NeuroImage 56 (2011) 651–661

Contents lists available at ScienceDirect

NeuroImage

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

Predicting subject-driven actions and sensory experience in a virtual world with Relevance Vector Machine Regression of fMRI data

Giancarlo Valente [*], Federico De Martino, Fabrizio Esposito, Rainer Goebel, Elia Formisano

Maastricht Brain Imaging Center, Department of Cognitive Neuroscience, Maastricht University, The Netherlands

ARTICLE INFO

Article history:
Received 1 December 2009
Revised 20 September 2010
Accepted 24 September 2010
Available online 1 October 2010

ABSTRACT

In this work we illustrate the approach of the Maastricht Brain Imaging Center to the PBAIC 2007 competition, where participants had to predict, based on fMRI measurements of brain activity, subject driven actions and sensory experience in a virtual world. After standard pre-processing (slice scan time correction, motion correction), we generated rating predictions based on linear Relevance Vector Machine (RVM) learning from all brain voxels. Spatial and temporal filtering of the time series was optimized rating by rating. For some of the ratings (e.g. *Instructions, Hits, Faces, Velocity*), linear RVM regression was accurate and very consistent within and between subjects. For other ratings (e.g. *Arousal, Valence*) results were less satisfactory. Our approach ranked overall second.

To investigate the role of different brain regions in ratings prediction we generated *predictive maps*, i.e. maps of the weighted contribution of each voxel to the predicted rating. These maps generally included (but were not limited to) "specialized" regions which are consistent with results from conventional neuroimaging studies and known functional neuroanatomy.

In conclusion, Sparse Bayesian Learning models, such as RVM, appear to be a valuable approach to the multivariate regression of fMRI time series. The implementation of the Automatic Relevance Determination criterion is particularly suitable and provides a good generalization, despite the limited number of samples which is typically available in fMRI. *Predictive maps* allow disclosing multi-voxel patterns of brain activity that predict perceptual and behavioral subjective experience.

Introduction

Machine learning and pattern recognition techniques are being increasingly employed in functional MRI data analysis. These methods outperform conventional univariate statistical models and allow detecting subtle, non-strictly localized effects. In typical fMRI applications, pattern recognition algorithms "learn" a functional relationship between multivoxel brain response patterns (Multi Voxel Patterns, MVP) and a perceptual, cognitive or behavioral state of a subject expressed in terms of a label, which may assume discrete (*classification*) or continuous (*regression*) values. This learned functional relationship is then used to predict the unseen labels from a new dataset ("brain reading").

This approach was first proposed by Haxby et al. (2001), showing that spatial multi-voxel patterns of BOLD responses evoked by a visual stimulus are informative with respect to the perceptual or cognitive state of a subject. Following this study, several other groups investigated, with different approaches, the multivariate relationship between sensory and cognitive stimuli and measured brain activity (Cox and Savoy, 2003; Haynes and Rees, 2005; Kamitani and Tong, 2005; Mitchell et al., 2004; Mourao-Miranda et al., 2005; LaConte et al., 2005; Kriegeskorte et al., 2006; De Martino et al., 2008; Formisano et al., 2008).

In this work we describe the approach of the Maastricht Brain Imaging Center (MBIC) group in the PBAIC 2007 competition, where our group ranked overall second, regressing continuous perceptual, behavioral and emotional ratings on multivariate patterns of activation (see next section and http://www.lrdc.pitt.edu/ebc/2007/competition.html for a description of the competition).

To predict the ratings, we employed a Sparse Bayesian Learning model (Relevance Vector Machine, RVM, Tipping, 2001). This method is particularly appealing in the context of fMRI "brain reading" as the large ratio between problem dimension (number of voxels) and training samples (scans) requires parsimonious models. Bayesian methods are therefore particularly suited, as they usually embed an estimation of model complexity and implement Ockham's razor on estimated parameters. Sparse Bayesian learning algorithms have been recently successfully employed in the context of multi-class classification for fMRI datasets in (Yamashita et al., 2008). The competition has received great visibility and approximately 50 different participants have submitted their predictions of the ratings based on the same brain data. Descriptions of two of these submissions, based on Elastic Net (Carroll et al., 2009) and on low-dimensional embedding (Shen and

[*] Corresponding author. University of Maastricht, Department of Cognitive Neurosciences, P.O. Box 616, 6200 MD Maastricht, The Netherlands. Fax: +31 433884125.
E-mail address: giancarlo.valente@maastrichtuniversity.nl (G. Valente).

1053-8119/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2010.09.062

Meyer, 2008) respectively, have already been published. A submission based on Relevance Vector Machine and Kernel Ridge Regression, from Chu and colleagues is described in this issue (Chu et al., 2011). Similarly to these submissions, we estimated a separate model for each subject and for each rating. While Carroll et al. (2009) used Elastic Net Regression, which is based on a linear combination of L1- and L2-norm penalizations, we used a kernel method, where model sparsity is enforced on the kernel representation rather than on the voxel space (like in L1-norm penalization methods). Furthermore, given the current Bayesian formulation, the parameters are estimated from all training data (using *evidence approximation*) and thus there is no need of cross-validation. Similarly to (Chu et al.) we employed Relevance Vector Machine to predict the final ratings; however, in order to be able to map voxel's relevance for each voxel, we did not employ non-linear kernels, which seemed to provide slightly higher accuracy on same ratings.

In the first part of this work we illustrate the PBAIC 2007 competition framework and dataset, and we describe several general aspects of the multivariate regression of fMRI time series with pattern recognition algorithms with particular emphasis on linear predictive models and Sparse Bayesian learning algorithms. We then describe in detail our analysis of the PBAIC 2007 data, including the analyses performed with BrainVoyager QX (www.brainvoyager.com). Besides theoretical considerations, our choice for this regression method was further motivated by a critical analysis of various regression methods that were employed on the PBAIC 2006 competition. We considered the five best ranking approaches, which included Recurrent Neural Networks (ITC-IRST, Italy, first), Ridge Regression (Princeton University, second and fourth), Dynamic Gaussian Markov Random Field (Stanford University, third), and Relevance Vector Machine (University College London, fifth). This analysis suggested that the application of RVM regression to fMRI 'brain reading' data provides a good compromise between accuracy in predicting the subjective/objective ratings and interpretability of the resulting models in terms of predictive maps, also at single-subject level.

Materials and methods

PBAIC 2007 competition

The PBAIC 2007 competition, organized by Walt Schneider and Greg Siegle of the University of Pittsburgh, was an open competition that involved fMRI data analysis of subject-driven behavior in a virtual world. Such behaviors included navigating, collecting objects, responding to cell phones, taking pictures and avoiding a threatening dog.

The aim of the competition was to evaluate and benchmark, in an objective framework, the capabilities of brain reading techniques to predict subjects' behavior based on fMRI data.

The data were collected on a Siemens Allegra 3T scanner in Pittsburgh University. Thirty-four axial 3.5 mm thick slices were acquired parallel to the AC-PC line using a reverse EPI sequence (TR = 1.75 s, TE = 25 ms, FOV = 210 mm, FA = 76degrees). Structural data were acquired with 1 mm spatial resolution. Three functional runs, each approximately 20 min in length, were collected per subject.

During each task, subjects performed a game in the virtual reality environment. In this game, subjects were paid by and anthropology department to gather information on urban culture, exploring a neighborhood and collecting samples of toy weapons, fruit, and collecting pictures of people's piercings. Subjects had to avoid a threatening dog, whose growl indicated when he was approaching. All the actions performed correctly during the game were rewarded with real money, unless the subject was bitten by the dog.

Based on Virtual Reality logfiles, eye movement as recorded by an eye-tracking system and subsequent subjective evaluation from the subject, a set of ratings of subjects' activity was associated with each functional run. These ratings were related to different components of the subjects' task such as actions (*Velocity*, *Hits*, *Search People*, *Search Weapons*, *and Search Fruit*), vision (*Body*, *Faces*, *Weapons and Tools*, *Interior and Exterior*, *Gender*, *Fruits and Vegetables*), audition (*Instructions*, *Dog*) and emotional valence (*Arousal*, *Valence*). Each rating was convolved with the double gamma model for hemodynamic response function (Friston et al., 1998) and downsampled to match the functional scans temporal resolution. For more details on the competition framework, please refer to http://www.lrdc.pitt.edu/ebc/2007/competition.html.

