Feature selection and classification of imbalanced datasets. Application to PET images of children with Autistic Spectrum Disorders

Edouard Duchesnay, Arnaud Cachia, Nathalie Boddaert, Nadia Chabane, Jean-Franois Mangin, Jean-Luc Martinot, Francis Brunelle, Monica Zilbovicius

▶ **To cite this version:**

Edouard Duchesnay, Arnaud Cachia, Nathalie Boddaert, Nadia Chabane, Jean-Franois Mangin, et al.. Feature selection and classification of imbalanced datasets. Application to PET images of children with Autistic Spectrum Disorders. NeuroImage, 2011, 57 (3), pp.1003-1014. 10.1016/j.neuroimage.2011.05.011 . cea-01328634

HAL Id: cea-01328634
https://cea.hal.science/cea-01328634

Submitted on 8 Jun 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Feature selection and classification of imbalanced datasets. Application to PET images of children with Autistic Spectrum Disorders

Edouard Duchesnay[a,b,d,*], Arnaud Cachia[g], Nathalie Boddaert[e,c,d], Nadia Chabane[f,b], Jean-Franois Mangin[a,b,d], Jean-Luc Martinot[b,d], Monica Zilbovicius[c,d]

[a] *CEA, Neurospin, LNAO, Gif-sur-Yvette, France*
[b] *INSERM-CEA U1000, Neuroimaging & Psychiatry Unit, SHFJ, Orsay, France*
[c] *INSERM-CEA U1000, Neuroimaging & Psychiatry Unit, Necker Hospital, Paris, France*
[d] *IFR49, Institut d'Imagerie Neurofonctionnelle, Paris, France*
[e] *Service de Radiologie Pédiatrique, AP-HP, Necker Hospital, Paris,France*
[f] *Service de Psychopathologie de l´Enfant et de l´Adolescent, AP-HP, Robert Debré Hospital, Paris, France*
[g] *UMR 894 INSERM and Paris Descartes University, Center for Psychiatry & Neurosciences, Sainte-Anne Hospital, Paris, France*

Abstract

Learning with discriminative methods is generally based on minimizing the misclassification of training samples, which may be unsuitable for imbalanced datasets where the recognition might be biased in favor of the most numerous class. This problem can be addressed with a generative approach, which typically requires more parameters to be determined leading to reduced performances in high dimension. In such situations, dimension reduction becomes a crucial issue. We propose a feature selection / classification algorithm based on generative methods in order to predict the clinical status of a highly imbalanced dataset made of PET scans of forty-five low-functioning

*Corresponding author: Edouard Duchesnay, CEA, Neurospin, LNAO, Bâtiment 145, 91191 Gif-sur-Yvette, France. *Email address:* edouard.duchesnay@cea.fr

children with autism spectrum disorders (ASD) and thirteen non-ASD low-functioning children. ASDs are typically characterized by impaired social interaction, narrow interests, and repetitive behaviours, with a high variability in expression and severity. The numerous findings revealed by brain imaging studies suggest that ASD is associated with a complex and distributed pattern of abnormalities that makes the identification of a shared and common neuroimaging profile a difficult task. In this context, our goal is to identify the rest functional brain imaging abnormalities pattern associated with ASD and to validate its efficiency in individual classification. The proposed feature selection algorithm detected a characteristic pattern in the ASD group that included a hypoperfusion in the right Superior Temporal Sulcus (STS) and a hyperperfusion in the contralateral postcentral area. Our algorithm allowed for a significantly accurate (88%), sensitive (91%) and specific (77%) prediction of clinical category. For this imbalanced dataset, with only 13 control scans, the proposed generative algorithm outperformed other state-of-the-art discriminant methods. The high predictive power of the characteristic pattern, which has been automatically identified on whole brains without any priors, confirms previous findings concerning the role of STS in ASD. This work offers exciting possibilities for early autism detection and/or the evaluation of treatment response in individual patients.

Keywords: multivariate classification, autism, features selection, dimension reduction

---

1. Introduction

Multivariate machine-learning methods offer a wide range of new applications in the neuroimaging field. In cognitive neurosciences, such methods can process fMRI intra-subject scans to decode subject's mental states (Thirion et al. (2006); Pereira et al. (2009)). In a clinical context, they can be applied on several subjects' scans that stem from different groups (e.g., case, control) in order to identify the imagery pattern associated with the group differences. The pattern, which is generally a combination of spatially-distributed imagery biomarkers, can be applied on individual scans to predict the subject group ownership. This enables a computer-aided diagnosis perspective in both neurological (Klöppel et al. (2008)) diseases or psychiatric (Fan et al. (2007)) disorders. This also offers possibilities of image-based phenotyping that can be associated with genetic data. A key feature of those methods is their potential to detect global, complex and distributed patterns of "abnormalities" that cannot be efficiently identified with univariate voxel-based methods whose sharp and localized view field (voxel) yield a reduced sensitivity.

Multivariate classification can be based on generative or discriminative approaches. Discriminative classifiers directly learn a mapping from the inputs $\mathbf{x}$ to the output label y. For example, a probabilistic method such as the logistic regression, models the posterior $p(y|\mathbf{x})$ distribution of the label, which require the estimation of P parameters (in a P dimensional space). Discriminative methods minimize the misclassification of training samples, which may be unsuitable for imbalanced datasets that may bias the recognition in favor of the most numerous class. Reweighting or resampling can

be used to "rebalance" samples of the least numerous class. However, at least for reweighting techniques, we demonstrate that such heuristics do not provide a satisfying solution on our dataset. On the other hand, generative classifiers learn the full joint distribution $p(\mathbf{x}, y)$. First, they learn the classes conditional densities $p(\mathbf{x}|y)$, which can be done independently for each class. Such approach is closely related to two separate one-class learning, which is a recent approach (Chawla et al. (2004)) to deal with in imbalanced situations. The issue here is that the parameters estimation of the minority class will be poorer, which does not imply a narrower distribution and thus a prediction bias toward the other class. Second, the predictive function $p(y|\mathbf{x})$ is obtained by combining $p(\mathbf{x}|y)$ with an explicit class priors $p(y)$ using Bayes rule and choosing the most probable label. Such models provide a better control over class disequilibrium. The main difficulty is the estimation of $p(\mathbf{x}|y)$ in high dimensional space. For typical generative classifier such as the linear discriminant analysis (LDA), the estimation of the class means and the pooled covariances matrix, i.e. $P(P+5)/2$ parameters, is required. This must be compared to the P estimated parameters of linear discriminative methods. Such high number of parameters leads to severe risk of over-fitting, in a high dimensional space, which require to associate generative classifiers with an efficient dimension reduction strategy.

Dealing with imbalanced class may be addressed by three main ways (see Japkowicz and Stephen (2002) for a review), resampling, reweighting and one class learning. In sampling strategies, either the minority class is oversampled or majority class is undersampled or some combination of the two is deployed. Undersampling (Zhang and Mani (2003)) the majority class

would lead to a poor usage of the left-out samples. We cannot afford such strategy since wee are also facing a small sample size problem even for the majority class. Indeed there are only 45 ASD subjects in a $P \approx 200,000$ dimensional space. Informed oversampling, which goes beyond a trivial duplication of minority class samples, require the estimation of class conditional distributions in order to generate synthetic samples. Here generative models are required. An alternative, proposed in Chawla et al. (2002) generate samples along the line segments joining any/all of the k minority class nearest neighbors. Such procedure blindly generalizes the minority area without regard to the majority class, which may be particularly problematic with high-dimensional and potentially skewed class distribution. Reweighting, also called cost-sensitive learning, work at an algorithmic level by adjusting the costs of the various classes to counter the class imbalance. Such reweighting can be implemented within SVM (Chang and Lin (2001)) or logistic regression (Friedman et al. (2010)) classifiers. One class learning is a recognition-based rather than discrimination-based learning, where classes are learned separately. It is then related to generative methods where classes conditional densities $p(\mathbf{x}|y)$ are estimated (almost) independently. However, non parametric methods such as SVMs can be used: Raskutti and Kowalczyk (2004) show that one class learning SVMs outperformed discriminative two class SVMs in high dimensional noisy feature space. Finally, it is essential to use appropriate performance evaluation measurements such as ROC (receiver operating characteristic) analysis and AUC (area under curve) measure.

In this paper we propose to evaluate the potential of multivariate machine-learning for identification of a functional PET imaging pattern shared by

children with autism spectrum disorders (ASD) in the specific context of an imbalanced dataset. Such a method may offer perspective in the context of early detection of autism or in evaluating the effects of treatments on a subject. Autism is a complex neurodevelopmental disorder characterized by deficits in social functioning and communication as well as restricted, repetitive behaviors and interests (DSM IV). The several findings in neuroimaging studies, briefly reviewed in the discussion section, support the hypothesis that ASD is linked with complex patterns of brain abnormalities that are distributed across the brain and possibly organized as networks. Consequently, multivariate methods are implemented to reach our overall goal: the identification of rest functional brain imaging networks of abnormalities associated with ASD and the evaluation of their efficiency in individual classification.

Previous neuroimaging studies of ASD using multivariate classifiers were based on global brain measurements, Akshoomoff et al. (2004) classified children with autism based on global *a priori* pre-selected cerebellar and cerebral white or gray matter volumes. Neeley et al. (2007) measured the white and gray matter volumes within manually identified regions including five temporal gyri, the amygdala and the hippocampus. They then used a CART (classification and regression trees) method that achieved high specificity in classifying autism subjects from matched IQ controls based on the relationship between the volume of the left fusiform gyrus gray and white matter and the volume of the right temporal stem and right inferior temporal gyrus gray matter. More recently, Ecker et al. (2010b) applied a SVM (Support Vector Machine) classifier on whole brain structural MRI in order to detect adults with ASD from healthy matched controls. In this context, the first

goal of this paper was to conduct a whole brain *a priori* free evaluation of multivariate classification on functional PET imaging.

