jarque-Bera hypothesis test of composite normality) value was: $\rho=0.01$ with an associated p of 0.86 (Fig. 10). These results are therefore not significant. This can be explained by the fact that the ages are linear values, whereas the probabilities associated to the classifier's predictions are non-linear values, generated by a "softmax" function. In other words, the 'perfect' correlation isn't a straight decreasing line. Therefore, the authors suggest rather using covariates when trying to account for more than one measurement or to 'unsquash' the model probabilities beforehand.

Question 5. Can we distinguish the imaging centers from which the preprocessed images were acquired?

Most institutions, whether at a local, national or international level, encourage collaborations between researchers, which often results in data acquired using the same design but in different centers and on different machines. In this context, an interesting question is to investigate whether the classification we might perform on the multi-center data is robust to data acquisition. Results reported in the literature suggest a decrease in performance when combining data across sites (Klöppel et al., 2008).

In the present work, this question was investigated using the IXI dataset, which consists in data acquired in three centers across London, United Kingdom. The 'young' versus 'old' classification was replaced by a 'center1' (96 subjects), 'center2' (142 subjects), 'center3' (31 subjects) classification. The scalar momentum features were mean centered and divided by their norm. A multiclass Gaussian Process classification was then performed in a "Leave One Subject Out" cross-validation. The obtained accuracies are high: balanced accuracy is 98.7 % and class accuracies are 96.9, 100.0 and 99.3 % for centers 1, 2 and 3, respectively. Figure 11 shows the confusion matrix for this classification.

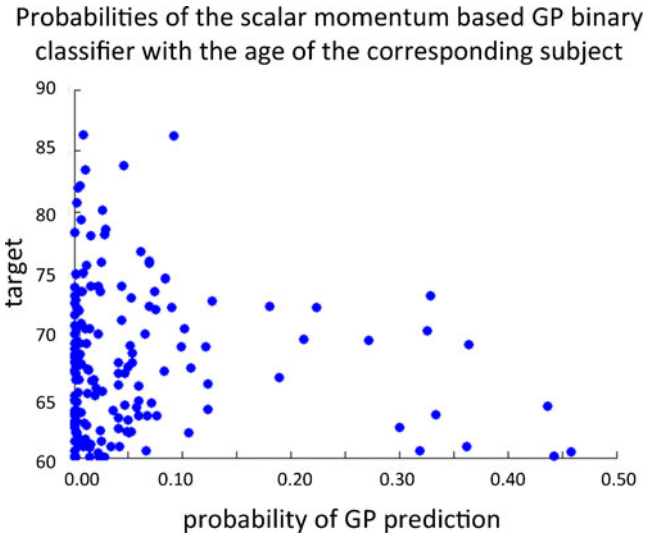

Fig. 10 Scatter plot of the probabilities of the scalar momentum based GP binary classifier with the age of the corresponding subject. This plot shows that no linear relationship could be derived from those values

These results show that the considered data contain sufficient information to classify almost perfectly the different centers. This shows a potential caveat of machine learning based modeling: all differences between categories are modeled, whether these correspond to differences in brain activity/structure or to confounding effects. It is therefore important to match subjects across groups for any potential confound effect such as age, gender or acquisition centers to prevent the classifier from learning confound effects that are correlated

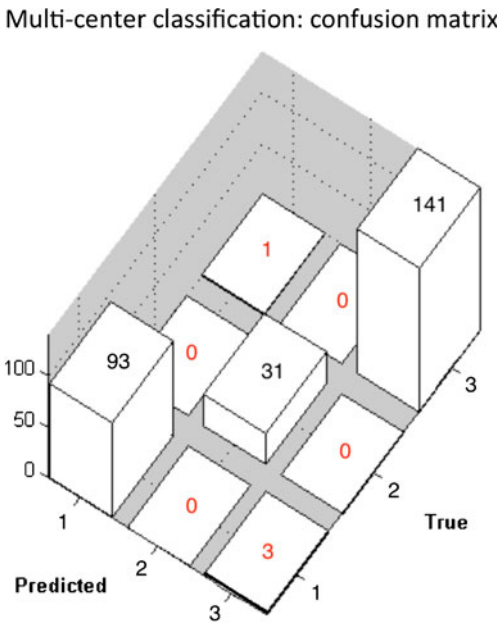

Fig. 11 Confusion matrix obtained from the multiclass GP model to distinguish between centers. The *diagonals* show the largest numbers (by far), which reveals an almost perfect classification of the centers