The dataset provided to the participants consisted of measurements from three subjects, with three functional runs each. For each subject, two runs were made available to participants, together with the associated ratings (training dataset), while the third run (test dataset) was provided without ratings. Participants were to learn the association between ratings and functional data from the training data (e.g. using multivariate regression) and submit the prediction for each rating on the test dataset to the organizers for scoring and performance assessment.

Multivariate regression in fMRI: general principles

Consider an fMRI data set **X** and some associated labels **t** (e.g. the subjects' feature ratings provided together with the PBAIC fMRI dataset).

Pattern analysis algorithms "learn" a functional relationship between data and labels in a *training dataset* and use this relationship to predict the unseen labels of a new dataset (*test dataset*).

A suitable performance metric, which can be expressed as an error (or loss) function ε (for instance, the sum of squares error, or some other suitable measures), is generally considered while evaluating and comparing different models. In very general words, pattern recognition algorithms aim at learning a model on the training dataset **D** that gives the minimum error on an unseen dataset test **D'**. Several models and several performance metrics can be introduced to perform this task.

Fig. 1 summarizes in a block diagram the steps that are generally performed. Pre-processing is usually employed in order to reduce the effects of noise. The first general steps are slice scan time correction, motion correction and removal of linear trends from the time series. In the context of multivariate pattern analysis, particular attention must be paid to the coregistration of different functional runs.

Spatial and temporal filtering, can be adjusted repeatedly contextually to the model training. To avoid removing useful information from the data by means of spatio-temporal filter, it is safer to evaluate different amounts of filtering on the training dataset and choose the

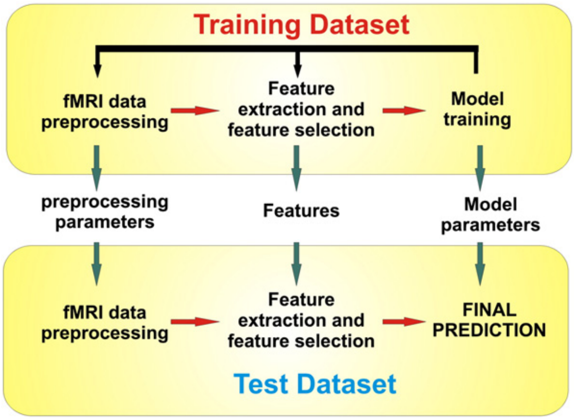

Fig. 1. A schematic description of the steps performed during the learning and the prediction phase in fMRI brain reading. Red arrows indicate temporally distinct phases of the learning procedure, while green arrows indicate parameters settings, voxel selection and model weights which are transferred from the training dataset to the test dataset. Black arrows indicate feedbacks during the learning phase.

settings that provide the highest performance metric (Formisano et al., 2008). Several filtering strategies are usually available, ranging from simple box-car smoothing functions to multi-resolution approaches.

Subsequently, features are extracted from the data, on which the model will be trained. The aim of this step is to transform the fMRI dataset into a different representation that may be more suitable for the regression scheme implemented. As pointed out in Duda et al. (2001), the boundary between feature extraction and pattern classification is not clear-cut, as there may be a classification scheme so powerful that it does not need any feature extraction, while there may be an ideal feature extraction scheme that makes the work of the classifier trivial.

In the context of regression for fMRI brain reading, different features can be considered. When using Kernel methods (that are better suited in handling fMRI data of large dimension), all voxels' time courses can be considered as features.

In many cases some dimension reduction scheme may be employed. In fact, the reduction of the number of features (dimensions) generally helps data understanding, reduces memory storage requirements and training times and helps mitigating the effects of the *curse of dimensionality* improving the overall performances (Guyon and Elisseeff, 2003).

Univariate ranking methods are generally employed due to their reduced computational demands. The number of considered features may be decided in advance, or it may be evaluated and assessed in cross-validation. In order to have better generalization performances, several pre-processing and feature extraction/selection possibilities may be evaluated during the training phase, hence the feedback arrows on top of Fig. 1.

Training and test dataset should be strictly separated, and once the pre-processing, feature extraction and selection parameters and model weights have been estimated *on the training dataset*, they are used for the final prediction on a new dataset.

The main part of any pattern analysis approach is the use of a suitable model and its estimation on the available data. After the preprocessing and the feature extraction/selection steps, the training dataset **D** consists of a collection of N pairs ($\mathbf{x_i}$, t_i), denoting with $\mathbf{x_i}$ a sample vector of dimension D and with t_i the corresponding one-dimensional label. The choice of the model is a crucial part of the learning process. A model too simple may fail to grasp the variability of the data, while a model too complex may also fit the noise component in the data (*overfitting*). Good generalization performance on a new dataset is based on finding a proper trade-off (given the training dataset **D**).

The choice of the best learning model is not trivial and–if no prior assumptions about the nature of the regression task are available–it is not possible to prefer a model to another, as stated by the *No Free Lunch Theorem* (Duda et al., 2001). This theorem indicates that there is no algorithm that can outperform any other (even random guessing) on *any* problem. Once some aspects of the problem, like data distribution, amount of training data and cost functions, are available, it is possible to compare different algorithms. In the context of the Brain Reading competition, in order to choose the model and the learning algorithm, we conducted an analysis on the best five ranking approaches on the previous year's dataset. This analysis suggested the use of Relevance Vector Machine due to the high accuracies at a single subject level and to the possibility of mapping voxels' contribution to the prediction.

Model parameters can be set using cross-validation approach. The training dataset is divided into m disjoint parts and a model is trained m times, each time leaving a part as a validation set. The mean of the errors of the models on the different datasets is then considered while comparing parameters (*m-fold cross-validation*) (Duda et al., 2001).

Bayesian methods generally do not require cross-validation to estimate model parameters (although sometimes cross-validation is used in Bayesian frameworks for model selection (Rasmussen and Williams, 2006)). In Bayesian analysis, model comparison involves the use of marginal probabilities of the choice of a suitable model (Duda et al., 2001; Bishop, 2006).

Once the training phase is completed, a set of pre-processing parameters, of features and of model parameters is available to perform the prediction on a new dataset. This dataset is preprocessed in the same way, and the prediction is performed using the trained model.

Linear models for multivariate regression

Model estimation

Linear models are a natural choice for fMRI "brain reading." There are both computational and interpretational reasons for this. In fact, being fMRI in an extremely high dimensional space, the use of non-linearities does not provide significant improvement over linear models, while making model estimation more demanding. Another reason for the widespread use of linear models lies in the straightforward interpretation of the results. In fact, together with generalization performances, it is possible to assess the relative "importance" of single features (i.e. voxels); in functional neuroimaging this aspect is as important as the generalization performance, as it helps improve the understanding of the brain and the functional role of different locations. With linear models this "brain mapping" step is straightforward and considerably easier, if compared with non-linear models, leading to a clearer interpretation of the ongoing neural processes. For these reasons we considered only linear models for the analyses of the PBAIC competition.

A standard linear model has the following form:

$$t = y(\mathbf{x}, \mathbf{w}) + \varepsilon \tag{1}$$

where y(**x**, **w**) is the deterministic input-output mapping part and ε accounts for the noise in the measurements. The deterministic mapping can be modeled as (Bishop, 2006):

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_D x_D = \mathbf{w}^{\mathbf{T}}\,\tilde{\mathbf{x}} \tag{2}$$

where $\mathbf{x} = (x_1,\ldots,x_D)^T$ denotes the training dataset (defined in a D-dimensional space), $\tilde{\mathrm{x}} = (1,\mathbf{x}^{\mathbf{T}})^{\mathbf{T}}$ and the $D+1$-dimensional weight vector **w** indicates the weights of the linear model (with w_0 being the bias term). The deterministic mapping in Eq. (2) can be regarded as a linear combination of the fMRI time-series of different voxels, each with a different weight. A feature space mapping is usually employed by many learning algorithms:

$$\phi(\mathbf{x}) = (\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \ldots, \phi_{M-1}(\mathbf{x}))^T \tag{3}$$

where $\phi : \Re^D \rightarrow \Re^M$, maps the original D-dimensional *feature space* **x** into an M-dimensional one. ϕ can be for instance a linear polynomial, or radial basis function. After this transformation, Eq. (2) can be written as:

$$y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^{\mathrm{T}}\phi(\,\tilde{\mathbf{x}}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) \tag{4}$$

where this time **w** is an M-dimensional vector of parameters.