However, such a multivariate strategy brings up a major methodological issue: combining all voxels allows many ways to distinguish two populations. Unfortunately, most of these distinctions cannot be generalized to larger populations because they exploit spurious differences embedded in the feature measurement noise. This well-known issue, called the curse of dimensionality, arises when the dimension of the underlying space ($\approx 200,000$ voxels in our case) is large compared to the number of subjects (58 in our case). To overcome this difficulty, a feature (voxel) selection step may be necessary in order to select a subset of useful features that constitute a characteristic pattern on which to build a robust classifier. Feature selection generally consist of two steps: the first step ranks subsets of features $[F_1, \ldots, F_k, \ldots, F_P]$, where F_k is the best combination of k features and P is the total number of available features, and the second step selects an optimal subset to be used in the final classifier.

The first step of feature subset ranking may be addressed with three categories of strategies: filters, wrapper and embedded methods (Guyon et al. (2006)). Filters rank feature subsets independently of the final predictor and are generally assimilated to mass-univariate feature ranking. In the context of neuroimaging data analysis, filters can thus be related to voxel-based analysis. They are computationally efficient and more robust to over-fitting than multivariate methods (Guyon and Elisseeff (2003)). However, filters raise the issue of determining a significance level that accounts for such a multiple testing procedure. Moreover, they are blind to feature interrelations,

a problem that can be addressed only with multivariate selection such as wrappers or embedded methods. Wrappers (Kohavi and John (1997)) are so called because they wrap around an objective function that is supposed to portray the predictor performances. Wrappers are generic stepwise-like optimizers that explore the features space with a greedy forward, backward or combined strategy. Finally, embedded methods are directly plugged into the predictor. Many predictive algorithms have such built-in procedures; for example, random forest (Breiman (2001)) or all L1 penalized (logistic) regressions such as Lasso (Tibshirani (1996)). Other predictor algorithms, like support vector machine (SVM), that were not originally designed with an embedded feature selection were modified to do so: the SVM-RFE (Guyon et al. (2002)) alternates the fit of a SVM with a recursive feature elimination (RFE) procedure based on the input features' weight. Most of those feature subset ranking strategies can be assimilated to feature ranking since they produce nested subsets of features. However, alternative strategies that mix forward and backward steps or other Lasso-based methods may yield to non-nested subsets that justify the term feature subset ranking to designate this first step of feature selection. Moreover, such generic procedures ignore the three dimensional structure of the images, thereby entailing the selection of a scattered and widespread discrimination voxels map as shown in Ecker et al. (2010b), where a generic SVM-RFE is used. Consequently, the second goal of this article is to push further the feature selection by proposing a new strategy that combines univariate and multivariate strategies and by exploiting the images' three dimensional structure by merging voxels of the same neighborhood (within a few regions), producing much more parsimonious

and interpretable results.

Surprisingly, the second step of model selection, which consists of determining the optimal subset (F_k) of features, is generally not addressed by authors, including Ecker et al. (2010b). Classification results are often presented as a function of a varying number of voxels, which may lead to an optimistic interpretation as remarked in Reunanen (2003). Indeed, the significance of the classification rates must be corrected with multiple repetitions of the classification experiments. An alternative but computationally-costly solution is to determine F_k based on cross-validation. For the third contribution of this article, we propose to address the choice of the optimal feature subset as a model selection problem using an automatically calibrated adaptive penalization of the likelihood, offering excellent performance with negligible computation overhead once the calibration has been done.

As a summary, we hypothesized that ASD is associated with global, complex and distributed patterns of abnormalities. In this context, our first goal is to identify rest functional brain imaging networks of abnormalities associated with ASD and evaluate their efficiency in individual classifications. Following this track, the second contribution of the paper is to propose a new feature selection algorithm that combines univariate and multivariate strategies to exploit the images' three-dimensional structure and produce regional features subsets of increasing size. For the third contribution of this article, we propose to address the choice of the optimal feature subset through an automatically calibrated model selection with the goal of identifying parsimonious syndrome-specific PET brain imaging indexes (biomarkers).

2. Feature selection and classification methods

Here, we briefly present our multi-stage feature selection and classification algorithm (Figure 1). It is based on the same approach we used in Duchesnay et al. (2007) but with specific adaptations to deal with the particular size (hundreds of thousands of dimensions) and the three dimensional topology of PET images.

Insert Figure 1 about here

2.1. Regional features extraction

The first step of the pipeline is a univariate feature selection based on p-values derived from two-sample F-tests (Figures 1 step 1 and 3 step 1.1). Only voxels with a p-value < 0.001, uncorrected for multiple comparisons, were retained (Figure 3 step 1.2). Thresholded-connected voxels were grouped into regions (or clusters), and the PET signal was averaged within each region, producing a set of new regional features (Figure 3 step 1.3). This threshold was empirically chosen among the three possible values of 10^{-2}, 10^{-3} and 10^{-4} as the value that produces clusters of reasonable size.

2.2. Feature subset ranking

The resulting regional features were used as the input of a wrapper; i.e., a multivariate stepwise feature selection (step 2 in Figures 1 and 3). This approach was based on a sequential floating forward selection (SFFS, Pudil et al. (1994)), which is a hybrid strategy that includes a backward loop that deletes the worst feature, within a forward loop that adds the best feature.

Characteristic combinations of features that lead to efficient separation of the two groups were evaluated using a multivariate F-statistic known as

the Pillai-Bartlett trace (Hand and Taylor (1987) $tr((\mathbf{V}^k)^{-1}\mathbf{B}^k)$, where $\mathbf{V}^k$ and $\mathbf{B}^k$ are the total and the between groups variance matrices evaluated on the features subset F_k. The output of this step was a list of subsets of features: $[F_1, \ldots, F_k, \ldots, F_P]$, where F_k was the best-identified combination of k features for a maximum of P regional features.

2.3. Model selection

The third stage aims at selecting the optimal subset of features (F_k) to be used by the final classifier (step 3 in Figures 1 and 3). This was addressed as a model selection problem using a penalized likelihood framework.

Many criteria have been proposed (Burnham and Anderson (2004)), and the two most commonly used are the Bayesian (BIC, Schwarz (1978)) and Akaike information criteria (AIC, Akaike (1974)). Despite their different theoretical foundations, both yield a similar linear penalization of the log-likelihood with the number of parameters. The use of fixed penalties in those criteria is mainly justified by asymptotic arguments that may be wrong in a non asymptotic context. This limitation motivated the development of data-driven methods to calibrate criteria whose penalties are known up to a multiplicative factor, e.g. "the slope heuristics" proposed by Birgé and Massart (2007). Moreover, our two-stages feature selection procedure prevents from a straightforward application of a fixed penalty as function of the number of regional features. Indeed, such penalty would ignore the overfit induced by the first step of region extraction. We demonstrated this under-penalization of fixed penalties criteria in Figure 2. A solution, to avoid such under-penalization, would have been be to include the number of voxels, used be to build the regions, into the penalization term. However this raises the

complex question of defining a good penalty "form" that would be a function of the number of voxels combined with the number of regional features.

Adaptive log-likelihood penalization. Instead of the previous solution and likewise the "the slope heuristics" of Birgé and Massart (2007), we retain the initial idea of a linear penalization as function the number regional features (BIC like) but we loosen the fixed link by adding a free parameter (noted "a"):

$$\ln p(\mathbf{y}|\mathbf{X}^k, F_k) \simeq \ln p(\mathbf{y}|\mathbf{X}^k, \boldsymbol{\theta}^k, F_k) - a\frac{1}{2}k \ln N \tag{1}$$

Where $\mathbf{y} = (y_1, \ldots, y_N)$ are the subjects labels ($y_i \in \{1, 0\}$, 1 denotes ASD and 0 denotes low-functioning) and $\mathbf{X}^k = (\mathbf{x}_1^k, \ldots, \mathbf{x}_N^k)^t$ are the F_k (regional) features ($\mathbf{x}_i^k \in \mathbb{R}^k$). The evidence $p(\mathbf{y}|\mathbf{X}^k, F_k)$ of the model built with the F_k features is approximated with the likelihood $p(\mathbf{y}|\mathbf{X}^k, \boldsymbol{\theta}^k, F_k)$, adding a penalization term ($k \ln N$) whose weight can be adapted with the free parameter a. The parameters $\boldsymbol{\theta}^k$ are obtained by maximizing the likelihood itself:

$$p(\mathbf{y}|\mathbf{X}^k, \boldsymbol{\theta}^k, F_k) = \prod_{i=1}^{N} p(y_i = 1|\mathbf{x}_i^k, \boldsymbol{\theta}^k, F_k)^{y_i} p(y_i = 0|\mathbf{x}_i^k, \boldsymbol{\theta}^k, F_k)^{(1-y_i)} \tag{2}$$

where $p(y_i = 1|\mathbf{x}_i^k, \boldsymbol{\theta}^k, F_k)$, which denotes the probability that an input vector ($\mathbf{x}_i^k$) belongs to the ASD group ($y_i = 1$); this is detailed in the Section 2.4. Finding the F_k that maximizes the evidence $p(\mathbf{y}|\mathbf{X}^k, F_k)$ (1) implies that the penalization on the likelihood yields a good approximation of the evidence.

Penalization calibration based on randomized data. In order to adapt this penalization value (a) to the over-fitting caused by the feature selections algorithm we proposed to calibrate a under the null hypothesis estimated from randomized datasets. Under such hypothesis we measured the increase of the

log-likelihood purely due to over-fitting of the training data and compared it to the theoretical log-evidence, which is supposed to be constant and equal to $\ln(1/2)^N$. A good penalization is supposed to fit this increase in order to approximate the (constant) log-evidence. The experiment, conducted on all samples, presented in Figure 2 clearly shows that our feature selection algorithm dramatically increases the over-fitting that cannot be balanced with the BIC or AIC penalization criteria. However, this experiment also suggest that a satisfying linear approximation can be estimated. Consequently, we repeated the latter experiment independently for every cross-validation fold excluding the test sample (column (a) in Figure 1). The estimated adaptive penalization values (a) (average across folds=2.62, SD=0.03) were then plugged in (1) (see Figure 1) and evaluated for all F_ks. The features subset (F_k) that maximized this penalized log-likelihood was selected for the classification step.