with the labels instead of the effect of interest. Collaborations involving multi-center data should thus be encouraged as they increase the size of the datasets, and hence their variability, leading to more realistic and useful diagnosis/prognosis models, as long as the subjects are matched across centers or scanner effects are modeled, as in Klöppel et al., 2008.

Discussion

Mass univariate statistical analyses of neuroimaging data, based on the GLM, have been recently complemented by the use of multivariate pattern analyses, based on machine learning models. These models allow an increased sensitivity and flexibility compared to univariate techniques but until recently they lacked an established and accessible software framework.

In this paper, we propose PRoNTo, a MATLAB toolbox for accessible and flexible machine learning based classification/regression of neuroimaging data. While MATLAB is not freely available software, its high-level language, as well as its widespread use in the neuroscientific community, makes it a favorable environment to analyze neuroimaging data. Furthermore, many MATLAB-based toolboxes already exist to import, preprocess and analyze brain data. PRoNTo is therefore fully compatible with the most widely used of them, SPM (Friston et al. 2007), and allows an easy integration of machine learning based MVPA into the currently used pipelines, especially via the MATLAB batch system. Thanks to its graphical interfaces, PRoNTo requires no programming ability whilst allowing fully flexible machine learning based classification/regression. In addition, developers can easily add new algorithms, with very little prior knowledge of the toolbox's functionalities.

One of the main advantages of the pattern recognition approach is its predictive power, which not only helps to investigate brain function but can also be used as a potential diagnostic tool for neurologic and psychiatric disorders (Kloppel et al. 2011). In this paper, we presented our software framework by exploring the type of questions that can be addressed using this methodology and that cannot be properly answered using the GLM approach. These questions include: does the pattern of activation in brain regions A, B and C encode information about a variable of interest? How do we account for the hemodynamic response function in pattern recognition analyses and how much does correcting for the HRF affect classification results? Which type of features best discriminates between groups? Can we predict continuous measures from brain scans, and how do we deal with continuous clinical values? How different are the images acquired in different centers and can we predict where they were acquired? These examples comprise only a limited subset of the variety of neuroscientific questions that can be addressed with pattern recognition for neuroimaging. However, while the assets of the field are becoming well known and recognized, its limitations and technical subtleties are sometimes not properly appreciated and even overlooked.

One common mistake, when using linear models, relates to the temptation of interpreting the model weights images as statistical parametric maps (SPMs). Contrary to SPMs, it is the combination of all weights that defines the model and therefore the weights at each voxel are dependent on one another. No voxel-wise statistical tests assuming independence can be performed on them. This leads to interpretability issues, since most neuroscientists look to find not only how information is encoded in the brain but also where in the brain this information resides. While this is still a topic of much debate, one of the proposed ways of addressing this issue is the use of sparse models (e.g. Zou 2005). These models, through their regularization properties, impose sparsity in the voxel dimension and force many voxel weights to zero. In theory, these models could potentially identify underlying brain networks responsible for the discrimination task and therefore improve model interpretability. However, in practice, there is no way of preventing a sparsity enforcing model from removing more than irrelevant voxels leading once again to difficulties in interpreting its results. Rasmussen et al. (2012) introduce an interesting discussion regarding the interpretability and generalization of the model weights. According to these authors, the most common way of optimizing the models, by maximizing their predictive accuracy, does not lead to the most accurate representation of underlying brain networks. The authors therefore propose to use both an accuracy based and reproducibility measure to assess the model and provide more reliable spatial patterns. Another way to optimize the models, in a probabilistic context, such as is the case of Gaussian Processes, is by maximizing the model evidence. However, the best way to assess model quality and generalization, e.g. using the predictive accuracy or model evidence, still remains an issue to be further investigated.