The training of the model aims at finding a set of optimal weights **w**, such that the error on the unseen dataset is minimized. A solution to this problem is to look for the model weights that minimize an error measure on the *training dataset.*

One common error function is the sum of squares:

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} \left\{ t_n - \mathbf{w}^T \phi(\mathbf{x}_n) \right\}^2 \tag{5}$$

The minimization of this function leads to the estimate of the model parameters:

$$\tilde{\mathbf{w}} = \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\mathbf{t} \tag{6}$$

with $\boldsymbol{\Phi} = (\phi(\mathbf{x}_1),\ldots,\phi(\mathbf{x}_N))^T$. It can be shown that the least-squares solution corresponds to the projection of the target t onto the subspace generated by the columns of $\boldsymbol{\Phi}$ (Bishop, 2006).

A perfect fit on the training dataset, however, may not be optimal for generalization purposes. In fact, with a small sample size (compared to the dimensions of the feature space) there is a high risk of fitting also the noise term as part of the model. A way to reduce the effects of overfitting is to have a smooth estimate by means of an additional regularization term.

The regularized error function can be expressed as a weighted sum of two terms:

$$E(\mathbf{w}) = E_{\mathrm{D}}(\mathbf{w}) + \lambda E_{\mathrm{W}}(\mathbf{w}) \tag{7}$$

where E_{D} is the same as in Eq. (5) and E_{W} is the regularization term. A simple form of regularizing term (L2-norm regularization) is the following (Bishop, 2006):

$$E_{\mathrm{W}}(\mathbf{w}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \tag{8}$$

that leads to the solution:

$$\tilde{\mathbf{w}} = \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\mathbf{I}\right)^{-1}\boldsymbol{\Phi}^T\mathbf{t} \tag{9}$$

that is also called *ridge regression* solution. Regularization is particularly effective while training on small datasets (as it reduces the model complexity and accordingly the risk of overfitting). The weighting coefficient λ is usually set in cross-validation over the training dataset.

Kernel methods

All the models presented so far are based on linear combination of the features (or of their mapped transform ϕ). In some situations, however, it is more convenient to use a different representation for data, using a *kernel function*.

Considering the mapping in Eq. (3), a kernel function is defined as:

$$k(\mathbf{x},\mathbf{x}') = \phi(\mathbf{x})^{\mathrm{T}}\phi(\mathbf{x}') \tag{10}$$

Considering the identity mapping (i.e. $\phi(\mathbf{x}) = \mathbf{x}$), Eq. (10) becomes:

$$k(\mathbf{x},\mathbf{x}') = \mathbf{x}^{\mathrm{T}}\mathbf{x}' \tag{11}$$

The models seen in the previous section can be reformulated in a dual form in the kernel representation (Bishop, 2006). A linear model in the kernel space can be described as:

$$y(\mathbf{x},\mathbf{w}) = \sum_{j=1}^{N} w_j k\left(\mathbf{x},\mathbf{x}_j\right) + b = K(\mathbf{x},\mathbf{x})\mathbf{w} + b \tag{12}$$

where the last part is a vector reformulation of the equation.

The advantage of using the dual representation is quite evident in the fMRI data analysis. In fact, in this case **w** is an *N*-dimensional vector, while in Eq. (2) it was a *D*-dimensional one. Recalling that *N* represents the number of available samples (typically few hundreds) and *D* the number of features (voxels, typically tens of thousands), the kernel representation is far more compact and it allows handling large datasets. Support Vector Machines (SVM, Vapnik, 1995), Relevance Vector Machine (RVM, Tipping, 2001) and Gaussian Processes (GP, Rasmussen and Willams, 2006) are commonly employed kernel-based approaches.

Once the model in Eq. (12) has been estimated on the training dataset D, the prediction is performed by considering a new kernel that accounts for the prediction dataset D':

$$t' = K(\mathbf{x}',\mathbf{x})\mathbf{w} + b \tag{13}$$

Therefore, to perform a prediction, all the samples of the training set are generally used. Sparse kernel machines (like SVM or RVM), instead, make use of a subset of the training set, selecting the most "important" samples according to different criteria.

Predictive mapping

An important aspect of Machine Learning techniques applied to fMRI data analysis is the possibility to assess the relative importance of the features (i.e. voxels) in generating the final prediction. Linear models allow extracting this information in a straightforward way.

Consider the linear model presented in Eq. (2). Once the model weights **w** have been estimated, it is possible to map the relative importance of feature *j* by simply considering the absolute amplitude of weight w_j. The sign of w_j denotes then a positive or negative contribution to the prediction.

Similar considerations can be done for the dual representation, in the linear case (Eq. (11)). Eqs. (12) and (13) can be rewritten as:

$$y(\mathbf{x},\mathbf{w}) = \mathbf{x}^{\mathrm{T}}\mathbf{x}\mathbf{w} + b = \mathbf{x}^{\mathrm{T}}\mathbf{m} + b \tag{14}$$

and

$$t' = \mathbf{x}'^{\mathrm{T}}\mathbf{x}\mathbf{w} + b = \mathbf{x}'^{\mathrm{T}}\mathbf{m} + b \tag{15}$$

where $\mathbf{m} = \mathbf{x}\mathbf{w}$ is a D-dimensional vector that represents the weights in the feature space (while **w** in this case refers to the sample space).

Relevance Vector Machine

Relevance Vector Machine is a probabilistic kernel method. Least square and L2-norm regularized solutions to multivariate regression problems can be also framed in the context of probabilistic models.

Consider Eq. (1), and assume that the noise term ε follows an independent, identically distributed Gaussian distribution with zero mean and precision (inverse of the variance) equal to β, that is $p(\varepsilon) = N(0,\beta^{-1})$. It follows that:

$$p(t|\mathbf{x},\mathbf{w}) = \mathcal{N}\left(y(\mathbf{x},\mathbf{w}),\beta^{-1}\right) \tag{16}$$

and, considering the i.i.d. assumption, it holds that

$$p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{w}^T\phi(\mathbf{x}_n),\beta^{-1}\right) \tag{17}$$

where **t** denotes the vector of all the *N* targets.

It can be shown (Bishop (2006)) that the ML estimation, in the case of i.i.d. Gaussian noise term, is *equivalent* to the least-squares solution (7) (and, as discussed previously, more prone to overfit the data). This aspect is an inconvenient of ML methods. Bayesian methods, on the other hand, are designed in such a way to reduce or avoid overfitting, by considering the posterior distribution of the parameters. Using Bayes' rule it holds:

$$p(\mathbf{w}|\mathbf{t},\mathbf{x},\beta) = \frac{p(\mathbf{t}|\mathbf{w},\mathbf{x},\beta)p(\mathbf{w})}{p(\mathbf{t}|\mathbf{x},\beta)} \tag{18}$$

where the prior term p(w) contains all the information one has on the model parameters. A full Bayesian perspective on the problem does not require the use of point estimate on the model. In fact the prediction is not performed by considering the most probable model parameter ($\mathbf{w}_{\mathrm{MAP}}$) and using that point estimate on the new data, but rather averaging across *all* the possible models with their probabilities. In other words, considering a new data point $\mathbf{x}^*$ then the predicted value t^* will be distributed according to the following:

$$p(t^*|\mathbf{x}^*,\mathbf{x},\mathbf{t}) = \int p(t^*|\mathbf{x}^*,\mathbf{w})p(\mathbf{w}|\mathbf{x},\mathbf{t})d\mathbf{w} \tag{19}$$

that, in the case of Gaussian prior, is still a Gaussian (Rasmussen and Williams, 2006; Bishop, 2006). For a review of Bayesian methods refer to (Duda et al., 2001; Bishop, 2006).