Insert Figure 2 about here

2.4. Classification

The last stage of the pipeline is a classifier based on a linear discriminant analysis (LDA) (step 4 in Figures 1 and 3). To compute the likelihood of (2), LDA provides the probability that an input vector $\mathbf{x}_i^k$ belongs to the group $\mathcal{G} \in \{1, 2\}$:

$$p(y_i = \mathcal{G}|\mathbf{x}_i^k, \boldsymbol{\theta}^k, F_k) = \frac{\pi_{\mathcal{G}} \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_{\mathcal{G}}^k, \boldsymbol{\Sigma}^k)}{\sum_{\mathcal{G}\in\{1,2\}} \pi_{\mathcal{G}} \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_{\mathcal{G}}^k, \boldsymbol{\Sigma}^k)} \tag{3}$$

where the estimated parameters $\boldsymbol{\theta}^k$ are the means of the two groups $\boldsymbol{\mu}_1^k, \boldsymbol{\mu}_2^k$ and the pooled covariance $\boldsymbol{\Sigma}^k$. Priors over the two classes were both set to

50% ($\pi_0 = \pi_1 = \frac{1}{2}$) in order to avoid any bias caused by imbalanced sizes of the two groups (45 ASD children vs. 13 non-ASD, low-functioning children).

3. Performances validation and comparison methodology

3.1. Classification performances

The classification accuracy of the entire pipeline (including the feature selection) was evaluated by leave-one-out cross-validation (LOO-CV), column (b) in Figure 1. This provided an almost unbiased estimate of the actual expected accuracy (Ambroise and McLachlan (2002); Kohavi (1995)). We note that the subject to be tested was set aside before any processing: all the feature selection steps and the final classification were performed for each LOO-CV iteration, taking into account only the (58-1 = 57) training subjects. The group was predicted for each test subject and was compared with the actual group. Finally, all the predictions were averaged to evaluate classification performance. We tested the significances of accuracy, sensitivity and specificity against the null hypothesis that the classification was made with random choice; i.e., $p(\text{ASD}) = p(\text{low-functioning}) = \frac{1}{2}$. The sensitivity and specificity are calculated as follow: sensitivity = TP/(TP+FN), specificity = TN/(TN+FP) where TP is the number of true positives, i.e., the number of ASD images correctly classified; TN is the number of true negatives, i.e., number of non-ASD images correctly classified; FP is the number of false positives, i.e., number of non-ASD images classified as ASD; and FN is the number of false negatives, i.e., number of ASD images classified as non-ASD. We also conducted ROC analysis and evaluated the significance of AUC scores using the Wilcoxon non-parametric test of rank (Mason and

Graham (2002)).

Furthermore, to confirm the significance of classification and prevent any bias that could stem from imbalanced groups sizes, we conducted 1000 randomized permutations. Subjects were randomly assigned to the low-functioning or ASD group while keeping the total number of subjects per group the same. Then for each permutation, the complete LOO-CV of the pipeline was performed exactly as mentioned herein as shown by column (c) in Figure 1.

3.2. Comparison with other dimension reduction and classification strategies

As a *post hoc* experiment, we compared the proposed algorithm with other strategies made of alternative options of dimension reduction and classification methods.

Alternative feature subsets ranking methods. We tested: (i) no ranking (using all features); (ii) univariate (F-test) and (iii) multivariate feature ranking based on RFE applied to both SVM and LDA (iv) finally, a multivariate feature selection based on Lasso. Indeed, solving the Lasso problem along an entire path of values for the regularization parameter (the λs of (5)) yields the selection of features subsets of various sizes, where each subset is the active set of features with nonzero coefficients.

Alternative model selection methods. We compared the proposed adaptive penalization (aPena) framework with a classic ten-fold cross-validation (CV).

Alternative classifiers. We compared the proposed generative methods with other discriminative-based settings:

(i) a linear SVM (based on libsvm Chang and Lin (2001)) with the constant of the regularization term set to one and an individual re-weighting (Veropoulos et al. (1999); Osuna et al. (1997)) of samples to compensate for an under-represented group of low-functioning subjects. This yielded to the following SVM primal formulation:

$$\min_{\mathbf{w}} ||\mathbf{w}||_2 + C \sum_{y_i=1} \xi_i + C \frac{\#\{y_i = 1\}}{\#\{y_i = 0\}} \sum_{y_i=0} \xi_i \tag{4}$$

where $\mathbf{w}$ is the weights vector, and ξ_i are the positive slack variables that measure the distance to the margin of samples that are on the wrong side of the margin: $(0 < \xi_i \leq 1)$ for correctly classified ones and $(1 < \xi_i)$ for misclassified samples. $\#\{y_i = 0\}$ (resp. $\#\{y_i = 1\}$) is the number of low-functioning (resp. ASD) samples in the current training set. The regularization parameter (C) was defined using two strategies: in a first experiment we used the default value i.e.: one. Then we used a ten-fold cross-validation (CV) on the training samples to select the value with the minimum error. We only retained the results with the default value (one) since this setting almost always outperformed the CV-based setting.

(ii) A sparse logistic regression based on a L1 (Lasso) penalization, which is a penalized version of (2):

$$\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \lambda) = \sum_{i=1}^{N} y_i \log p(y_i = 1|\mathbf{x}_i, \boldsymbol{\beta})(1 - y_i) \log p(y_i = 0|\mathbf{x}_i, \boldsymbol{\beta}) - \lambda ||\boldsymbol{\beta}||_1 \tag{5}$$

where $p(y_i = 1|\mathbf{x}_i, \boldsymbol{\beta}) = 1/(1+\exp(\boldsymbol{x_i}^T \boldsymbol{\beta}))$ is the logistic link function with the linear regression of parameter $\boldsymbol{\beta}$ and λ is the regularization parameter.

As evoked herein, we solved this Lasso problem with various values of λ leading to a path made of active set of features of various size. This penalized log-likelihood was optimized with a coordinate descent procedure (Friedman et al. (2010)). As with SVM, samples were re-weighted to compensate for imbalanced groups sizes.

By combining a selected choice of those alternatives, we formed four strategies:

1. No feature selection combined with a linear (re-weighted) SVM classifier. This strategy acts as a baseline to highlight the specific contributions of feature selection.
2. Univariate F-test feature subset ranking, model selection with CV combined with a linear (re-weighted) SVM classifier. This strategy acts as a baseline to highlight the specific contributions of the two last strategies based on multivariate feature selection.
3. Multivariate RFE-based feature subset ranking, model selection with CV combined with a linear (re-weighted) SVM. This strategy is commonly used as a state-of-the-art multivariate feature selection with a kernel-based classifier (Guyon et al. (2006); Ecker et al. (2010b)).
4. Multivariate Lasso-based feature subset ranking, model selection with CV combined with a linear (re-weighted) Lasso logistic regression classifier. This strategy is a state-of-the-art representative of recent advances in L1-constraint based methods.

Those four strategies were first directly applied on entire brain images (hundreds of thousands of voxels), and then on the regional features ($\approx$10). The low dimension space made of regional features allowed us to test a gen-

erative (LDA) classifier, replacing the SVM by the LDA within the strategies evoked previously. Additionally, we tried the computationally expensive wrapper feature subsets ranking (SFFS), that we found to be efficient in a low dimensional space (Duchesnay et al. (2007)). Finally, in order to evaluate the contribution of the proposed model selection method (aPena), we combined it with all multivariate feature subsets ranking methods, by replacing the CV-based model selection with a proposed aPena framework.

As a summary, we had the following design of experiments (see Table 2): four voxels-based analyses, four regional-based analyses, four regional-based analyses with LDA and three regional-based analyses with aPena model selection. As recommended in Dietterich (1998), we compared the prediction rates of these alternatives strategies using the McNemars test which has a low type I error.

4. Dataset

Forty-five children with idiopathic autism spectrum disorder (37 boys) were selected among patients attending a specialized autism consultation sessions at a university hospital in France. They were aged from 5 to 12 years (mean = 7.9, SD = 2.2). They were recruited at two university hospitals with referral centers dedicated to the assessment and treatment of autism by the French Health Ministry. Diagnosis was performed in these units by a multidisciplinary team including child psychiatrists, child psychologists and speech therapists, during 3-7 days of extensive evaluation. ASD diagnosis was based on DSM IV-R criteria (APA (2000)) and confirmed by ADI-R scores (Lord et al. (2001)). Exclusion criteria included infectious, metabolic,

neurological or genetic diseases, chromosomal abnormalities and seizures. All children were also evaluated by a pediatric neurologist and a clinical geneticist. The recommended biological and medical screenings for ASD were performed, including high-resolution karyotyping, DNA analysis of FRA-X and normal standard metabolic testing (plasma and urine amino and organic acid analysis, urine glycosaminoglycans (GAG) quantitation, urine oligosaccharide, purine and pyrimidine analysis, and creatinine guanidoacetate urine analysis). Mental capacity was assessed by an intelligence quotient (IQ) determined with the Wechsler Intelligence Scale for Children (WISC-III). The Developmental Quotient (DQ) was obtained in all children younger than 6 years ($N = 11$). Developmental quotient (DQ) was determined with the Psycho-Educational-Profile Revised (PEP-R) and the Brunet-Lzine developmental tests (IQ = 45 $\pm$ 22; DQ = 44 $\pm$ 23). As a comparison group, we selected 13 non-ASD low-functioning children (9 boys). They were aged from 5 to 15 years (mean = 8.6, SD = 2.7 years). Their mean IQ was 48 (SD = 14.5). They had idiopathic mental retardation according to the DSM-IV criteria with no associated neurological disorder. The following conditions were excluded: known infectious, metabolic or chromosomal diseases, epilepsy or recognizable neurological syndromes. This comparison group was chosen in order to detect abnormalities specifically related to autism and to evaluate features that cannot be attributed to mental retardation, taking into account the fact that mental retardation is often associated with autism, and that we have studied a low functioning group of children with autism. No specific etiology for either ASD or non-ASD children was found following extensive clinical and laboratory investigations. All children had been medication-free for

at least one month prior to imaging. Written informed consent was obtained from all the children's parents. Cerebral blood flow (rCBF) was measured using positron emission tomography (PET) (Siemens ECAT Exact HR+ 962) following intravenous injection of $H_2{}^{15}O$. Attenuation-corrected data were reconstructed into 63 slices with a resulting resolution of 5 mm, full width at half maximum. Fifteen seconds before each scan, 7 mCi of $[^{15}O]H_2O$ was administrated by an intravenous bolus injection. Data were collected over a period of 80 seconds. In all children, PET studies were performed during sleep induced by premedication with rectal pentobarbital (7 to 10 mg/kg) to obtain perfect motionlessness. A previous study showed that sedation does not change either the global rCBF or local rCBF distribution (Zilbovicius et al. (1992)). The study was approved by the local ethics committee. PET scans were first spatially normalized (Friston et al. (1995)) to a standard stereotactic space and smoothed with an isotropic Gaussian filter (full width at half maximum of 15 mm). Global intensity differences between subjects were corrected using proportional scaling. Normalized, smoothed and scaled images were used for subsequent processing.