The attempt to find where in the brain information is located, and in particular, which features (voxels) to use in a classification problem, can lead to another common mistake called 'peeking'. This issue occurs when the choice of features depends on the labels of the entire dataset. This permits information from the test set to influence the learning of the classifier in the training set and, as a result, it can lead to optimistic accuracy estimates. The solution to this issue is to restrict the features to one or more regions of interest (based on apriori anatomical hypotheses) or proceed by doing recursive feature elimination (RFE - Guyon & Elisseeff, 2007; De Martino et al., 2008) in a nested cross-validation context. Once the data has been split into training and testing datasets one can further split the training data and use the labels within this fold to select the most discriminative voxels. In each

feature elimination step, a small proportion of voxels is discarded until a core set of voxels remains with the highest discriminative power. Although more computationally intensive, this technique leads to a sparse solution.

Another sensitive issue, which has not been comprehensively explored, is the effect of the preprocessing steps (such as realignment, normalization, smoothing…) in the classification/regression output. Although some results exist (LaConte et al. 2003), it is still important for further analyses to find out which and how preprocessing parameters affect pattern recognition. Each preprocessing strategy essentially tries to encode some measure of similarity among the scans. Principled ways to derive similarities among scans, which are based upon the generative models to which the data are fit during preprocessing (e.g. Fisher kernels (Jaakkola & Haussler 1998)), are therefore needed. In addition, most experimental designs so far have been 'recycled' from previous GLM analyses. Even though they might be ideal for finding activations in the brain they may be suboptimal for discrimination tasks, and the field would benefit from further investigations into this area.

Another issue is the lack of a clear strategy on how to remove confounds from pattern recognition analyses. This issue can be particularly worrying in clinical settings, where one of the most common confounds is the patients' medication. In other words, in many pattern recognition studies (e.g. Mourao-Miranda et al. 2012), the groups of subjects (healthy controls and patients) differ not only on the presence of the disease but also on the fact that patients are taking medication at the time of the experiment. New methodological advances need therefore to be introduced into the field in order to properly account for this confound. For instance, using Multi-Kernel Learning, one could ask the question: "does the image data plus medication status predict more accurately than medication status alone?"

Finally, even though some aspects of pattern recognition for neuroimaging are still under much debate, for instance regarding the interpretability of its output spatial patterns, it is a very promising methodology for studying the brain and has proven to answer questions that go beyond the scope of existing statistical approaches. With the development of PRoNTo, we hope not only to provide a working tool for neuroscientists but also a platform to motivate the development of the techniques and contribute to the resolution of some of its current limitations. The authors therefore hope to facilitate the interaction between the neuroscientific and machine learning communities. On one hand, the machine learning community should be able to contribute to the toolbox with novel published machine learning models. On the other hand, the toolbox should provide a variety of tools for the neuroscience and clinical neuroscience communities, enabling them to ask new questions that cannot be easily investigated using existing statistical analysis tools.

Future Work

As aforementioned, PRoNTo has a modular design which allows the easy addition of new classifiers/regression algorithms (e.g. sparse models) wrapped as machines. When calling a machine, the main code refers to a key name (e.g. 'svm_bin' for a binary SVM classifier) and provides a structure comprising the training and test data in matrix form, the training labels or values (for classification or regression, respectively), a flag expressing whether the method uses kernels or features and arguments which are specific to the machine (e.g. soft-margin parameter for SVM). To implement a new machine, one would thus need to create an interface between the input structure and the new method, not caring about cross-validation or performance estimation. To fully interface the new machine, a key name should be generated and linked to the main batching/GUI system. More details are provided in the manual. Regarding feature selection algorithms, once nested cross-validation is available, their addition will also be possible and straightforward.