Compared to SVM, the RVM typically leads to much sparser models and does not require that the kernel is positive definite. In the RVM formulation, $N+1$ parameters have to be estimated. Under the assumption of Gaussian noise, Eq. (17) is used also for this algorithm.

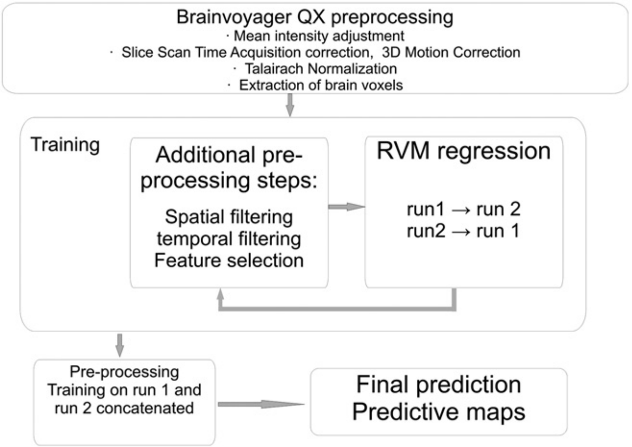
Brainvoyager QX preprocessing
· Mean intensity adjustment
· Slice Scan Time Acquisition correction, 3D Motion Correction
· Talairach Normalization
· Extraction of brain voxels
Training
Additional pre-processing steps:
Spatial filtering
temporal filtering
Feature selection
RVM regression
run1 → run 2
run2 → run 1
Pre-processing
Training on run 1 and run 2 concatenated
Final prediction
Predictive maps

Fig. 2. A schematic description of the approach used by Maastricht University for PBAIC 2007 competition. Filter settings were optimized for each rating and each subject separately.

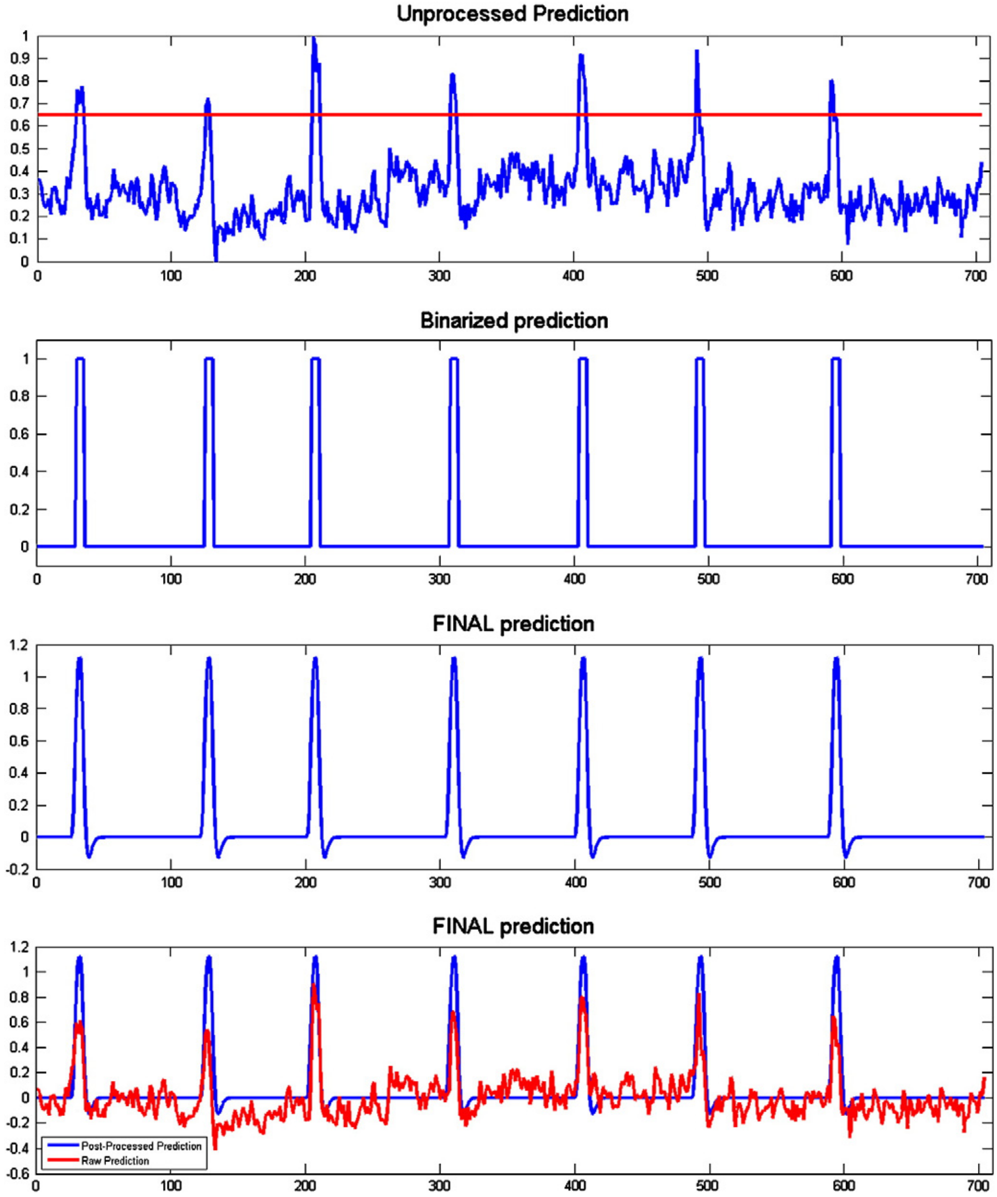

Fig. 3. Scheme of the post-processing strategy for the rating *Instructions*. Using this strategy it is possible to achieve a 0.99 correlation between the real and the predicted rating (the correlation between the “raw” prediction and the rating is approximately 0.8). This post-processing step is particularly effective due to Fisher *z*-transform of the correlation.

The key point of RVM is the prior on the model weights **w**. Each of them is assumed to have a Gaussian distribution with zero mean and precision (inverse of the variance) α_i:

$$p(\mathbf{w}\,|\,\boldsymbol{\alpha}) = \prod_{i=1}^{N+1} \mathcal{N}\left(0, \alpha_i^{-1}\right) \tag{20}$$

The parameters α_i are called *hyperparameters* and it can be shown that posterior distribution of the weights is again Gaussian (Tipping, 2001), with mean and covariance given by:

$$\mathbf{m} = \beta\Sigma\Phi^T\mathbf{t} \tag{21}$$

$$\Sigma = \left(\mathbf{A} + \beta\Phi^T\Phi\right)^{-1} \tag{22}$$

denoting with $\mathbf{\Phi}$ the design matrix as in Eq. (6) and $\mathbf{A}=\mathrm{diag}(\alpha_i)$. A common choice is to use a *log-uniform* hyperprior over α, that in combination with Eq. (20) implements the *Automatic Relevance Determination* (ARD) (MacKay, 1994; Neal, 1996). In fact, during the estimation of the model, many hyperparameters α_i will grow to infinity, so that the corresponding model weight w_i will have a posterior distribution concentrated around zero. In other words, only the model weights (and therefore the basis functions associated with these parameters) that are "relevant" given the training data will remain, pruning out the unnecessary ones and leading to a sparse model. *Relevance vectors* can be interpreted similarly to *support vectors* in the SVM formulation, where the support vectors represent the most "difficult" cases, while the relevance vectors represent the most "archetypical" ones.