5. Results

5.1. Regions involved in the classification

Figure 3 and Table 1 show regions that were selected by the multi-step feature selection method. The first step of regional feature extraction led to few characteristic regions which formed a network of abnormalities (Figure 3 before step 2).

Insert Figure 3 about here

Insert Table 1 about here

Four of those regions featured hypoperfusion in the ASD group. The first hypoperfused region (i) concerned the right temporo-parietal junction (RTPJ). Two additional hypoperfused regions were found in the right temporal lobe: (ii) the Superior Temporal Sulcus (STS), (iii) the middle temporal gyrus. The fourth region (iv) was found in the posterior zone of the corpus callosum where it overlaps with the right posterior cingulum and bilateral thalami. Finally, two hyperperfused regions in the ASD group were identified in (v) the left post-central and (vi) the right pre-central areas. The second and third steps of identifying characteristic regions selection took into account characteristic interrelations among those regions. This led to the selection of the final characteristic pattern in the ASD group that featured a hypoperfused region in the right superior temporal sulcus and a hyperperfused region in the left post-central (Figure 3 after step 3 and Figure 4).

Insert Figure 4 about here

5.2. Performance evaluation of individual classification

The performance of the whole pipeline was evaluated by LOO-CV (column (b) in Figure 1). The rate of correct recognitions was 88% (51 correct recognitions across 58 subjects). The method achieved 91% sensitivity (41 correct recognitions across 45 ASD subjects) and 77% specificity (10 correct recognitions over 13 non-ASD, low-functioning subjects). At a chance level of 50%, both sensitivity and specificity recognition rates were significant with, respectively, $p < 0.0001$ and $p = 0.046$. In order to take into account the highly imbalanced dataset we also tested the significances with

the chance levels being the respective proportions of the two classes (77.6% for the ASD group and 22.4% for the low-functioning group). Both sensitivity and specicity recognition rates were significant with, respectively $p = 0.0162$ and $p < 0.0001$. Significances of the prediction rates were also confirmed with a random permutation of the group label. The average prediction rates of the pipeline in over 1000 random permutations achieved nearly the expected rates of random choice (51% accuracy, 53% sensitivity and 45% specificity). Comparing the random prediction rates to true ones (88%, 91%, 77%) yield significances of $p < 0.001$ for sensitivity and $p = 0.02$ for specificity. Finally, the ROC curve of the LOO validation (Figure 5), confirmed a good sensitivity/specificity trade-off: the area under curve (AUC) for the proposed method reached a significant ($p < 0.001$, with a Wilcoxon test of ranks) score of 0.81.

5.3. Comparison with other dimension reduction and classification strategies

Table 2 shows that the proposed algorithm outperformed the fourteen alternative strategies of feature selection and classification. Better performances were systematically obtained for accuracy, sensitivity and specificity. Those results were confirmed with a comparison using the McNemars test: in ten over fourteen cases we found a significantly better accuracy. In two (#1 and #2) of the four remaining cases the proposed algorithm was found to be significantly more specific. Finally, improvements were not found to be significant for the last two cases (#12 and #14).

Insert Table 2 about here

ROC curves in Figure 5 and significant AUC scores in Table 2 highlight

that the three last strategies (#13, #14 and #15) reached good sensitivity while keeping an acceptable specificity (low false positive rate). This demonstrates the relevance of a strategy that combines regional feature extraction, multivariate feature subsets ranking and calibrated (aPena) model selection.

Insert Figure 5 about here

SVM failed to obtained specific predictions: Experiments based on SVM classifier were significantly ($p < 0.01$, Welch two-sample t-test) less specific (mean = 32%, SD = 7%)) than those based on a generative LDA classifier (mean = 54%, SD = 17%)). This suggests the relevance of generative models to deal with imbalanced datasets.

6. Discussion

This paper presents a novel method of classifying brain images that allowed us to detect a characteristic pattern of abnormalities from resting PET images in patients with ASD. This pattern is composed of: (i) a hypoperfusion (in the ASD group) in the right Superior Temporal Sulcus (STS) and (ii) a hyperperfusion (in the ASD group) in the contralateral postcentral area (Figure 4). The algorithm that identified this pattern achieved significant clinical classification accuracy. The prediction performance was sensitive and specific. The results were validated with a leave-one-out procedure: trained classifiers that included a selection of features were tested on unseen PET images. Moreover, the significance of those cross-validated classification rates was confirmed with random permutations. This strongly suggests that similar results may be generalized to other datasets. These first

classifications of ASD on the basis of functional brain imagery suggest great promise for future applications of early detection of autism or in evaluating the effects of treatments on a subject.

6.1. Brief review of neuroimaging findings in ASD

Early brain imaging studies focused on manual and *a priori* defined regions of interest (ROIs) such as the cerebellum (Courchesne et al. (1988)), the amygdala (Howard et al. (2000)), the hippocampus (Aylward et al. (1999)), the corpus callosum (Egaas et al. (1995)) and the cingulate (Haznedar et al. (1997)). Despite methodological design limitations (e.g., IQ heterogeneity, age range of ASD subjects, inclusion of epileptic ASD subjects), failures of these classical studies to replicate localized brain anomalies in ASD may be attributed to the ROI approach that is inherently subjective and operator-dependent.

The development of automated voxel-based methods allowed whole-brain and operator-independent analyses. Such methods have been used for PET / SPECT imaging at rest, structural MRI, diffusion and functional MRIs. With PET / SPECT imaging at rest, significant functional hypoperfusion in autism has been identified in the STS and in the superior temporal gyrus (Ohnishi et al. (2000); Zilbovicius et al. (2000)). Voxel-based morphometry (VBM) could identify many gray matter abnormalities associated with ASD in frontal, superior temporal, parietal and striatal regions (Abell et al. (1999); Boddaert et al. (2004); McAlonan et al. (2005); Hadjikhani et al. (2006)). Using a voxel-based analysis of diffusion MRI, reduced fractional anisotropies (FA) were found in the ventromedial prefrontal cortex, anterior cingulate gyri, temporoparietal junctions, bilateral STS (Barnea-Goraly

et al. (2004)) and corpus callosum (Keller et al. (2007); Alexander et al. (2007)). Functional MRI (fMRI) studies have focused primarily on social cognition. Most have shown activation abnormalities in the regions involved in language/voice and face perception, theory of mind and the mirror neuron system. At least five fMRI studies have shown that ASD subjects exhibit reduced levels of activity in the fusiform face area (FFA) during face perception (for a review, see Schultz (2005)). However, some studies report fusiform activation in autism when comparing faces to non-facial stimuli (Hadjikhani et al. (2004)). This seems to be associated with identifying familiar faces (Pierce et al. (2004)) and is correlated with the degree of eye gaze fixation on faces for the autism group (Dalton et al. (2005)). A voice perception fMRI study (Gervais et al. (2004)) identified significant differences in the pattern of brain activation along the upper bank of the STS. Concerning the theory of mind, several studies have found abnormal activation patterns. An fMRI study (Baron-Cohen et al. (2000)) found that ASD subjects did not activate the amygdala, and two other studies (Castelli et al. (2002); Pelphrey et al. (2005)) found that ASD subjects exhibit reduced activation in the STS and frontal regions. Concerning the mirror neuron system, a fMRI study (Dapretto et al. (2006)) found no activation in the inferior frontal gyrus.

Recently, (Ecker et al. (2010b)) applied a multivariate SVM classifier with RFE feature selection on an anatomical MRI. They found a widespread pattern that involved the limbic, frontal-striatal, fronto-temporal, fronto-parietal and cerebellar systems. It must be noted that they found a gray matter increase (in the ASD group) in the bilateral STS and a white matter increase (in the ASD group) in bilateral postcentral gyri. Ecker et al. (2010a)

also used a SVM classifier without feature selection on several parameters extracted from the cortical surface issued from anatomical MRI. Concerning the cortical thickness, the excess pattern in the ASD group was comprised of predominantly occipito-temporal regions, while the pattern displaying a relative thinning of the cortex in ASD versus controls (i.e., deficit pattern) included mainly frontal and parietal regions.

6.2. Relations with regions identified with the proposed multivariate method

Our feature selection procedure applied to hundreds of thousands of voxels identified a parsimonious characteristic pattern made of only two regions (right STS, left postcentral). The selection of hypoperfusion of the right STS corroborates with findings from other studies (Ohnishi et al. (2000); Zilbovicius et al. (2000)). Further investigations must be done to explain the seeming contradiction between our hypoperfusion finding and the increase of gray matter (in the ASD group) found by Ecker et al. (2010b) in the same temporal area. We note that when considered individually using a classical univariate voxel-based method, the STS is the least significant temporal region (Table 1). However, our feature selection method, that took advantage of characteristic co-variations, selected this only temporal area. This suggests that the abnormalities of autism are not simple additions of local differences but may constitute more complex and distributed patterns. The other region involved in the classification is the left postcentral. In this region we observed a hyperperfusion in the ASD group. Since a similar and highly correlated (the correlation of 0.73 is displayed as a link in Figure 3) abnormality is also found in the right hemisphere (Table 1, #vi), this finding can be related to two fMRI studies (Pierce et al. (2004); Müller et al.

(2004)) that report a greater BOLD response in the right postcentral gyrus in ASD patients relative to a normal group. Moreover, those functional abnormalities may be related to the structural findings of Ke et al. (2008) that identified a gray matter volume enlargement in the right postcentral gyrus in high-functioning ASD children compared with matched controls.