Although PRoNTo already proposes spatial and temporal compression, which are particularly useful in a within-subject context (Mourao-Miranda et al. 2006), feature selection could bring valuable advantages such as reducing memory storage requirements (Formisiano et al. 2008) and possibly improving model accuracy (Guyon and Elisseeff 2007). However, some authors have questioned the benefits of feature selection in neuroimaging applications (Cuingnet et al. 2011; Chu et al., 2012). In addition, approaches that evaluate the information contained within local multivariate patterns, such as the searchlight approach (Kriegeskorte et al., 2006), might also provide new insights about brain functions. Future versions of PRoNTo are likely to comprise methods such as Recursive Feature Elimination (RFE, De Martino et al., 2008) and the searchlight approach. The next release of PRoNTo will also include the possibility of optimizing model parameters through nested cross-validation.

When more than one modality is available, it would be interesting to consider their joint information instead of treating them independently. Therefore, we plan to include different methods of concatenating the features from different imaging techniques, to achieve multimodal classification/regression of the data.

In a clinical context, acquiring data from new patients after the analysis of previous datasets is not rare. In the same way, the label might not be available for some patients/healthy subjects. Therefore, being able to predict the label/continuous value of a new feature vector from a previously built model would be highly desirable and should be included in the near future. Please note that this feature could be applied to validate the model (using what is referred to as a *validation set*), after the training and testing phases.

Finally, in PRoNTo, the weight maps built from linear classifiers are displayed without a threshold or statistical test. This results from the fact that due to the multivariate nature of the patterns, spatial inference on the weights cannot be performed using univariate thresholds or statistics. However, the authors intend to develop an original and more easily interpretable way of inferring relevant brain areas from the model weights.

Conclusions

PRoNTo aims at providing a comprehensive and user-friendly software framework for multivariate analysis based on machine learning models for neuroimaging data. While built to be compatible with SPM, non-SPM users will have no difficulty in using the graphical interfaces. Thanks to its modular design, PRoNTo can easily be extended via the addition of new feature selection and extraction approaches, validation procedures or classification/regression models, therefore aiming to improve the interaction between the neuroimaging and machine learning communities.

Information Sharing Statement

PRoNTo, and all its documentation, are available to download from: http://www.mlnl.cs.ucl.ac.uk/pronto/. The toolbox code is distributed for free, but as copyright software under the terms of the GNU General Public License as published by the Free Software Foundation. PRoNTo is written for MATLAB 7.5 (R2007b) and onwards, and needs an installed version of SPM (versions 8 or above, including the latest updates) to work. Some routines may need to be compiled for your specific OS. For further information on how to use the software, please consult PRoNTo's manual, available here: http://www.mlnl.cs.ucl.ac.uk/pronto/manual.htm. The datasets used in this paper are freely available to the general public and we have provided the websites where they can be downloaded from throughout the manuscript.

Acknowledgments PRoNTo is the deliverable of a Pascal Harvest project coordinated by Dr. J. Mourao-Miranda and its development was possible with the financial and logistic support of:

- the Department of Computer Science, University College London (http://www.cs.ucl.ac.uk);
- the Wellcome Trust support under grant no. WT086565/Z/08/Z (http://www.wellcome.ac.uk/);
- PASCAL2 (http://www.pascal-network.org) and its HARVEST programme;
- the Fonds de la Recherche Scientifique-FNRS (http://www.fnrs.be), Belgium;
- the Portuguese Foundation for Science and Technology (http://www.fct.pt), Portugal;
- Swiss National Science Foundation (PP00P2-123438) and Center for Biomedical Imaging (CIBM) of the EPFL and Universities and Hospitals of Lausanne and Geneva.
- The King's College London Centre of Excellence in Medical Engineering, funded by the Wellcome Trust and EPSRC under grant no. WT088641/Z/09/Z
- Marie Curie Actions (project #299500).

Conflict of interest The authors declare that they have no conflict of interest.

References

Ashburner, J., & Friston, K. J. (2000). Voxel-based morphometry—The methods. *NeuroImage, 11*(6), 805–821.

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *NeuroImage, 26*, 839–851.

Ashburner, J., & Friston, K. J. (2009). Computing average shaped tissue probability templates. *NeuroImage, 45*, 333–341.

Ashburner, J., & Friston, K. J. (2011). Diffeomorphic registration using geodesic shooting and Gauss-Newton optimisation. *NeuroImage, 55*, 954–967.

Bishop, C. (2006). Pattern recognition and machine learning. Springer.