The values of α and β are determined using type-II Maximum Likelihood (known also as *evidence approximation*) (Bishop, 2006). Once these parameters have been estimated, the prediction over a new data point $\mathbf{x}^*$ can be estimated using Eq. (19), having a predictive distribution that is still Gaussian, with mean and variance given by (Tipping, 2001):

$$\mathbf{m}(\mathbf{x}^*) = \mathbf{m}^T\phi(\mathbf{x}^*) \tag{23}$$

$$\sigma^2(\mathbf{x}^*) = (\beta)^{-1} + \phi(\mathbf{x}^*)^T\Sigma\phi(\mathbf{x}^*) \tag{24}$$

Compared to SVM, relevance vector machine provides in many applications a much sparser model, typically an order of magnitude more compact, with little or no reduction of generalization error (Tipping, 2001). Furthermore, no parameter has to be estimated in cross-validation (like C and ε in SVM). RVM suffers from having high confidence in making predictions in regions far from the training data. Other Bayesian regression techniques, like Gaussian Processes (Rasmussen and Williams, 2006), do not suffer from this drawback, at the expenses of an increased computational time.

PBAIC 2007 competition: Maastricht group approach

The approach employed by MBIC group is summarized in Fig. 2. We performed the analyses for *each subject separately*. This choice was motivated by computational complexity (training simultaneously on multiple subjects increases considerably memory requirements in manipulating the kernel (Eq. (11)). Also, training simultaneously on multiple subjects without properly accounting for inter-subject variability in the spatial patterns may reduce the performance of trained models.

Furthermore, we decided to predict each rating separately from the others. This was due to the need of achieving a high score and investigating the neural activation patterns that reliably predict a specific rating.

We pre-processed the raw data provided by the PBAIC organization using BrainVoyager QX (Brain Innovation, Maastricht, The Netherlands). In the training phase different pre-processing and

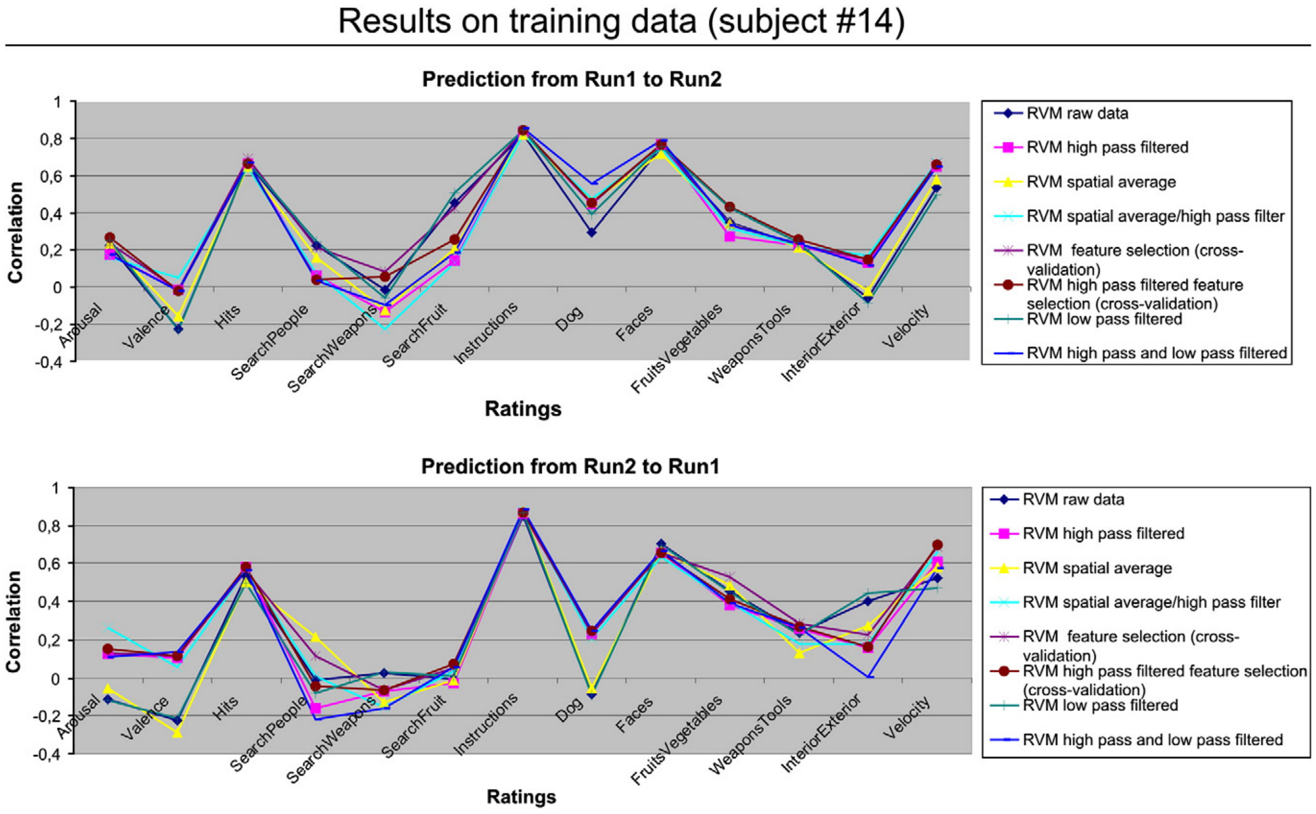

Fig. 4. Results of the training on the first two runs for subject #14. Correlation values for different pre-processing strategies and feature selection are reported.

feature selection schemes were explored, and the final predictions and predictive maps were obtained from the two concatenated runs.

Preprocessing

We performed a standard fMRI pre-processing, including slice scan time correction (with sinc interpolation) and 3-D motion correction (with trilinear/sinc interpolation). To minimize the effects of subjects' head movements between two runs, we realigned both runs 2 and 3 to the first functional run. Due to the presence of some spiking artifacts in subject #1 and subject #13, we performed mean intensity adjustment on those functional runs. To mitigate the effects of drifts in the measurements, we removed linear trends in all time series. As the evaluation metric of the competition was based on Pearson correlation, no information on the mean of the prediction is considered, and therefore, we removed the temporal mean of each run.

Functional volumes were interpolated to 3 mm × 3 mm × 3 mm resolution, co-registered to the anatomical images, and both anatomical and functional data were normalized to Talairach space (Talairach and Tournoux, 1998). We subsequently removed the voxels of spatial locations outside the brain by means of a volume mask, which lead to approximately 45000 voxels per subject.

Training

We performed assessment of parameters (filtering and feature selection) and training of a linear RVM model (see above) on the first two runs. We employed a 2-fold (*split half*) cross-validation strategy: training was performed on run 1 (*training set*) and accuracy was