We note that the results presented in Figures 3 and 4 and in Table 1 were obtained using all the subjects, and that these yield slightly different values from the classification rates presented in the results section. Indeed, the latter were calculated using a cross-validation method: within each iteration of the cross-validation, a different subject was left out, leading to 58 slightly different analyses. Nevertheless, across all the 58 iterations of the LOO-CV, the automatic region selection algorithm consistently selected the same two regions. This reproducibility throughout the re-sampling strongly confirms a stable characteristic pattern (PET biomarkers of autism) that comprises hypoperfusion (in the ASD group) in the right STS and hyperperfusion in the left postcentral (Figure 4).

Finally, concerning the regional feature extraction, we used a very simple supervised procedure based on the thresholding of an univariate statistical map. First, this raises the issue of choosing the threshold. Second, this procedure may be advantageously replaced by more sophisticated methods such as the one proposed in Fan et al. (2007).

6.3. Specific issues of imbalanced datasets

The imbalanced sizes of the two groups (45 children with ASD versus 13 non-ASD, low-functioning children) raises specific issues: the two discriminant-based approaches that we tried (linear SVM, Lasso logistic re-

gression) obtained poor predictions on low-functioning samples (low specificity). This issue could not be solved with re-weighting techniques as described in (4). Conversely, LDA is a fully generative model, based on Bayes rules that provide a formal framework to fix and control the priors of the two groups. In a simple experiment we loosened the control on the class priors, instead of that, we estimated them from the data. The high (97%) sensitivity and the low (30%) specificity highlight that is a crucial aspect to accommodate the imbalanced class design. The choice of such a generative model across the whole pipeline ensures that subjects were not classified on the basis of their proportion in the training set (45 vs. 13) but were identified solely on the basis of functional differences. Nevertheless, the high number of parameters estimated with the LDA leads to severe over-fitting in a high dimensional space (big P). This problem has motivated the development of a parsimonious feature selection through regional feature extraction.

Considering the comparison experiments that we presented, it is essential to understand that we did not blindly apply all those strategies on the data. Instead, we made just one initial experiment using a strategy based on a univariate F-test combined with a CV-based feature subset selection and a SVM classifier. The low specificity (38%, Table 2) motivated the development of the proposed strategy. Then, as a *post hoc* experiment, we were willing to evaluate the quality of the proposed strategy in comparison with other standard and state-of-the-art alternatives.

The chance levels of 50% that we used to tests for the significance instead of the proportion of each group. Testing against groups' proportions could produce some artificially significant specificities since the chance level is ar-

tificially low. Indeed a score of at least 53% of specificity (7/13) would be declared significant. On the other hand, it is true that such setting would require a very high sensitivity, above 88% (40/45) to be declared significant. Nevertheless, we applied such setting on all features selection/classification strategies presented in Table 2). Only the two last ones (#14 & #15) reach significant scores of both sensitivity and specificity. This confirms our conclusions on the superiority, on this dataset, of a methods based on: regional feature extraction + multivariate feature selection + adaptive model selection with a final generative classifier.

6.4. Limitations

It is difficult to investigate children with ASD because of the heterogeneity in intellectual efficiency associated with this disorder. Here we have specifically studied children with low functioning autism, which represents a large sub-group of children with ASD. In addition, most brain imaging studies have been performed in adults with high-functioning autism, so we believe it is important to investigate a more representative group of patients. However, these results need to be replicated in children and adults with ASD and normal intelligence performance.

7. Conclusion

In conclusion, this paper presents a feature selection algorithm that identifies a parsimonious pattern of regional features. We focus on a novel model selection procedure based on an adaptive penalized likelihood. This procedure outperformed the classical cross-validation while lessening computational burden. The identified pattern associated with a linear generative

classifier achieved an accurate (sensitive and specific) individual prediction of the clinical status despite an imbalanced training dataset. Moreover, we present an extensive comparison study with other state-of-the-art discriminant methods and demonstrate the superiority of the proposed generative algorithm.

We aimed to identify a shared pattern that discriminates all ASD subjects from controls. The STS is a critical part of this pattern, confirming previous multimodal brain imaging findings regarding STS in ASD. However, the multiple etiology of ASD and the numerous findings in neuroimaging studies suggest that several brain patterns may exist across the autistic spectrum. The next step would be to look for the multiple patterns that may be associated with the multiple etiologies.

References

Abell, F., Krams, M., Ashburner, J., Passingham, R., Friston, K., Frackowiak, R., Happé, F., Frith, C., Frith, U., Jun 1999. The neuroanatomy of autism: a voxel-based whole brain analysis of structural scans. Neuroreport 10 (8), 1647–1651.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19 (6), 716–723.

Akshoomoff, N., Lord, C., Lincoln, A. J., Courchesne, R. Y., Carper, R. A., Townsend, J., Courchesne, E., Mar 2004. Outcome classification of preschool children with autism spectrum disorders using mri brain measures. J Am Acad Child Adolesc Psychiatry 43 (3), 349–357.

Alexander, A. L., Lee, J. E., Lazar, M., Boudos, R., DuBray, M. B., Oakes, T. R., Miller, J. N., Lu, J., Jeong, E.-K., McMahon, W. M., Bigler, E. D., Lainhart, J. E., Jan 2007. Diffusion tensor imaging of the corpus callosum in autism. Neuroimage 34 (1), 61–73.
URL http://dx.doi.org/10.1016/j.neuroimage.2006.08.032

Ambroise, C., McLachlan, G. J., May 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci U S A 99 (10), 6562–6566.
URL http://dx.doi.org/10.1073/pnas.102102699

APA, 2000. Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR Text Revision. American Psychiatric Publishing, Inc.

Aylward, E. H., Minshew, N. J., Goldstein, G., Honeycutt, N. A., Augustine, A. M., Yates, K. O., Barta, P. E., Pearlson, G. D., Dec 1999. Mri volumes of amygdala and hippocampus in non-mentally retarded autistic adolescents and adults. Neurology 53 (9), 2145–2150.

Barnea-Goraly, N., Kwon, H., Menon, V., Eliez, S., Lotspeich, L., Reiss, A. L., Feb 2004. White matter structure in autism: preliminary evidence from diffusion tensor imaging. Biol Psychiatry 55 (3), 323–326.

Baron-Cohen, S., Ring, H. A., Bullmore, E. T., Wheelwright, S., Ashwin, C., Williams, S. C., May 2000. The amygdala theory of autism. Neurosci Biobehav Rev 24 (3), 355–364.

Birgé, L., Massart, P., 2007. Minimal penalties for gaussian model selection. Probability theory and related fields 138, 33–73.

Boddaert, N., Chabane, N., Belin, P., Bourgeois, M., Royer, V., Barthelemy, C., Mouren-Simeoni, M.-C., Philippe, A., Brunelle, F., Samson, Y., Zilbovicius, M., Nov 2004. Perception of complex sounds in autism: abnormal auditory cortical processing in children. Am J Psychiatry 161 (11), 2117–2120.
URL http://dx.doi.org/10.1176/appi.ajp.161.11.2117

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Burnham, K. P., Anderson, D. R., 2004. Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach. Springer-Verlag New York Inc.

Castelli, F., Frith, C., Happé, F., Frith, U., Aug 2002. Autism, asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. Brain 125 (Pt 8), 1839–1849.

Chang, C.-C., Lin, C.-J., 2001. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Chawla, N., Japkowicz, N., Kolcz, A., 2004. Editorial: Special issue on learning from imbalanced data sets. In: ACM SIGKDD Explorations.

Chawla, N. V., Bowyer, K. W., Moore, T. E., Kegelmeyer, P., 2002. Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357.

Courchesne, E., Yeung-Courchesne, R., Press, G. A., Hesselink, J. R., Jernigan, T. L., May 1988. Hypoplasia of cerebellar vermal lobules vi and vii in autism. N Engl J Med 318 (21), 1349–1354.

Dalton, K. M., Nacewicz, B. M., Johnstone, T., Schaefer, H. S., Gernsbacher, M. A., Goldsmith, H. H., Alexander, A. L., Davidson, R. J., Apr 2005. Gaze fixation and the neural circuitry of face processing in autism. Nat Neurosci 8 (4), 519–526.
URL http://dx.doi.org/10.1038/nn1421

Dapretto, M., Davies, M. S., Pfeifer, J. H., Scott, A. A., Sigman, M., Bookheimer, S. Y., Iacoboni, M., Jan 2006. Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. Nat Neurosci 9 (1), 28–30.
URL http://dx.doi.org/10.1038/nn1611

Dietterich, T. G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation 10, 1895–1923.

Duchesnay, E., Cachia, A., Roche, A., Rivière, D., Cointepas, Y., Papadopoulos-Orfanos, D., Zilbovicius, M., Martinot, J.-L., Régis, J., Mangin, J.-F., Apr 2007. Classification based on cortical folding patterns. IEEE Trans Med Imaging 26 (4), 553–565.
URL http://dx.doi.org/10.1109/TMI.2007.892501

Ecker, C., Marquand, A., Mouro-Miranda, J., Johnston, P., Daly, E. M., Brammer, M. J., Maltezos, S., Murphy, C. M., Robertson, D., Williams,

S. C., Murphy, D. G. M., Aug 2010a. Describing the brain in autism in five dimensions–magnetic resonance imaging-assisted diagnosis of autism spectrum disorder using a multiparameter classification approach. J Neurosci 30 (32), 10612–10623.
URL http://dx.doi.org/10.1523/JNEUROSCI.5413-09.2010

Ecker, C., Rocha-Rego, V., Johnston, P., Mourao-Miranda, J., Marquand, A., Daly, E. M., Brammer, M. J., Murphy, C., Murphy, D. G., Consortium, M. R. C. A., Jan 2010b. Investigating the predictive value of whole-brain structural mr scans in autism: a pattern classification approach. Neuroimage 49 (1), 44–56.
URL http://dx.doi.org/10.1016/j.neuroimage.2009.08.024

Egaas, B., Courchesne, E., Saitoh, O., Aug 1995. Reduced size of corpus callosum in autism. Arch Neurol 52 (8), 794–801.