Borroni, B., Di Luca, M., & Padovani, A. (2006). Predicting Alzheimer dementia in mild cognitive impairment patients. Are biomarkers useful? *European Journal of Pharmacology, 545*(1), 73–80.

Burges, C. J. C. (1998). *A tutorial on support vector machines for pattern recognition*. Boston: Kluwer Academic Publishers.

Chadwick, M. J., Hassabis, D., Weiskopf, N., & Maguire, E. A. (2010). Decoding individual episodic memory traces in the human hippocampus. *Current Biology, 20*(6), 544–547.

Chu, C., Ni, Y., Tan, G., Saunders, C. J., & Ashburner, J. (2011). Kernel regression for fMRI pattern prediction. *NeuroImage, 56*, 662–673.

Chu, C., Hsu, A. L., Chou, K. H., Bandettini, P., & Lin, C. (2012). Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *NeuroImage, 60*(1), 59–70.

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

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M. O., et al. (2011). Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. *NeuroImage, 56*(2), 766–781.

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., & Formisano, E. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fmri spatial patterns. *NeuroImage, 43*, 44–58.

Formisiano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). "Who" is saying "What"? brain-based decoding of human voice and speech. *Science, 322*(5903), 970–973.

Franke, K., Ziegler, G., Klöppel, S., Gaser, C., et al. (2010). Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. *NeuroImage, 50*, 883–892.

Friston, K. J., et al. (2007). *Statistical parametric mapping: the analysis of functional brain images*. London: Elsevier Academic Press.

Golland, P., & Fischl, B. (2003). Permutation tests for classification: Towards statistical significance in image-based studies. LNCS. *Proceedings of IPMI: International conference on information processing and medical imaging*, 2732. Springer, 330–341.

Guyon, I., & Elisseeff, A. (2007). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., & Pollmann, S. (2009a). PyMVPA: a python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics, 7*, 37–53.

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Olivetti, E., Fründ, I., Rieger, J. W., et al. (2009b). PyMVPA: a unifying approach to the analysis of neuroscientific data. *Frontiers in Neuroinformatics, 3*, 3.

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

Hastie, T., Tibshirani, R., & Friedman, J.H. (2003). *Elements of statistical learning*. Springer.

Haxby, J., Gobbini, M. I., Furev, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*(5539), 2425–2430.

Haynes, J., & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience, 8*, 686–691.

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

Haynes, J., Sakai, K., Rees, G., Gilbert, S., Frith, C., & Passingham, R. E. (2007). Reading hidden intentions in the human brain. *Current Biology, 17*(4), 323–328.

Henson, R. N. A., Shallice, T., Gorno-Tempini, M.-L., & Dolan, R. J. (2002). Face repetition effects in implicit and explicit memory tests as measured by fMRI. *Cerebral Cortex, 12*, 178–186.

Jaakkola, T., & Haussler, D. (1998). Exploiting generative models in discriminative classifiers. *In Advances in Neural Information Processing Systems, 11*, 487–493.

Jeffreys, H. (1961). *The theory of probability* (3 ed.). Oxford.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience, 8*, 679–685.

Kapur, S., Craik, F. I. M., Brown, G. M., Houle, S., & Tulving, E. (1995). Functional role of the prefrontal cortex in memory retrieval: a PET study. *Neuroreport, 6*, 1880–1884.

Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., et al. (2008). Automatic classification of MR scans in Alzheimer's disease. *Brain, 131*, 681–689.

Kloppel, S., Abdulkadir, A., Jack, C. R., Jr., Koutsouleris, N., Mourao-Miranda, J., & Vemuri, P. (2011). Diagnostic neuroimaging across diseases. *NeuroImage, 61*(2), 457–463.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *PNAS, 103*, 3863–3868.

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*(1), 10–27.

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

Lemm, S., Blankertz, B., Dickhaus, T., & Müller, K. (2010). Introduction to machine learning for brain imaging. *NeuroImage, 56*, 387–399.

Marquand, A., Howard, M., Brammer, M., Chu, C., Coen, S., & Mourao-Miranda, J. (2010). Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes. *NeuroImage, 49*(3), 2178–2189.