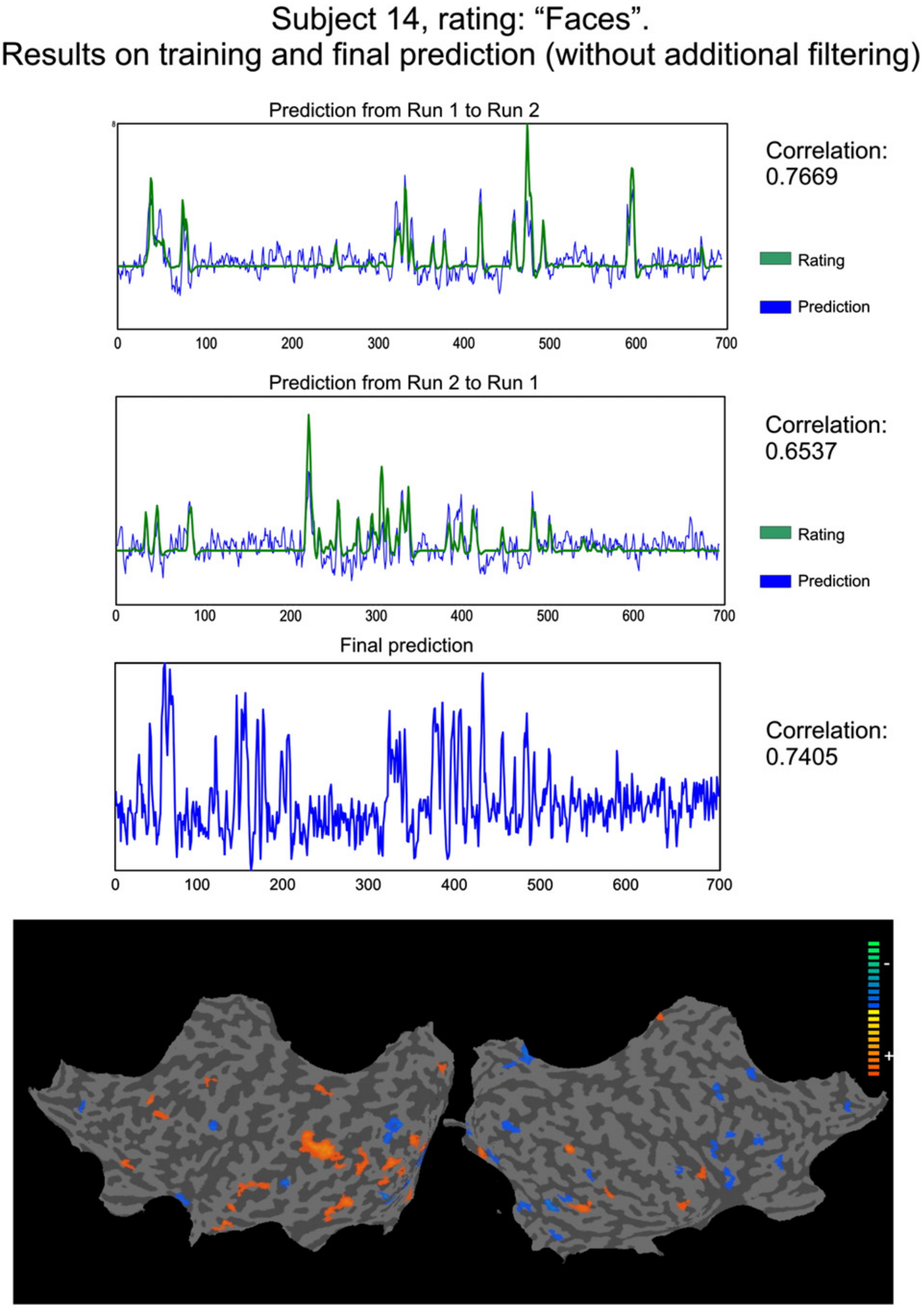

Fig. 5. Example of prediction and mapping for the rating *Faces* of subject 14. First two top panels: comparison between predicted and real rating on the two training runs. Third panel: predicted rating for run 3. Bottom panel: individual predictive map obtained training on the first two runs.

assessed on run 2 (*validation set*) and vice versa. We did not divide further the available datasets into a higher number of folds to have a realistic estimate of performance degradation due to inter-run variability of BOLD signal.

Within the training, we considered additional, rating-specific spatial and temporal filtering of fMRI datasets. Temporal filtering aims at reducing the spectrum of the time series to the one of the ratings. However, due to non-stationarity of the signal, both low and high pass filter may remove informative content. We therefore performed different filtering approaches and evaluated the average prediction performance of the pre-processed dataset. Spatial smoothing was also considered, to mitigate the effects of spatial isotropic noise. The three filtering steps can be summarized as follows:

Temporal low-pass filtering: each time series was filtered using a box-car having a width of 3 samples. Temporal high-pass filtering: this step removed up to 3 cycles per time-series. Spatial low-pass filtering: each voxel time course was averaged with the ones of the neighbor voxels within a cube comprising 27 voxels.

We also explored the use of more intense filtering, but it did not give significant and consistent benefit on the training, therefore we did not consider it for the final prediction.

The feature selection strategy we employed was based on univariate ranking of the features. We ranked all the voxels on the training dataset by means of linear correlation (absolute value). Subsequently we considered the *n* voxels ranked first, with n ranging from a hundred of voxels to the whole set, and tested the model on the validation dataset.

The RVM training was performed by first constructing the kernel (as described in Eq. (11)), and then estimating the model (see Eqs. (21) and (22)). Once the model had been estimated, the prediction was carried out according to Eqs. (13), (15). The average accuracy for the two predictions was then considered.

Once the "optimal" pre-processing had been estimated on the training data, the final prediction was carried out by concatenating the two training runs and each rating was predicted separately.

Post-processing of ratings

Due to the competition evaluation metric, the post-processing of selected ratings, for which the prediction was expected to be highly accurate, might have affected relevantly the final score (in fact, the Fisher z-transform of the scores magnifies the weighting of accuracies close to one). Based on results from the training, we performed - for the rating *Instructions* - additional post-processing (see Fig. 3) Each predicted rating was linearly scaled between 0 and 1. We subsequently used a threshold and binarized the prediction according to this threshold. This binary predictor was then convolved with the standard estimate of the HRF (Friston et al., 1998), and shifted back of 4-5 TRs. Threshold value and shift were estimated on the training dataset. Moreover, since this rating was the same for the three subjects, an averaging of the prediction among the three subjects prior to the post-processing appeared as the most robust choice.

For the other ratings, we did not employ other post-processing strategies and did not exploit any dependency among predictors, although it was possible (as in the "*search*" ratings).

Results and predictions

On training data, our approach with RVM regression predicted accurately and robustly (with respect to different pre-processing options) many of the target ratings. Fig. 4 shows for subject #14 the correlation coefficients between predictions and provided ratings, obtained for different types of filtering and feature selection, after training on run 1 and testing on run 2 (top panel) and vice versa (low panel). It is noticeable that for several of the ratings the correlation values were above 0.6 for most of the tested approaches. These results were consistent with those obtained in the other two subjects (#1 and #13), even though the overall correlation values for these latter were lower. This was likely due to the presence of larger movements and of spiking artifacts in some runs.

Results of the predictions on run 3 confirmed the expectations from the training and the first two submission scores reflected expected scores from the training, thus suggesting that our approach did not overfit the training data. Individual ratings' correlations for subject #14 are reported in Fig. 4. Our final submission ranked second in the PBAIC 2007 competition.

Overall the best-predicted rating was *Instructions*, whose accuracy was 0.99—after the described post-processing. However, prediction accuracy on the training data, without any post-processing, was around

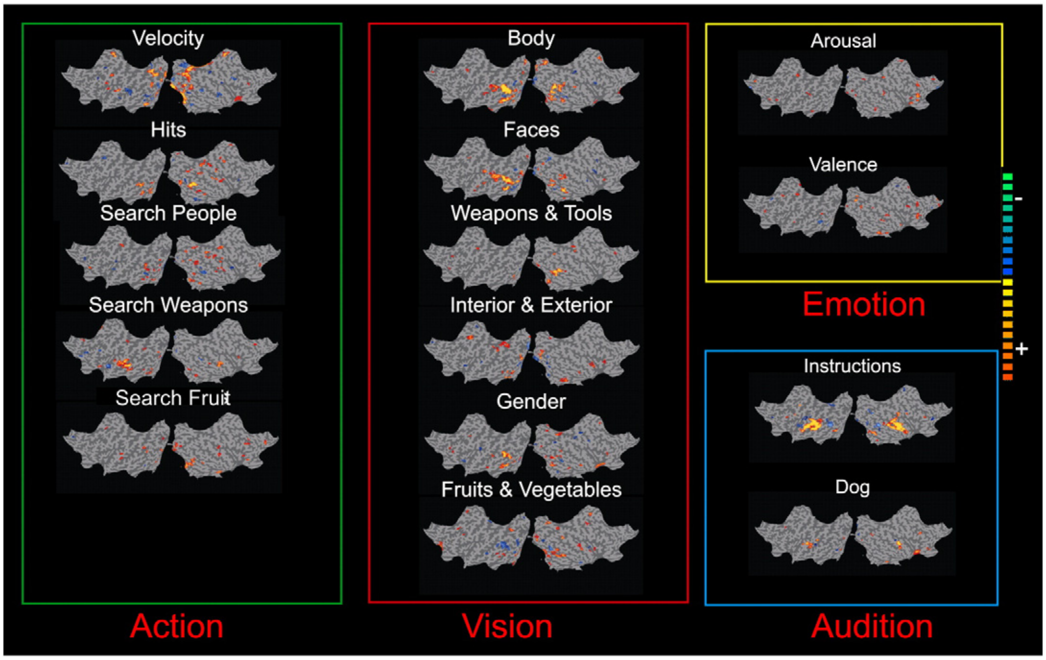

Fig. 6. Group maps related to the different ratings of PBAIC 2007 competition. Ratings were grouped in four categories according to their content.