Fan, Y., Shen, D., Gur, R. C., Gur, R. E., Davatzikos, C., Jan 2007. Compare: classification of morphological patterns using adaptive regional elements. IEEE Trans Med Imaging 26 (1), 93–105.
URL http://dx.doi.org/10.1109/TMI.2006.886812

Friedman, J. H., Hastie, T., Tibshirani, R., 2 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33 (1), 1–22.
URL http://www.jstatsoft.org/v33/i01

Friston, K. J., Ashburner, J., Frith, C. D., Poline, J.-B., Heather, J. D.,

Frackowiak, R. S. J., 1995. Spatial registration and normalization of images. Human Brain Mapping 3, 165–189.

Gervais, H., Belin, P., Boddaert, N., Leboyer, M., Coez, A., Sfaello, I., Barthélémy, C., Brunelle, F., Samson, Y., Zilbovicius, M., Aug 2004. Abnormal cortical voice processing in autism. Nat Neurosci 7 (8), 801–802. URL http://dx.doi.org/10.1038/nn1291

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. The Journal of Machine Learning Research 3, 1157–1182.

Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L., 2006. Feature extraction : foundations and applications. Springer-Verlag.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine Learning 46, 389–422.

Hadjikhani, N., Joseph, R. M., Snyder, J., Chabris, C. F., Clark, J., Steele, S., McGrath, L., Vangel, M., Aharon, I., Feczko, E., Harris, G. J., Tager-Flusberg, H., Jul 2004. Activation of the fusiform gyrus when individuals with autism spectrum disorder view faces. Neuroimage 22 (3), 1141–1150. URL http://dx.doi.org/10.1016/j.neuroimage.2004.03.025

Hadjikhani, N., Joseph, R. M., Snyder, J., Tager-Flusberg, H., Sep 2006. Anatomical differences in the mirror neuron system and social cognition network in autism. Cereb Cortex 16 (9), 1276–1282. URL http://dx.doi.org/10.1093/cercor/bhj069

Hand, D. J., Taylor, C., 1987. Multivariate Analysis of Variance and Repeated Measures: A Practical Approach for Behavioural Scientists. Chapman & Hall.

Haznedar, M. M., Buchsbaum, M. S., Metzger, M., Solimando, A., Spiegel-Cohen, J., Hollander, E., Aug 1997. Anterior cingulate gyrus volume and glucose metabolism in autistic disorder. Am J Psychiatry 154 (8), 1047–1050.

Howard, M. A., Cowell, P. E., Boucher, J., Broks, P., Mayes, A., Farrant, A., Roberts, N., Sep 2000. Convergent neuroanatomical and behavioural evidence of an amygdala hypothesis of autism. Neuroreport 11 (13), 2931–2935.

Japkowicz, N., Stephen, S., October 2002. The class imbalance problem: A systematic study. Intell. Data Anal. 6, 429–449.
URL http://portal.acm.org/citation.cfm?id=1293951.1293954

Ke, X., Hong, S., Tang, T., Zou, B., Li, H., Hang, Y., Zhou, Z., Ruan, Z., Lu, Z., Tao, G., Liu, Y., Jun 2008. Voxel-based morphometry study on brain structure in children with high-functioning autism. Neuroreport 19 (9), 921–925.
URL http://dx.doi.org/10.1097/WNR.0b013e328300edf3

Keller, T. A., Kana, R. K., Just, M. A., Jan 2007. A developmental study of the structural integrity of white matter in autism. Neuroreport 18 (1), 23–27.
URL http://dx.doi.org/10.1097/01.wnr.0000239965.21685.99

Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., Fox, N. C., Jack, C. R., Ashburner, J., Frackowiak, R. S. J., Mar 2008. Automatic classification of mr scans in alzheimer's disease. Brain 131 (Pt 3), 681–689.
URL http://dx.doi.org/10.1093/brain/awm319

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International Joint Conference on Artificial Intelligence. Morgan Kaufmann, pp. 1137–1143.

Kohavi, R., John, G., 1997. Wrappers for feature selection. Artificial Intelligence 97, 273–324.

Lord, C., Leventhal, B. L., Cook, E. H., Jan 2001. Quantifying the phenotype in autism spectrum disorders. Am J Med Genet 105 (1), 36–38.

Mason, S. J., Graham, N. E., 2002. Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: Statistical significance and interpretation. Quarterly Journal of the Royal Meteorological Society 128, 2145–2166.

McAlonan, G. M., Cheung, V., Cheung, C., Suckling, J., Lam, G. Y., Tai, K. S., Yip, L., Murphy, D. G. M., Chua, S. E., Feb 2005. Mapping the brain in autism. a voxel-based mri study of volumetric differences and intercorrelations in autism. Brain 128 (Pt 2), 268–276.
URL http://dx.doi.org/10.1093/brain/awh332

Müller, R.-A., Cauich, C., Rubio, M. A., Mizuno, A., Courchesne, E., Sep 2004. Abnormal activity patterns in premotor cortex during sequence

learning in autistic patients. Biol Psychiatry 56 (5), 323–332.
URL http://dx.doi.org/10.1016/j.biopsych.2004.06.007

Neeley, E. S., Bigler, E. D., Krasny, L., Ozonoff, S., McMahon, W., Lainhart, J. E., Aug 2007. Quantitative temporal lobe differences: autism distinguished from controls using classification and regression tree analysis. Brain Dev 29 (7), 389–399.
URL http://dx.doi.org/10.1016/j.braindev.2006.11.006

Ohnishi, T., Matsuda, H., Hashimoto, T., Kunihiro, T., Nishikawa, M., Uema, T., Sasaki, M., Sep 2000. Abnormal regional cerebral blood flow in childhood autism. Brain 123 (Pt 9), 1838–1844.

Osuna, E. E., Freund, R., Girosi, F., 1997. Support vector machines: Training and applications. Tech. rep., MIT.

Pelphrey, K. A., Morris, J. P., McCarthy, G., May 2005. Neural basis of eye gaze processing deficits in autism. Brain 128 (Pt 5), 1038–1048.
URL http://dx.doi.org/10.1093/brain/awh404

Pereira, F., Mitchell, T., Botvinick, M., Mar 2009. Machine learning classifiers and fmri: a tutorial overview. Neuroimage 45 (1 Suppl), S199–S209.
URL http://dx.doi.org/10.1016/j.neuroimage.2008.11.007

Pierce, K., Haist, F., Sedaghat, F., Courchesne, E., Dec 2004. The brain response to personally familiar faces in autism: findings of fusiform activity and beyond. Brain 127 (Pt 12), 2703–2716.
URL http://dx.doi.org/10.1093/brain/awh289

Pudil, P., Novovicov'a, J., Kittle, J., 1994. Floating search methods in feature selection. Pattern Recogn Lett. 15, 1119–1125.

Raskutti, B., Kowalczyk, A., 2004. Extreme re-balancing for svms: a case study. In: ACM SIGKDD Explorations Newsletter - Special issue on learning from imbalanced datasets.

Reunanen, J., 2003. Overfitting in making comparisons between variable selection methods. The Journal of Machine Learning Research 3, 1371–1382.

Schultz, R. T., 2005. Developmental deficits in social perception in autism: the role of the amygdala and fusiform face area. Int J Dev Neurosci 23 (2-3), 125–141.
URL http://dx.doi.org/10.1016/j.ijdevneu.2004.12.012

Schwarz, G., 1978. Estimating dimension of a model. Ann. Statist. 6, 461–464.

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J.-B., Lebihan, D., Dehaene, S., Dec 2006. Inverse retinotopy: inferring the visual content of images from brain activation patterns. Neuroimage 33 (4), 1104–1116.
URL http://dx.doi.org/10.1016/j.neuroimage.2006.06.062

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B. 58, 267–288.

Veropoulos, K., Campbell, C., Cristianini, N., 1999. Controlling the sensitivity of support vector machines. In: Proceedings of the International Joint Conference on AI. pp. 55–6.

Zhang, J., Mani, I., 2003. knn approach to unbalanced data distributions: A case study involving information extraction. In: In Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Datasets.

Zilbovicius, M., Boddaert, N., Belin, P., Poline, J. B., Remy, P., Mangin, J. F., Thivard, L., Barthélémy, C., Samson, Y., Dec 2000. Temporal lobe dysfunction in childhood autism: a pet study. positron emission tomography. Am J Psychiatry 157 (12), 1988–1993.

Zilbovicius, M., Garreau, B., Tzourio, N., Mazoyer, B., Bruck, B., Martinot, J. L., Raynaud, C., Samson, Y., Syrota, A., Lelord, G., Jul 1992. Regional cerebral blood flow in childhood autism: a spect study. Am J Psychiatry 149 (7), 924–930.

Acknowledgments:

We thank the staff at NeuroSpin, especially Denis Rivire and Alexis Roche, for their technical and scientific assistance. This work was supported by INSERM, CEA, and grants from the National Agency for Research (PSY-MARKER and AGIR), France Foundation and Orange Foundation. Arnaud Cachia was supported by Fondation Deniker. Monica Zilbovicius had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Tables and figures captions

Table 1.

Regions identified during the feature selection process. In bold, the two regions finally selected for the classification. We provide the p-values ($df = 56$), uncorrected for multiple comparisons, of the maximum of each region. Positive t-values indicate hypoperfusion of the ASD group. We also provide the regions size in voxels (voxels are 2 mm isotropic) and the location of the maximum in MNI coordinates.

Table 2.

Summary of a comparison with the alternatives feature selection and classification strategies. They are organized by (i) the type of the input features (Feat.) which can be voxels (Vox.) or regional (Reg.) features; (ii) the method of feature subset ranking (Rank.); (iii) the method of feature subset selection (Sel.); and (iv) the classifier (Clf). Precisions: SVM are linear; LLR: Lasso Logistic Regression; CV: ten-fold Cross-Validation; RFE: linear SVM or LDA Recursive Feature Elimination depending on the final classifier. Concerning the performances, we provide the correct prediction rates (%) of accuracy (Acc.) sensitivity (Sensi.) and specificity (Speci.). We also provide the Area under Curve (AUC), and its significance was evaluated with a Wilcoxon test of ranks. Significances are reported as: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. Finally we reported whether (Yes or No) the proposed strategy (reported on the last line) significantly outperformed others using a McNemar's test of accuracy (Acc.), sensitivity (Sensi.) and

specificity (Speci.).