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

Mourao-Miranda, J., Hardoon, D. R., Hahn, T., Marquand, A. F., Williams, S. C., Shawe-Taylor, J., et al. (2011). Patient classification as an outlier detection problem: an application of the one-class support vector machine. *NeuroImage, 58*(3), 793–804.

Mourao-Miranda, J., Almeida, J., Hassel, S., de Oliveira, L., Versace, A., Marquand, A., et al. (2012). Pattern recognition analyses of brain activation elicited by happy and neutral faces in unipolar and bipolar depression. *Bipolar Disorders, 14*(4), 451–460.

Mourão-Miranda, J., Oliveira, L., Ladouceur, C. D., Marquand, A., Brammer, M., et al. (2012). Pattern recognition and functional neuroimaging help to discriminate healthy adolescents at risk for mood disorders from low risk adolescents. *PLoS One, 7*(2), e29482. doi:10.1371/journal.pone.0029482.

O'Toole, A. J., Jiang, F., Abdi, H., & Haxby, J. V. (2005). Partially distributed representations of objects and faces in ventral temporal cortex. *Journal of Cognitive Neuroscience, 17*(4), 580–590.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *Journal of Machine Learning Research, 999888*, 2825–2830.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage, 45*, 199–209.

Phillips, C. L., Bruno, M. A., Maquet, P., Boly, M., Noirhomme, Q., Schnakers, C., et al. (2011). "Relevance vector machine" consciousness classifier applied to cerebral metabolism of vegetative and locked-in patients. *NeuroImage, 56*(2), 797–808.

Polyn, S. M., Natu, V. M., Cohen, J. D., & Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science, 310*(5756), 1963–1966.

Rasmussen, C.E., & Williams, C.K.I. (2006). *Gaussian processes for machine learning*. The MIT Press.

Rasmussen, P. M., Hansen, L. K., Madsen, K. H., Churchill, N. W., & Strother, S. C. (2012). Model sparsity and brain pattern interpretation of classification models in neuroimaging. *Pattern Recognition, 45*, 2085–2100.

Richiardi, J., Gschwind, M., Simioni, S., Annoni, J.-M., Greco, B., Hagmann, P., Schluep, M., Vuilleumier, P., Van De Ville, D. (2012). Classifying minimally-disabled multiple sclerosis patients from resting-state functional connectivity, NeuroImage

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge University Press.

Schölkopf, B. & Smola, A. J. (2002). *Learning with kernels*. MIT Press.

Schrouff, J., Kussé, C., Wehenkel, L., Maquet, P., & Phillips, C. (2012a). Decoding semi-constrained brain activity from fMRI using support vector machines and gaussian processes. *PLoS One, 7*(4), e35860. doi:10.1371.

Schrouff, J., Kussé, C., Wehenkel, L., Maquet, P., & Phillips, C. (2012b). Decoding spontaneous brain activity from fMRI

using Gaussian Processes: tracking brain reactivation, *Proceedings of International Workshop on Pattern Recognition in Neuroimaging*.

Shinkareva, S. V., Mason, R. A., Malave, V. L., Wang, W., Mitchell, T. M., et al. (2008). Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. *PLoS One, 3*(1), e1394. doi:10.1371/journal.pone.0001394.

Singh, N., Fletcher, P., Preston, J., Ha, L., King, R., Marron, J., et al. (2010). *Multivariate statistical analysis of deformation momenta relating anatomical shape to neuropsychological measures*. Springer: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2010.

Spiridon, M., & Kanwisher, N. (2002). How distributed is visual category information in human occipito-temporal cortex? An fMRI study. *Neuron, 35*, 1157–1165.

Tipping, M. E. (2001). Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research, 1*, 211–244.

Vemuri, P., Gunter, J.L., Senjem, M.L., Whitwell, J.L., Kantarci, K., Knopman, D.S., Boeve, B.F., (…), & Jack Jr. C.R. (2008). Alzheimer's disease diagnosis in individual subjects using structural MR images: Validation studies. *NeuroImage*, *39*(3), 1186–1197.

Zou, T. H. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*, 301–320.