0.8 (see Fig. 4). Other ratings that were predicted with high accuracy were "visual" ratings, such as *Faces*, *Body* and *Velocity*. The scores were significantly lower for the prediction of the two subjective ratings (*Arousal* and *Valence*). The accuracy for these scores, however, was considerably higher on subject #14 then in the other subjects. The scores for the three *Search* ratings were also considerably lower than the average (even with some negative values). As the *Search* ratings were identical across subjects, we used–for the final submission–the best among the three subjects' predictions.

The analysis on feature selection showed that limiting the training to a subset of all voxels may have led to improvement in the total score (Fig. 4). However, results on the third dataset submission did not confirm this suggestion. Indeed, using this approach we obtained a final score = 0.495 (no filtering, second submission), which is lower than the score obtained without feature selection (no filtering, first submission, score = 0.512). This might be due to several reasons. The principal problem of the implemented feature selection scheme is that the selection process is univariate. This may be too simple for the problem considered, leading to an overconfidence on single voxel relevance. Multivariate feature selection schemes (De Martino et al., 2008; Yamashita et al., 2008) may offer a valid alternative to this problem. Computational requirements, however, limit the use of these more complex feature selection schemes with datasets of such large dimensions.

The analysis on spatio-temporal filtering (Fig. 4) indicated that the pre-processing strategy affected the results in a way that was dependent of rating (and subject), and therefore the final predictions of run 3 ratings were based on subject- and rating-specific settings. The filtering approaches used for the final submission are reported in the Supplementary Material.

Although the filtering had some influence on the final score, spatially and temporally filtering of the data did not appear to have a systematic effect on the quality of the predictions. This may be related to using all voxels' time courses for the prediction. As many time series are considered together, the averaging effect makes low-pass filtering less relevant.

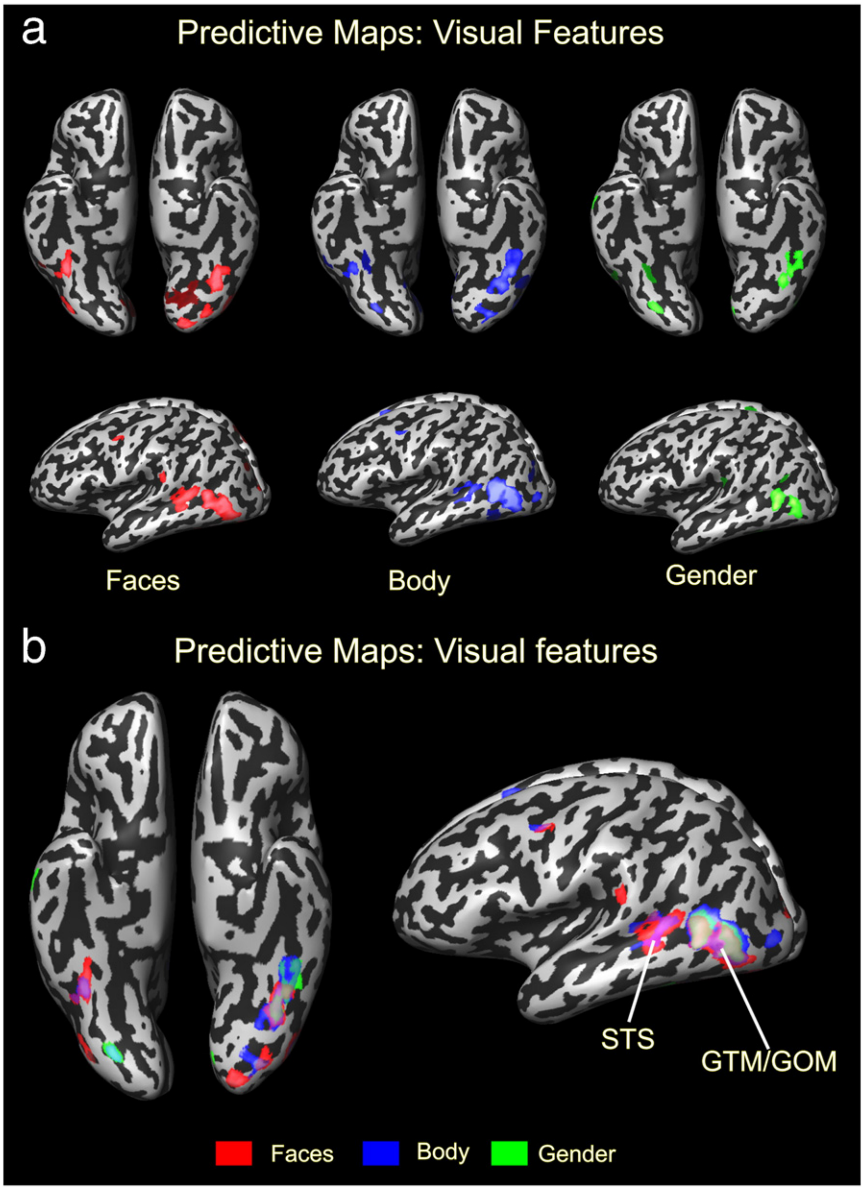
a
Predictive Maps: Visual Features
Faces
Body
Gender
b
Predictive Maps: Visual features
STS
GTM/GOM
🟥 Faces
🟦 Body
🟩 Gender

Fig. 7. (a) Predictive maps for visual features: *Faces*, *Body* and *Gender*. The positive values of the maps are in light color, while the negative in dark color. (b) Predictive maps related to visual features: *Faces*, *Body* and *Gender*. The predictive map of each rating is considered in its *absolute value* and superimposed to those of the others.

Predictive mapping

To investigate the role of different brain regions in ratings' prediction we generated maps indicating the weighted contribution of each voxel. For each subject and each rating, we computed the corresponding *predictive* map using Eqs. (14) and (15) and normalized it to the standard deviation. As functional and anatomical data were co-registered, predictive maps could be displayed on the folded, inflated or flattened representation of cortical surface, obtained from the anatomical data set (Goebel et al., 2006). An example is illustrated in Fig. 5, where we considered the rating "Faces" for subject #14. In the top panel the intermediate and final prediction are plotted, while in the bottom panel we illustrate the spatial map associated with this prediction.

We also considered 'group' predictive maps obtained by averaging (in Talairach space) the three single-subject maps after a spatial smoothing with a Gaussian kernel (FWHM = 4 mm). Fig. 6 shows these "group" maps projected on the reconstructed and flattened cortical surface of an individual brain. Maps are grouped in four categories: Action (*Velocity*, *Hits* and the three *Search* ratings), Vision (*Body*, *Faces*, *Weapons* & *Tools*, *Interior* & *Exterior*, *Gender* and *Fruits* & *Vegetables*), Audition (*Dog* and *Instructions*) and Emotion (*Arousal* and *Valence*).

For visual ratings, regions of the occipito-temporal cortex contributed most consistently to the predictions (red box in Fig. 6, see also Fig. 7). As expected, obtained maps included (but were not limited to) 'specialized' regions which conventional neuroimaging studies report to be maximally active during the perception of certain features (e.g. fusiform face area for faces (Kanwisher et al., 1997) or lateral occipital complex for objects (Malach et al., 1995). The three maps of *Face*s, *Body* and *Gender* show considerable overlap in the occipito-temporal cortex (OTC), in the superior temporal sulcus and gyrus (STS/STG) and the middle occipital gyrus (MOG, see Fig. 7b). While OTC and MOG seem to be involved in predicting Faces, Body and the Gender, STS/STG does not seem to be used to infer the gender while watching a person.

Interestingly, for each of these features, predictive maps also included additional regions in the frontal or temporal cortex which seem to contribute in generalizing the predictions to new data. Future analyses on a larger number of subjects are needed to assess the consistency of these additional regions.

Maps corresponding to auditory ratings are illustrated in Fig. 8. In particular, predictive maps for *Instructions* included a large expanse of bilateral superior temporal cortex (STG/STS), extending toward the temporal pole. Maps for *Dog* included a subset of the auditory regions also present for *Instructions*. Maps for *Instructions* also included a negative contribution from regions in MOG, which overlap with regions that positively predict *Faces*. It may be speculated, that this overlap reflects the inhibition of the attention to relevant visual targets (faces) while listening to the auditory instructions.

Conclusions

In this work we have described the approach employed by the Maastricht Brain Imaging Center in the PBAIC 2007 brain reading competition. The use of Relevance Vector Machine models trained on the whole brain proved reliable in predicting the final ratings. Spatio-temporal filtering was optimized for each rating and subject, since the results obtained on the training data indicated that the filtering strategies had no systematic effect on all the predictions.

Results on the training dataset indicated that some ratings (e.g. *Instruction*, *Faces*, *Hits*, *Velocity*) could be accurately and consistently predicted after training on one run and predicting on the other. On the other hand, for other ratings (*Arousal*, *Valence*, *Search*) the prediction accuracy was considerably lower. The results of the final prediction were in agreement with the results obtained on the training data, suggesting that the learned model, based on the use of Automatic Relevance Determination, did not overfit the training dataset.

Using a linear model, we could address the problem of assessing the relevance of individual voxels in the final prediction. Therefore a predictive map was associated with each prediction. The areas involved in the prediction included but were not limited to the 'specialized' regions highlighted by conventional neuroimaging studies. These results suggest that predictive maps are a valuable tool to disclose multivoxel patterns of brain activity that predict perceptual and behavioral experience and to

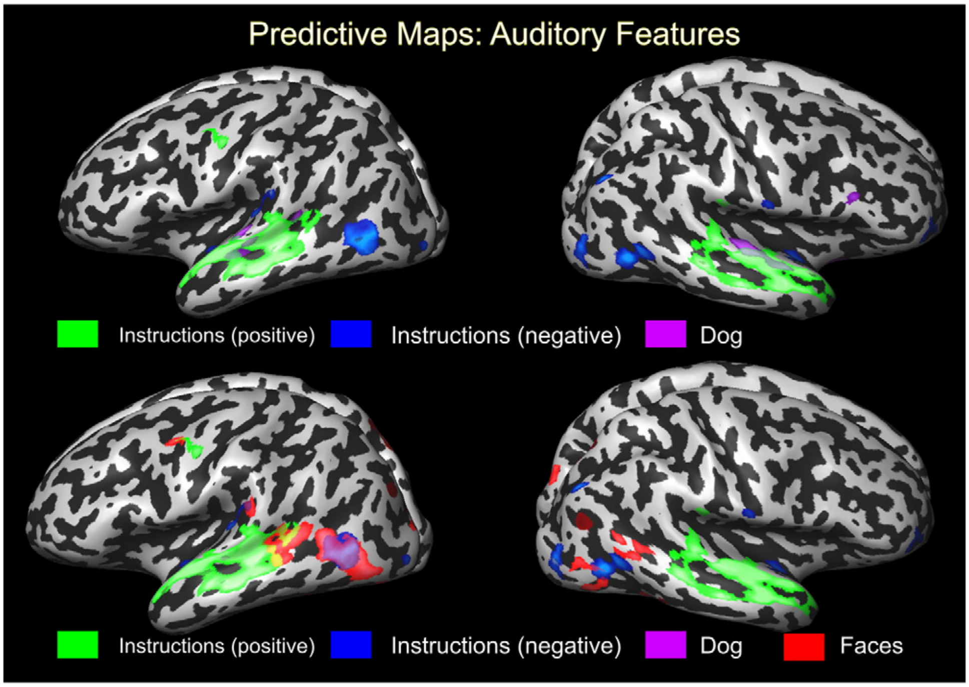

Fig. 8. Top panel: predictive maps for auditory features: *Instructions, Dog* (top panel; bottom panel: auditory features compared with *Faces*. Regions in MOG negatively predict auditory cues (*Instructions*) while being positively predictive for *faces*.

reveal candidate regions hitherto not implicated in the processing of specific categories.

Supplementary materials related to this article can be found online at doi:10.1016/j.neuroimage.2010.09.062.

Acknowledgments

Financial support from NWO (MaGW-VIDI grant 452-04-330) to EF and from the Experience Based Cognition project (University of Pittsburgh) to EF and GV is gratefully acknowledged. The dataset used in this article is provided by the Pittsburgh Brain Activity Interpretation Competition (PBAIC) 2007, organized by Walter Schneider (University of Pittsburgh).

References

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer.

Cox, D., Savoy, R., 2003. Functional magnetic resonance (fMRI) “brain reading”: detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage 19 (2), 261–270.

De Martino, F., Valente, G., 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 (1), 44–58.

Carroll, M.K., Cecchi, G.A., Rish, I., Garg, R., Ravishankar Rao, A., 2009. Prediction and interpretation of distributed neural activity with sparse models. Neuroimage 44 (1), 112–122.

Chu, C., Ni, Y., Tan, G., Saunders, C.J., Ashburner, J., 2011. Kernel regression for fMRI pattern prediction. Neuroimage 56 (2), 662–673.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification, 2nd ed. John Wiley & Sons.

Formisano, E., De Martino, F., Valente, G., 2008. Multivariate analysis of fMRI time series: classification and regression of brain responses using Machine Learning. Magn. Reson. Imaging 26 (7), 921–934.

Friston, K.J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M.D., Turner, R., 1998. Event-Related fMRI: Characterizing Differential Responses. Neuroimage 7 (1), 30–40.

Goebel, R., Esposito, F., Formisano, E., 2006. Analysis of functional image analysis contest (FIAC) data with BrainVoyager QX: From single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. Hum. Brain Mapp. 27 (5), 392–401.

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

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

Haynes, J.D., Rees, G., 2005. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat. Neurosci. 8 (5), 686–691.

Kamitani, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. Nat. Neurosci. 8 (5), 679–685.

Kanwisher, N., McDermott, J., Chun, M.M., 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. 17, 4302–4311.

Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. U. S. A. 103 (10), 3863–3868.

LaConte, S., Strother, S., Cherkassky, V., Anderson, J., Hu, X., 2005. Support vector machines for temporal classification of block design fMRI data. Neuroimage 26 (2), 317–329.

MacKay, D.J.C., 1994. Bayesian methods for backpropagation networks. In: Domany, E., van Hemmen, J.L., Schulten, K. (Eds.), Models of Neural Networks III, ch. 6, 211–254. Springer-Verlag.

Malach, R., Reppas, J.B., Benson, R.R., Kwong, K.K., Jiang, H., Kennedy, W.A., Ledden, P.J., Brady, T.J., Rosen, B.R., Tootell, R.B.H., 1995. Object-Related Activity Revealed by Functional Magnetic Resonance Imaging in Human Occipital Cortex. Proc. Natl. Acad. Sci. 92, 8135–8139.

Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X., 2004. Learning to decode cognitive states from brain images. Mach. Learn. 57, 145–175.

Mourao-Miranda, J., Bokde, A.L., Born, C., Hampel, H., Stetter, M., 2005. Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data. Neuroimage 28 (4), 980–995.

Neal, R.M., 1996. Bayesian Learning for Neural Networks. Springer.

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes for Machine Learning. MIT Press.

Shen, X., Meyer, F., 2008. Low-dimensional embedding of fMRI datasets. Neuroimage 41 (3), 886–902.

Talairach, J., Tournoux, P., 1998. Co-Planar Stereotaxic Atlas of the Human Brain. Thieme, Stuttgart, Germany.

Tipping, M.E., 2001. Sparse Bayesian Learning and the Relevance Vector Machine. J. Mach. Learn. Res. 1 (3), 211–244.

Vapnik, V.N., 1995. The nature of statistical learning theory. Springer-Verlag, New York.

Yamashita, O., Sato, M., Yoshioka, T., Tong, F., Kamitani, Y., 2008. Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. NeuroImage 42 (4), 1414–1429.