Figure 1.

Overview of the proposed methods: input parameters (black ovals), estimated parameters (dashed oval) and output results (thick contour ovals). The three input parameters are the univariate threshold and the two priors. (a) The left column describes the calibration of the adaptive penalization based on a randomization of training samples. The calibration yield to the estimation of the penalization value used by the model selection step. (b) The middle column describes the leave-one-out cross-validation loop of the multi-stage feature selection and classification algorithm. The estimated parameters (that were not reported on the figure) are: (i) the groups means and the pooled variances over all voxels, for the first-stage (Section 2.1); (ii) the multivariate group means μ_1^k, μ_2^k and the pooled covariance matrix Σ^k over the k regional features for the three other stages (Sections 2.2, 2.3 and 2.4) As a results, on the bottom of this column, we obtain the classification rates (accuracy, sensitivity and specificity) described in Section 3.1. (c) Finally, the right column, describes the assessment of significance of the cross-validated classification rates based on random permutation of the group label.

Figure 2.

The multi-stages feature selection algorithm has been repeated on randomly permuted datasets. We can observe the increase of the log-likelihood (2) for a varying number of regional features (k). We reported the theoretical log-evidence (baseline), which is supposed to be constant and equal to $\ln(1/2)^N$.

We also reported the penalizations obtained with the BIC and AIC criteria. This experiment shows that those fixed penalty criteria lead to a severe under-penalization of the log-likelihood. However, it also demonstrates that a good linear approximation can be obtained leading to an adaptive penalization criterion noted (aPen) with a penalty term of $2.67\ \frac{1}{2}\ k \ln N$ as noted in (1).

Figure 3.

Regions selected by the three steps of feature selection, illustrated on a single subject and displayed consistent with neurological convention. (1) voxel-based two samples (ASD versus low-functioning) t-statistics. (1.2) The thresholding led to six characteristic regions: Four of those regions featured hypoperfusion in the ASD group: (i) the right temporo-parietal junction (RTPJ); (ii) the right Superior Temporal Sulcus (STS); (iii) middle temporal gyrus; and (iv) the posterior zone of the corpus callosum where it overlaps with the right posterior cingulum and bilateral thalami. Two hyperperfused regions in the ASD group were identified in (v) the left post-central and (vi) the right pre-central areas. The network of abnormalities formed by those six regions is represented with links that code for significant correlations between pairs of regions. The radius of each link is proportional to the absolute value of the correlation; purple and orange respectively indicate negative and positive correlations. The second (2) and third (3) steps led to our selection of two regions: (ii) the right STS and (v) the left postcentral.

Figure 4.

The characteristic pattern is composed of two regions: (i) a hypoperfusion (in the ASD group, blue dots) in the right STS and (ii) a hyperperfusion in the contralateral postcentral area. We added the density plots from these two regions. The top-right corner plot is a combined density of the two regions that was obtained from a projection of the most discriminant axis as identified through LDA. This combination of regions clearly shows reduced overlap and a better separation between the two populations.

Figure 5.

ROC curve of alternative popular classification strategies. They are organized by (i) the type of the input features which can be voxels (Vox.) or regional (Reg.) features; (ii) the method of feature subset ranking (Ranking); (iii) the method of feature subset selection (Selection); and (iv) the classifier. Precisions: SVM are linear; LLR: Lasso Logistic Regression; CV: ten-fold Cross-Validation; RFE: linear SVM or LDA Recursive Feature Elimination depending on the final classifier. Area under Curve (AUC) significances, evaluated with a Wilcoxon test of ranks, were reported as: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. It clearly shows an improvement of the sensitivity vs. specificity trade-off while moving toward a strategy that combines regional feature extraction, multivariate feature subsets ranking and calibrated (aPena) model selection.

Tables

#	Region name	p-val.	t-val.	#vox.	x	y	z
i	Right temporo-parietal junction	5.3×10^{-5}	4.17	271	56	-50	30
ii	**Right superior temporal sulcus**	1.0×10^{-4}	3.96	299	58	-10	-6
iii	Right middle temporal gyrus	5.3×10^{-5}	4.17	236	42	-40	-6
iv	Right post. part of corpus callosum	2.0×10^{-4}	3.77	90	2	-32	10
v	**Left post-central**	8.0×10^{-6}	-4.71	401	-38	-36	66
vi	Right pre-central	3.0×10^{-4}	-3.58	26	42	-28	68

Table 1:

	Strategy				Performance				Comparison		
#	Feat.	Rank.	Sel.	Clf	Acc.	Sensi.	Speci.	AUC	Acc.	Sensi.	Speci.
1	Vox.	no	all	SVM	77***	93***	23	0.64	N	N	**Y***
2	Vox.	t-test	CV	SVM	75***	86***	38	0.63	N	N	**Y***
3	Vox.	RFE	CV	SVM	65*	80***	15	0.5	**Y****	N	**Y***
4	Vox.	Lasso	CV	LLR	68**	82***	23	0.48	**Y****	N	**Y***
5	Reg.	no	all	SVM	65*	73**	38	0.43	**Y****	**Y***	N
6	Reg.	t-test	CV	SVM	63*	71**	38	0.58	**Y****	**Y***	N
7	Reg.	RFE	CV	SVM	74***	86***	30	0.47	**Y***	N	**Y***
8	Reg.	Lasso	CV	LLR	65*	73**	38	0.5	**Y****	**Y***	N
9	Reg.	no	all	LDA	74***	84***	38	0.58	**Y***	N	N
10	Reg.	t-test	CV	LDA	63*	73**	30	0.52	**Y****	**Y***	**Y***
11	Reg.	RFE	CV	LDA	75***	82***	53	**0.65***	**Y***	N	N
12	Reg.	SFFS	CV	LDA	81***	91***	46	0.61	N	N	N
13	Reg.	Lasso	aPena	LLR	74***	75***	69	**0.81*****	**Y***	**Y***	N
14	Reg.	RFE	aPena	LDA	84***	88***	69	**0.74****	N	N	N
15	Reg.	SFFS	aPena	LDA	87***	91***	**77***	**0.81*****	-	-	-

Table 2:

Figures

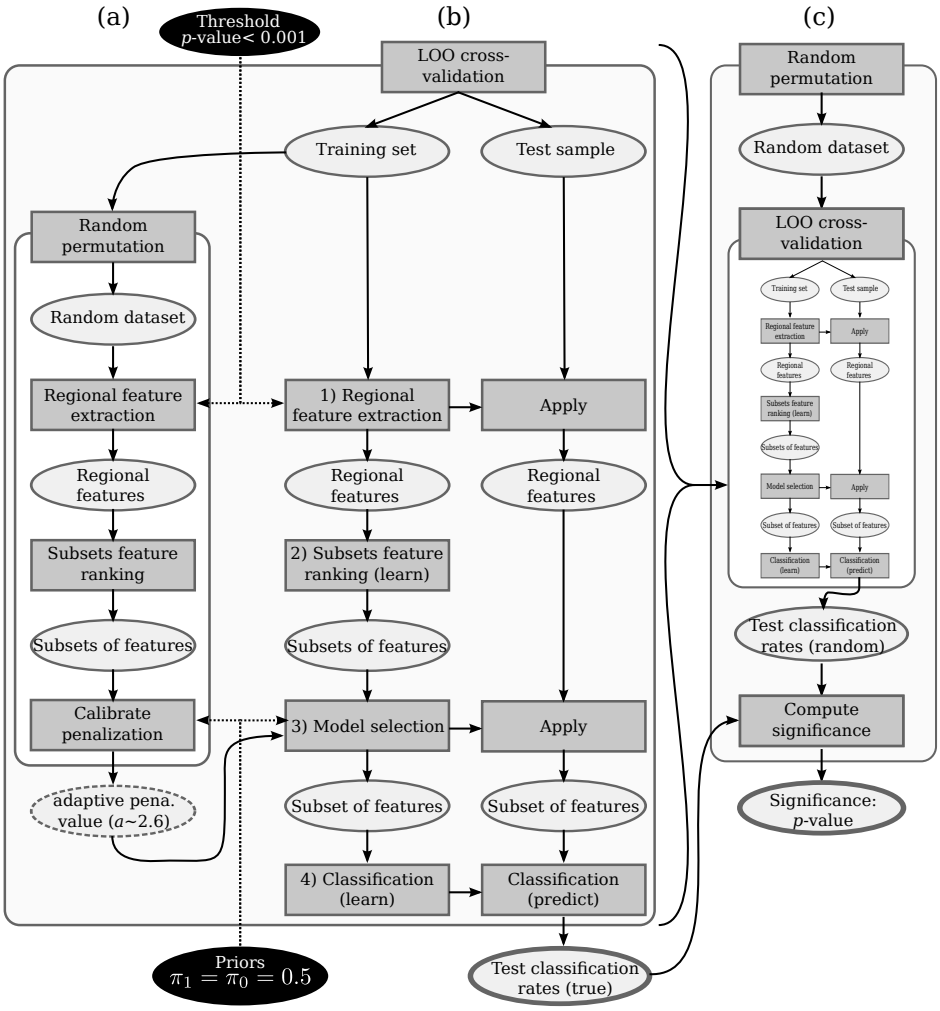

Figure 1:

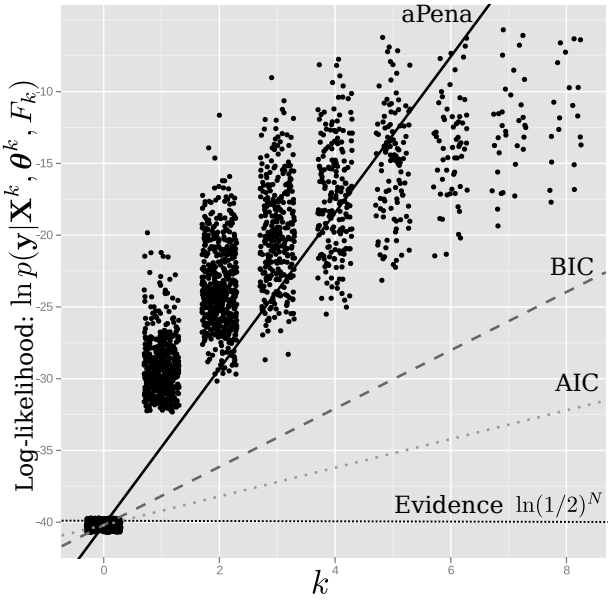
aPena
BIC
AIC
Evidence $\ln(1/2)^N$
Log-likelihood: $\ln p(\mathbf{y}|\mathbf{X}^k, \boldsymbol{\theta}^k, F_k)$
k
-10
-15
-20
-25
-30
-35
-40
0
2
4
6
8

Figure 2:

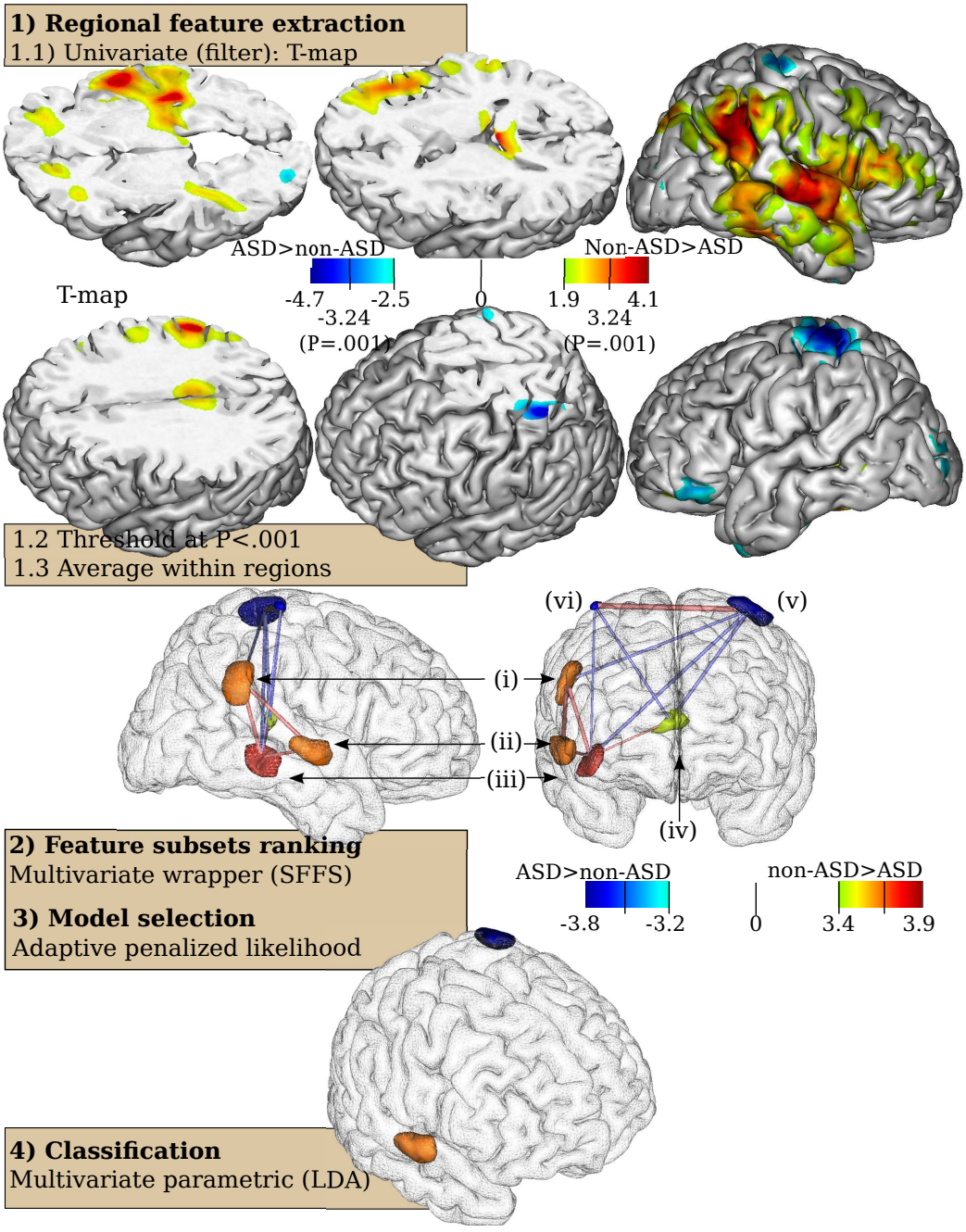
1) Regional feature extraction
1.1) Univariate (filter): T-map
ASD>non-ASD
Non-ASD>ASD
T-map
-4.7
-2.5
0
1.9
4.1
-3.24
(P=.001)
3.24
(P=.001)
1.2 Threshold at P<.001
1.3 Average within regions
(vi)
(v)
(i)
(ii)
(iii)
(iv)
2) Feature subsets ranking
Multivariate wrapper (SFFS)
3) Model selection
Adaptive penalized likelihood
ASD>non-ASD
non-ASD>ASD
-3.8
-3.2
0
3.4
3.9
4) Classification
Multivariate parametric (LDA)

Figure 3:

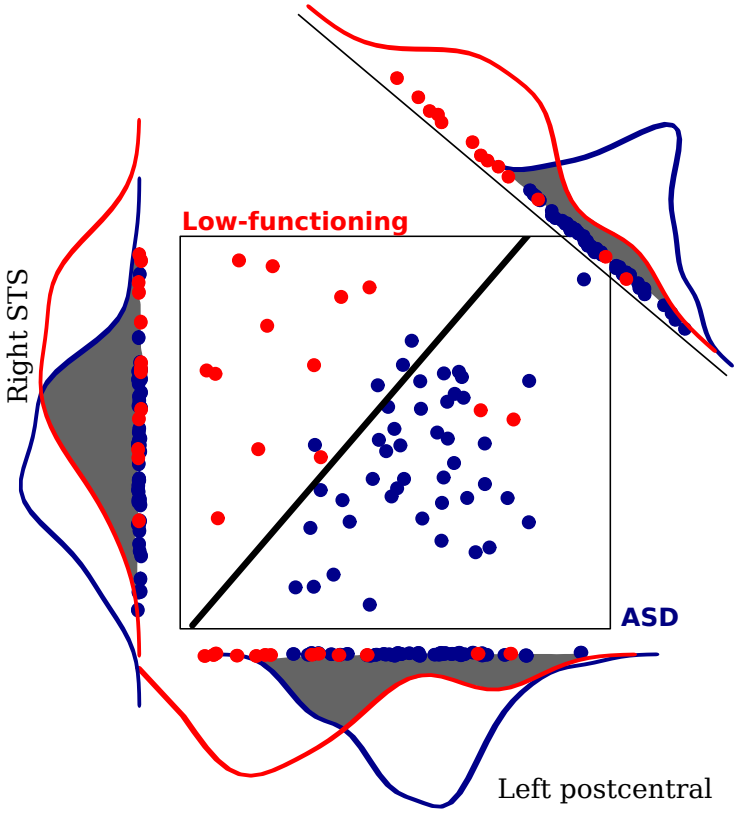

Figure 4:

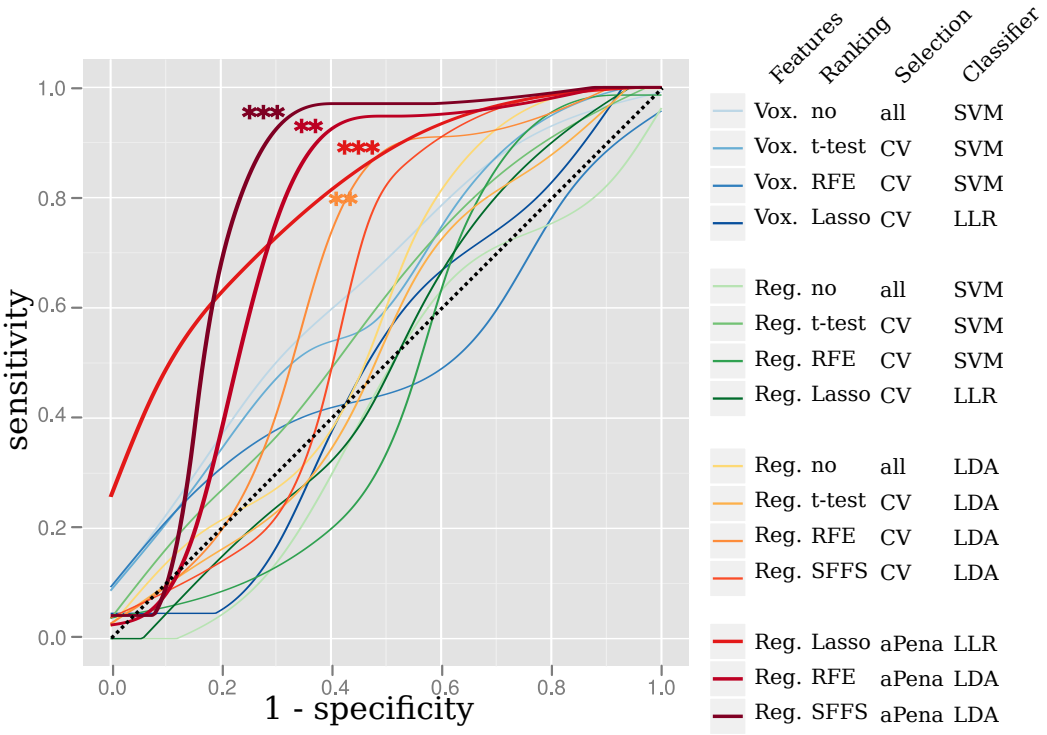
sensitivity
1 - specificity
0.0
0.2
0.4
0.6
0.8
1.0
***
**
***
**
Features
Ranking
Selection
Classifier
Vox. no all SVM
Vox. t-test CV SVM
Vox. RFE CV SVM
Vox. Lasso CV LLR
Reg. no all SVM
Reg. t-test CV SVM
Reg. RFE CV SVM
Reg. Lasso CV LLR
Reg. no all LDA
Reg. t-test CV LDA
Reg. RFE CV LDA
Reg. SFFS CV LDA
Reg. Lasso aPena LLR
Reg. RFE aPena LDA
Reg. SFFS aPena LDA

Figure 5: