Fast Gaussian Naïve Bayes for searchlight classification analysis

NeuroImage 163 (2017) 471–479

Contents lists available at ScienceDirect

NeuroImage

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

Fast Gaussian Naïve Bayes for searchlight classification analysis

Marlis Ontivero-Ortega [a], Agustin Lage-Castellanos [a,c], Giancarlo Valente [c], Rainer Goebel [c], Mitchell Valdes-Sosa [b,*]

[a] Department of NeuroInformatics, Cuban Center for Neuroscience, Cuba
[b] Department of Cognitive Neuroscience Cuban, Center for Neuroscience, Cuba
[c] Department of Cognitive Neuroscience, Maastricht University, Netherlands

ARTICLE INFO

Keywords:
Searchlight MVPA
Gaussian Naïve Bayes
Support vector machine
Permutation tests

ABSTRACT

The searchlight technique is a variant of multivariate pattern analysis (MVPA) that examines neural activity across large sets of small regions, exhaustively covering the whole brain. This usually involves application of classifier algorithms across all searchlights, which entails large computational costs especially when testing the statistical significance of the accuracies with permutation methods. In this article, a new implementation of the Gaussian Naive Bayes classifier is presented (henceforth massive-GNB). This approach allows classification in all searchlights simultaneously, and is faster than previously published searchlight GNB implementations, as well as other more complex classifiers including support vector machines (SVM). To ensure that the gain in speed for GNB would be useful in searchlight analysis, we compared the accuracies of massive-GNB and SVM in detecting the lateral occipital complex (LOC) in an fMRI localizer experiment (26 subjects). Moreover, this region as defined in a meta-analysis of many activation studies was used as a gold standard to compare error rates for both classifiers. In individual searchlights, SVM was somewhat more accurate than massive-GNB and more selective in detecting the meta-analytic LOC. However, with multiple comparison correction at the cluster-level the two classifiers performed equivalently. Thus for cluster-level analysis, massive-GNB produces an accuracy similar to more sophisticated classifiers but with a substantial gain in speed. Massive-GNB (available as a public Matlab toolbox) could facilitate the more widespread use of searchlight analysis.

1. Introduction

Multivariate pattern analysis (MVPA) is increasingly applied to functional magnetic resonance imaging (fMRI) data (Haynes, 2015). Although all voxels in the brain can be used in MVPA (e.g. Valente et al., 2014), it is frequent to ask if informative patterns are present in smaller and more localized regions. One approach for this is to perform MVPA in predefined regions of interest (ROIs), delineated either by anatomical (e.g. a certain gyrus or sulcus), or by functional criteria (i.e. clusters of activated voxels in an independent localizer task) (Saxe et al., 2006). Unfortunately, one does not always have clearly predefined criteria for ROIs in all situations -especially in new cognitive paradigms- nor can one always obtain all the potentially interesting localizers in a single subject (due to time constraints or simply the absence of a functional ROI in some subjects) (Spiridon et al. 2006).

Searchlight MVPA (Kriegeskorte et al., 2006) allows testing for localized informative patches in the brain without demanding the "a priori" knowledge mandatory for traditional ROI definition. This method usually consists in training, and testing, a classification algorithm within a small region that is displaced around the brain. These regions can be spheres (Kriegeskorte et al., 2006) in the brain volume or disks over the cortical surface (Oosterhof et al., 2011), which effectively comprise small, overlapping ROIs covering the brain exhaustively. This helps overcome the curse of dimensionality (Bishop, 2006) by reducing the number of features included in each classification problem.

Despite its appeal, searchlight MVPA presents a number of challenges (see reviews by Jimura and Poldrack, 2012, and Etzel et al. 2013), including the problem of the number of searchlights. This number can be very large especially when high-resolution fMRI data is collected (e.g. more than 300 000 in the volume for 7 T fMRI). This entails lengthy computations when classification algorithms are applied in a sequential loop over all searchlights. Furthermore, cross-validation is necessary to avoid overfitting (usually repeatedly splitting data into train and test sets), which implies a further computational burden (Pereira et al., 2009).

Extra costs arise from when assessing if classification accuracy across

[*] Corresponding author.
E-mail address: mitchell@cneuro.edu.cu (M. Valdes-Sosa).

http://dx.doi.org/10.1016/j.neuroimage.2017.09.001
Received 22 May 2017; Received in revised form August 2017; Accepted 1 September 2017
Available online 4 September 2017

searchlights is significantly above chance. At first glance, a binomial or multinomial test would be sufficient. However, as discussed in several articles (Pereira and Botvinick, 2011; Stelzer et al., 2013; Noirhomme et al., 2014; Jamalabadi et al., 2016) the distribution of MVPA accuracies are usually ill-behaved (i.e. in binary choices the results do not follow a binomial distribution, and sometimes the chance success rate is off 50%). The preferred solution is to obtain an empirical null distribution of accuracies by repeatedly applying the classifier after permuting the labels with respect to the trials. Thus many computations of the classifier in each searchlight are needed (usually around 10^3). But for more precise p value estimation, a larger number of permutations are desirable (e.g. $>10^5$) (Ojala and Garriga, 2010). These permutation tests were originally conceived only for within-subject analyses. Nevertheless, recent proposals incorporate the results of intra-individual permutations into second level (group) tests. This was motivated by inadequacies of traditional methods (e.g. Student t tests on accuracies in a group), the assumptions of which are usually not met in MVPA classification studies (see Allefeld and Haynes, 2014). Examples of this new approach are a fixed-effect analysis on group median accuracy (Stelzer et al., 2013), and more recently a random effect analysis for the prevalence of subjects carrying information (Allefeld et al., 2016).

All these considerations mandate fast implementation of classifiers to overcome this large computational burden, especially when the analyses are performed in personal PCs outside of computer grids. An ideal algorithm for rapid searchlight calculations is the Gaussian Naive Bayes (GNB) classifier (Bishop, 2006), which is several orders of magnitude faster than the popular Support Vector Machine (SVM) or Logistic Regression classifiers. In GNB one assumes a diagonal covariance matrix between features. This simplistic assumption is especially useful in high dimensional scenarios, since it avoids the estimation of a full covariance matrix. The estimation of the covariance matrix is problematic when the number of samples is smaller than the number of features (a frequent situation in fMRI experiments). Additionally, under this assumption, the contribution of each voxel to the classification function is always the same regardless of the different searchlights to which it belongs. This means that the parameters for each voxel are estimated only once. A further simplification is assume equal variance across different classes.

GNB was originally introduced for fast searchlight MVPA by Pereira and Botvinick (2011) in their Searchmight toolbox, with a C code implementation. Another speeded-up GNB implementation is part of the CoSMoMVPA toolbox (Oosterhof et al., 2016). Here, a new computational (algorithmic) framework for the acceleration of GNB is presented in which this classifier can be trained and evaluated at all searchlights simultaneously in a few seconds. The computational framework is based on the sparse relationship between searchlights and the space of voxels. This allows summation of voxel contributions within each searchlight, simultaneously across the brain, with sparse matrix multiplications. Hence, it is possible to circumvent sequential calculations over the set of searchlights. To gauge the gain in speed produced by massive-GNB, its computational time was compared with the time of the GNB in both the Searchmight and the CoSMoMVPA toolbox. Additionally, the combination of this new implementation with simple hardware-based parallelization was examined.

Greater speed would not be useful if the GNB classifier has poor performance in MVPA, which is a concern some studies have raised. If the data covariance matrix deviates from the diagonal or the variance is not equivalent across classes, then the classification hyper-surface estimated with GNB might separate the different classes poorly, with a loss in power for detecting information (Krzanowski, 1988). This possibility is consistent with reports asserting that GNB is less sensitive than other classifiers such as SVM (Ku et al., 2008; Misaki et al., 2010), even though other studies find a similar performance in this comparison (Wang et al., 2004; Pereira and Botvinick, 2011). Moreover, Raizada and Lee (2013) argue that the informative patches found by GNB are smoother, and more reproducible across different fMRI datasets, than those found by SVM. Thus, despite some doubts about its performance, speeding up the GNB for searchlight analysis could be of real practical importance.

Previous evaluations of classifiers for MVPA have only compared their relative accuracies under the assumption that larger values indicate better performance, which means in the searchlight context that larger accuracies in more units is better. But this ignores the possibility of false positive searchlights. It would be better to rate different classifiers against a "ground truth". Here GNB performance was compared with that of the SVM classifier in fMRI data from a lateral occipital complex (LOC) localizer experiment. The cortical patches identified as informative in the GNB and SVM searchlight analyses were compared with each other, but also with the LOC defined by a meta-analysis of a large number of univariate activation studies. This meta-analytic LOC was taken as "ground truth". In the real fMRI data, the simplicity of the GNB did not thwart detection of the same informative clusters found with SVM and both classifiers performed equivalently in reference to the meta-analysis.

2. Methods

2.1. Sparse representation of searchlights structures

A searchlight is defined by a central voxel and a set of voxels in its neighborhood. The neighborhood structure of all searchlights can be encoded in a sparse binary matrix S, whose size is number of voxels v x number of searchlights s. Usually (but not necessarily) the magnitude of v and s are equal. The S matrix has a non-zero entry at S_{ij} if the voxel i is included in searchlight j, and zero otherwise. A toy matrix formed by 4 voxels and 4 searchlights is presented in equation (1). For example, searchlight 3 is formed by voxels 3 and 4:

$$S_{v,s} = \begin{array}{c} \begin{array}{cccc} s_1 & s_2 & s_3 & s_4 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}_{v,s} \end{array} \qquad [1]$$

A whole brain MVPA analysis for high resolution fMRI data is typically in the order of 10^5 searchlights, which leads to a $10^5 \times 10^5$ sparse matrix S in which less than 0.5% of the elements are non-zero. The level of sparseness depends on the number of neighbors within each searchlight, which is usually less than a few hundreds of voxels. This is more than three orders of magnitude smaller that the number of searchlights to compute (i.e. 10^5).

The data in a searchlight classification problem consists of a matrix of features which is split in training (Tr) data $X^{Tr}_{n,v}$ (where n is number of training trials, and v number of features or voxels) and test (Te) data $X^{Te}_{m,v}$ (where m is the number of test trials), as well as the corresponding vector of labels $y^{Tr}_{n,1}, y^{Te}_{m,1}$. The most common features used in searchlight analysis are voxel-wise beta or t values (see Misaki et al., 2010 for a comparison of these two measures) obtained from a first level linear model, although the maximal or average amplitude of the fMRI signal within a time window posterior to the stimulus onset can be used as well.

2.2. Massive Gaussian Naïve Bayes (massive-GNB)

GNB is one of the simplest classification algorithms (Bishop, 2006). It consists in assigning the label of the class that maximizes the posterior probability of each sample, under the assumption that the voxel contributions are conditionally independent and obey a Gaussian distribution. The GNB decision rule is written in terms of the discriminant function for each class k at each searchlight s (the searchlight index is omitted in the next equation to avoid visual clutter). The discriminant function is defined as the sum of the squared distances to the centroid of each class, across all voxels in the searchlight, weighted by the variance, and the logrithm of the a-priori probability (p_k) computed in the training set, according to the Bayes rule (see equation (2)). The predicted class for sample i in the test set is assigned by selecting the label of the class having the discriminant functionδ_i^k with the largest value, which implies maximal posterior probability, within discriminant functions of all classes:

$$\hat{c}_i = \text{arg max}_{k=\{a,b\}}\left\{\delta_i^k\right\}$$

$$\delta_i^k = -\sum_{j=1}^{v}\left(\frac{\left(x_{ij}^{Te} - \hat{\mu}_j^k\right)^2}{2\hat{\sigma}_j^2}\right) + \log(p_k) \qquad [2]$$

This equation for a binary classification problem with classes $\{a, b\}$ can be easily generalized for multiclass classification problems. The mean and standard deviation $\{\hat{\mu}_j^k, \hat{\sigma}_j\}$ in each voxel j are computed over the training set. When GNB is generalized to multiple searchlights the former equation is written in one matrix product, using the sparse binary matrix S that selects and sums the voxels contributions within each searchlight in one operation:

$$D_{m,s}^k = -F_{m,v}^k S_{v,s} + \log(p_k) \qquad [3]$$

The elements of the matrix$F^k$are the voxel contributions to the discriminant function (i.e. squared z-score distance) for each sample i in the test set:$f_j^k = \frac{(x_{ij}^{Te} - \hat{\mu}_j^k)^2}{2\hat{\sigma}_j^2}$, which corresponds to the voxel-wise contributions of equation (2). Note that the contribution of each voxel is always the same for all the searchlights it belongs to. The matrix D^k contains the discriminant functions for the class k, and for each sample at each searchlight. The massive-GNB method is numerically stable and produces exactly the same results like as running the MATLAB classify function sequentially (see Text S1 and Fig. S1 in Supplementary Information (SI) for more details).

The GNB classifier is identical to a Linear Discriminant Analysis (LDA) in which a diagonal covariance matrix between variables (voxels) and identical across-class standard deviations are assumed. The assumption of different standard deviations between classes, corresponds to a Quadratic Discriminant Analysis (QDA), again under the assumption of a diagonal covariance matrix. The computation of QDA in our massive computational framework is straightforward, producing a discriminant function analogous to the linear case.

2.3. *The GNB in the Searchmight and CoSMoMVPA toolboxes*

The Searchmight toolbox (Pereira and Botvinick, 2011) contains a fast GNB implementation that uses mex compiled C code that can be called from Matlab (http://www.princeton.edu/~fpereira/Searchmight/). This variant of GNB has been recognized as one of the fastest options for MVPA analysis (Hebart et al., 2015). This function (SearchmightGNB.c) implements internally the cross-validation loop and the permutation test loop, which notably reduces computation times compared with the Matlab classify function. Nevertheless, the searchmight-GNB algorithm computes the GNB classifier serially across the searchlights. Likewise, the CoSMoMVPA toolbox has a rapid implementation of the GNB classifier (cosmo-GNB) in which the cross-validation loop is also computed internally (Oosterhof et al., 2016). This classifier is not numerically identical to the massive-GNB, neither to searchmight-GNB, since it assumes different variance across classes.

2.4. *Run time analysis*

The computation times of the massive-GNB were compared with those of searchmight-GNB and cosmo-GNB in high-resolution fMRI data (voxel size = 1.1 × 1.1 × 1.1 mm), that was obtained in an ultrahigh-field 7 T scanner (case 7 from Emmerling et al., 2016). On different trials the subject was requested to imagine dots moving in one of four directions (0°, 90°, 180°, and 270°), although only two directions were used here. The time taken to compute classifiers discriminating these two directions was measured. Three parameters were varied in separate analyses: 1) the number of volumetric searchlights, 2) the number of neighbors within each searchlight (as a function of its radius), and 3) the number of samples (trials) included in the training and in the test set. The number of searchlights varied from 100 up to 234067 (the total number of voxels, as in a whole brain analysis). The searchlights were obtained using the function computeNeighboursWithinRadius from the Searchmight toolbox. The radius (excluding the central voxel) was varied from 1 to 8 voxels, which produced searchlights of sizes that varied from 27 to 4913 neighbors. To explore the effect of sample size, the beta values for 100 trials were taken from the original fMRI dataset. This sample was expanded in 10 trial steps by recurrent inclusion of random samples from the original data set (to which standard normal white noise was added). This was repeated until 190 trials were reached. When one parameter was varied, the other two parameters were kept constant (at either 234067 searchlights, a radius of 2 voxels −125 neighbors-, or 100 samples). The computational times include all 5-fold cross-validations.

The gains in speed achieved by the massive-GNB algorithm can be boosted if it is combined with hardware parallelization. Hardware based parallelization is an art in itself, and significant progress have been made in its application in neuroscience (Eklund et al., 2014). Here a simple procedure was used to test acceleration of permutation tests based on massive-GNB. The Matlab parfor instruction with a computer grid was applied to send permutations across CPUs (workers), even though the cross-validations were computed serially in each worker. The duration of this computation was compared with the case in which both the permutations and cross-validations were performed sequentially. This analysis was performed for massive-GNB and Searchmight-GNB using 23407 searchlights (178-voxel in each), 100 samples (50 for each class) and 10^3 permutations. These, and all other, calculations in this article were performed on a grid containing 12 CPU cores (Intel(R) Xeon(R) E5-2670 v3) each with 2.30 GHz Clock Speed and 48 Gb RAM.

2.5. *Comparison of massive-GNB and SVM in real fMRI data*

Massive-GNB and SVM classifiers were compared in data from a standard localizer task for the LOC region (an area related to visual object recognition, Grill-Spector et al., 2001), since this paradigm yields easily reproducible findings. The data was obtained as part of a larger study, currently underway as a collaboration between the Cuban Center for Neuroscience and the University of Electronic Science and Technology of China (UESTC). LOC is usually identified as the voxels in which larger BOLD activations are produced for intact objects than for scrambled objects. This analysis is performed on spatially smoothed data, and evinces a cortical activation that is larger than usual searchlights sizes. However, differences in activation patterns have been found between different classes of objects in these same regions with unsmoothed data, even when using correlation-based MVPA that eliminates the effects of mean activations (e.g. Golomb and Kanwisher, 2011; MacEvoy and Yang, 2012). Thus, the existence within the LOC (defined by traditional activation methods) of fine-grained patterns distinguishing intact and scrambled objects was expected. Consequently, LOC was used as “ground truth” in order to compare the massive-GNB and SVM.

2.5.1. *Participants*

Twenty-six healthy students (9 females), with ages from 23 to 28 years (mean = 25.7 sd = 1.6), participated in the experiment. All were university graduates, native Chinese speakers, and fluent English readers. All had normal, or corrected-to-normal, vision and were right handed (except for two cases). None had a history of neurological or psychiatric disease. The experimental procedures were previously approved by the UESTC ethics committees, were carried out in accordance with the declaration of Helsinki, and all participants gave written informed consent.

2.5.2. Stimuli

Stimuli were generated using the Cogent Matlab toolbox (http://www.vislab.ucl.ac.uk/cogent.php), projected on a screen near the subject's feet in the scanner, and viewed through an angled mirror fixed to the MRI head-coil. To identify the LOC region, twenty black & white drawings of objects (see Fig. S2 for examples) (Snodgrass and Vanderwart, 1980), and their corresponding scrambled versions, were used. These were projected at visual angles of about 5.5° × 5.5°. Scrambling was achieved by dividing each object picture into 100 sectors, which were randomly re-positioned. Four blocks of intact and 4 blocks of scrambled pictures were alternated in each of 3 runs (for a total of 12 blocks of each type). A block included 20 stimuli, each lasting 300 ms and separated from each other by a 500 ms fixation point. Thus blocks lasted 16 s. They were separated from each other by a 16 s fixation point. The subjects were instructed to detect a 1-back repetition of a randomly selected stimulus within each block.

2.5.3. Data acquisition

Recordings were obtained at UESTC with a GE Discovery MR750 3 T scanner (General Electric Medical Systems, Milwaukee, WI, USA), using an 8 channel receiver head coil. Functional images were acquired with 35 slices covering all the head (except the vertex). A T2*- weighted echo planar imaging sequence was used with the parameters: TR = 3 s, TE = 40 ms, flip angle = 90°, voxel size = 3 × 3x3 mm, a gap between slices of 3 mm, and an acquisition matrix = 64 × 64. There were 90 images per run, from which the initial 5 vol were discarded to stabilize T1 magnetization. A 262 slice anatomical T1-weighted image was also obtained with the following parameters: voxel size = 1 × 1x0.5 mm, TR = 8.10 ms, TE = 3.16 ms, acquisition matrix = 256 × 256, and flip angle = 12.

2.5.4. Image preprocessing and univariate analysis

White matter and pial surfaces were reconstructed from the anatomical image for each subject using Freesurfer (http://surfer.nmr.mgh.harvard.edu), then registered to the FsAverage template and subsampled to 81924 nodes (vertices). A mid-gray surface was calculated as the mean of white and pial surface node coordinates. A set of 5 mm discs was defined around all nodes in the surface by means of the Surfing toolbox (http://surfing.sourceforget.net). For all functional series, artifact correction was performed with the ArtRepair toolbox (http://cibsr.stanford.edu/tools/ArtRepair/ArtRepair.htm). Pre-processing was performed with SPM8 (http://www.fil.ion.ucl.ac.uk/spm/) and included slice-timing, head motion correction (with extraction of motion parameters) and unwarping. Each T1 image was co-registered with the mean preprocessed-functional image, and the transformation matrix generated in this step was used to project the mid-gray surface into each subject's functional native space. Volume BOLD signals were interpolated at the coordinates of the mid-gray nodes, producing surface time-series (without spatial smoothing) that were high-pass filtered with cutoff of 128 s. A general linear model (GLM) was fit to the time-series of each surface node using regressors for each stimulation block (i.e. square-waves convolved with the canonical hemodynamic function), plus the head movement parameters and mean signal in each session as nuisance covariates. The beta parameters estimated for each block were used as features in the subsequent MVPA analyses. Only data from the left hemisphere was included in the analyses to reduce the duration of computations.

2.5.5. Comparison of massive-GNB and SVM

Due to the possible disadvantages of the GNB classifier, examined in the introduction, a comparison of the accuracy of our method in with the widely used SVM was carried out. This classifier is considered by many authors to be more sensitive than GNB in detecting informative brain regions in MVPA (e.g. Ku et al., 2008). The libSVM library (Chang and Lin, 2011) was used to calculate the SVM (with the default value of the box-constraint, C = 1). A permutation procedure was used, in which trial labels were swapped with a Monte Carlo scheme (n = 1000), with the restriction that exchanges were limited to the same experimental run. For each permutation, classification accuracy was calculated in a leave one-run out cross-validation scheme, and the results across the 3 folds averaged. The significance of observed accuracies, in each searchlight and in each subject, was estimated from the empirical null distribution. This null distribution was built from the maximum accuracy across all searchlights in each permutation (searchlight-wise correction for multiple comparison, Nichols and Holmes, 2002). The sizes of clusters surviving a threshold of 75% correct classifications (equivalent to $p < 0.01$ in the binomial test) in each permutation were measured, and their empirical null distribution in each subject was used to estimate the significance of the observed cluster sizes (i.e. cluster-level correction for multiple comparison, Hayasaka and Nichols, 2003). In addition, to guarantee that discrimination between conditions was not based on a simple difference in mean activation, the data in each trial and searchlight were corrected by subtracting the mean activation in the searchlights from the contributing cortical nodes in each trial (Coutanche, 2013). Thus the distinction between intact and scrambled objects could not be based on lower spatial frequency patterns based on broad activations over the cortex.

To compare the two classifiers the kernel density of the bivariate distribution of searchlight accuracies for massive-GNB and SVM was calculated in all subjects and for the group median accuracies. The concordance between massive-GNB and SVM maps of searchlights significantly above chance (for both searchlight-wise and cluster-level corrections) was compared across subjects using the Dice coefficient (Dice, 1945). In addition to these within-subject comparisons, group maps of significant effects were obtained using a random-effect prevalence method for accuracies (Allefeld et al., 2016), that incorporates the maximum statistic across searchlights to correct for multiple comparisons. The SVM and massive-GNB group prevalence maps were also compared with the Dice coefficient.

Furthermore, the sensitivity of both classifiers in identifying clusters that overlapped the LOC was estimated, using as ground truth coordinates identified from a meta-analysis of visual object recognition based on coordinates from maxima of univariate activations in 708 published studies (http://www.neurosynth.org/). This meta-analytic LOC was defined as the cortical voxels surviving FDR thresholding ($q = 0.01$) in the Neurosynth reverse inference map for the term ‘Object’. These voxels were mapped onto the left hemisphere FsAverage surface template, and surface clusters with areas smaller than 100 mm^2 were eliminated. This procedure isolated a meta-analytic LOC (see Fig. S3). Although the match between the LOC multivariate patterns in our subjects and the LOC identified in the meta-analysis of activations may not be perfect, there should be a large overlap between them, and any disjunction would equally challenge both classifiers. The bivariate kernel density was also estimated only for the searchlights within the meta-analytic LOC in each subject as well as for the group median values.

The selectivity of the searchlight analysis for localizing the ground truth of this experiment was evaluated within a Bayesian framework. The intuition behind this analysis is that for a classifier to be selective of a region, large accuracies should be observed at this ROI but not at other regions. The posterior probability of a cortical node to be included in the LOC region given its accuracy, Pr(LOC|a), was estimated for the group median values (equation (4)). In this equation, Pr(a|LOC) is the probability distribution of accuracies in the meta-analytic LOC, Pr(LOC) is the prior probability of being a LOC node, and the denominator of the equation is the probability distribution of accuracies across all nodes. The probability densities were modeled using univariate kernel models for each classifier separately.

$$\Pr(LOC|a) = \frac{\Pr(a|LOC)x\Pr(LOC)}{\Pr(a|LOC)x\Pr(LOC) + \Pr(a|\sim LOC)x\Pr(\sim LOC)} \qquad [4]$$

This Bayesian analysis assumes that all searchlights in the LOC are governed by a common accuracy distribution under this cognitive task, which is probably not valid. However, this is a useful approximation since ROIs from fMRI localizers are considered homogenous from a functional point of view.

3. Results

3.1. Run time analysis

Computational times for the massive-GNB, the searchmight-GNB and the cosmo-GNB are shown in Table 1, for combinations of values of two parameters. Massive-GNB was faster than the other two algorithms when the largest number of searchlights was used. Since the cosmo-GNB was always slower than the searchmight-GNB in this comparison, our method was benchmarked only against the latter in the subsequent analyses.

The runtimes as a function of three parameters of searchlight analyses are presented in Fig. 1, which shows that massive-GNB is always faster than searchmight-GNB classifier except when very few searchlights were included. The dependence of CPU time on the number of searchlights showed slopes of 0.21×10^{-4} and 1.9×10^{-4} s/searchlight for the massive-GNB and searchmight-GNB respectively (Fig. 1A). When all searchlights were involved (the rightmost point), the massive-GNB classifier was 6 times faster than the searchmight-GNB. The respective times were 7.55 and 43.0 s. The dependence of the CPU runtime with the searchlight radius (Fig. 1B) can be described with a power function with exponents: 2.15 for the searchmight-GNB 1.82 and for the massive-GNB. Both power models fitted the data for an alpha level of 0.05. This power behavior is a consequence of the rapid increase in the number of neighbors included in the searchlight when the searchlight radius is augmented. The dependence of runtimes on the number of samples is shown in Fig. 1C. As in the previous plots, the growth of runtime with the parameter is much slower for massive-GNB than for searchmight-GNB.

Multithreading parallelization of the permutation loop reduced computation time by a factor of 7.9 for the massive-GNB: from 139 min (2.3 Hrs.) with no parallelization to 17.6 min when parallelized (see Fig. 2). In similar conditions, the Searchmight- GNB computation time was reduced from 691 min (11.5 Hrs) to 172 min (2.9 Hrs): a factor of 4.0.

3.2. Comparison of massive-GNB and SVM

The permutation tests for the twenty-six participants in the LOC fMRI localizer experiment took a total of about 4.3 days of computation, whereas the massive-GNB only took about 7 h (0.13 days). The bivariate-distribution of the group median accuracies (upper left panel of Fig. 3) allows direct comparison of massive-GNB and the SVM performance across searchlights. This bivariate distribution was well modeled by a gaussian mixture model with two density components. The first one (about 81% of the cortical nodes and probably corresponding to non-informative searchlights), had only a moderate correlation ($r = 0.49$) between classifiers and mean accuracies near chance for SVM (0.539) and GNB (0.542). The second component (about 19% of the nodes and probably corresponding to informative searchlights), presented a high correlation ($r = 0.92$) between classifiers and somewhat larger mean accuracies for both SVM (0.672) and GNB (0.676). This component extended above the significance thresholds for searchlight-wise correction. In Fig. S4 equivalent plots for each of the participants are shown, in which the concordance between the two classifiers in the supra-threshold region was confirmed in most of the subjects.

Table 1
Times in seconds for three implementations of the GNB algorithm. Parameters: number of searchlights (S) and mean number of neighbors within each searchlight (N) which correspond to radii 1 and 2 respectively in volume space.

	S = 10000
	N = 27
		S = 234067
		N = 27
			S = 10000
			N = 125
				S = 234067
				N = 125

massive-GNB	1.82	3.76	1.99	6.82
searchmight-GNB	0.32	10.59	1.21	34.99
cosmo-GNB	2.62	22.85	4.46	115.39

The bivariate distribution of median accuracies of the searchlights within the meta-analytic LOC are shown in the upper right panel of Fig. 3. Most of the distribution values were above threshold for both massive-GNB and SVM. Consequently, both classifiers identify a large proportion of the meta-analytic LOC. However, the distribution extended somewhat more below the main diagonal, indicating slightly larger accuracies for SVM than GNB. To confirm these conclusions, the selectivity (posterior probability) of both classifiers for the localizing the meta-analytic LOC as a function of their accuracy is shown in the lower panel of Fig. 3. Although for both classifiers all cortical nodes with accuracies over 90% had a large posterior probability of falling within the LOC, this probability was almost 1.0 for the SVM classifier but nearer 0.92 for the massive-GNB. More generally, when the accuracy was larger than 75% SVM exhibited slightly more selectivity for LOC than GNB. This posterior probability should be interpreted as the change in our belief that a cortical node is part of LOC (which here had a prior probability smaller than 0.3), after one has known the classifier accuracy.

The group prevalence maps for both classifiers are shown in the upper row of Fig. 4. The threshold was $p < 0.05$ level (searchlight-wise corrected), which corresponds to an above chance classification of at least 60% of the participants. In addition to the LOC, small informative patches were also observed in other regions (principally the parietal lobe). There was a moderate degree of overlap between the two searchlight maps. About 34% more searchlights were significant with SVM ($n = 993$) than with the massive-GNB ($n = 740$). However, traditional maps only plot searchlight centers, ignoring the substantial overlap between searchlights. Correspondingly, smoothed versions of these maps (obtained by painting all nodes belonging to significant searchlights) exhibited an almost perfect agreement between the two classifiers (lower row Fig. 4), with a larger Dice index than for the unsmoothed version. The disagreement between these maps was concentrated at the borders of LOC, with SVM occupying slightly more area (additional views of the smoothed maps are presented in Fig. S5). These results support the idea that SVM is a more sensitive classifier, thus producing broader informative patches.

The prevalence method uses only the searchlight-wise threshold correction for multiple comparisons. Nonetheless, it is usual to make inferences on topological features, not individual elements (see Chumbley and Friston, 2009). Thus, the two classifiers were compared under cluster-level corrections for multiple comparisons. Fig. 5 shows boxplots of Dice indexes across participants for both searchlight-wise and cluster-level corrections. This analysis was constrained to the anatomical areas known to contain the LOC proper.

The first four rows of Fig. 5 show the comparison of the meta-analytic LOC with searchlights maps thresholded by the permutation tests (separately for searchlight-wise and cluster-level corrections). The SVM and massive-GNB maps agreed only moderately with the meta-analytic LOC (medians respectively 0.48 and 0.45) under searchlight-wise correction, but were not significantly different from each other. Under cluster-level correction, SVM and massive-GNB maps agreed significantly more with the LOC (medians respectively 0.52 and 0.53) than their searchlight-wise counterparts (SVM: $p < 0.05$; massive-GNB: $p < 0.015$), while not disagreeing from each other.

The Dice indexes between searchlight-wise corrected SVM and massive-GNB maps showed substantially agreement with each other (median = 0.71). Moreover, the corresponding cluster-level maps showed almost perfect agreement (median = 0.81), which was a highly significant ($p < 0.007$) increase relative to the case for searchlight-wise correction. A final comparison (see Fig. S7 and Text S2) showed that the hit rate for LOC nodes in the permutation tests across participants was equivalent for the two classifiers under searchlight-wise threshold correction, but significantly higher for SVM, and even higher for massive-GNB, under cluster-level correction.

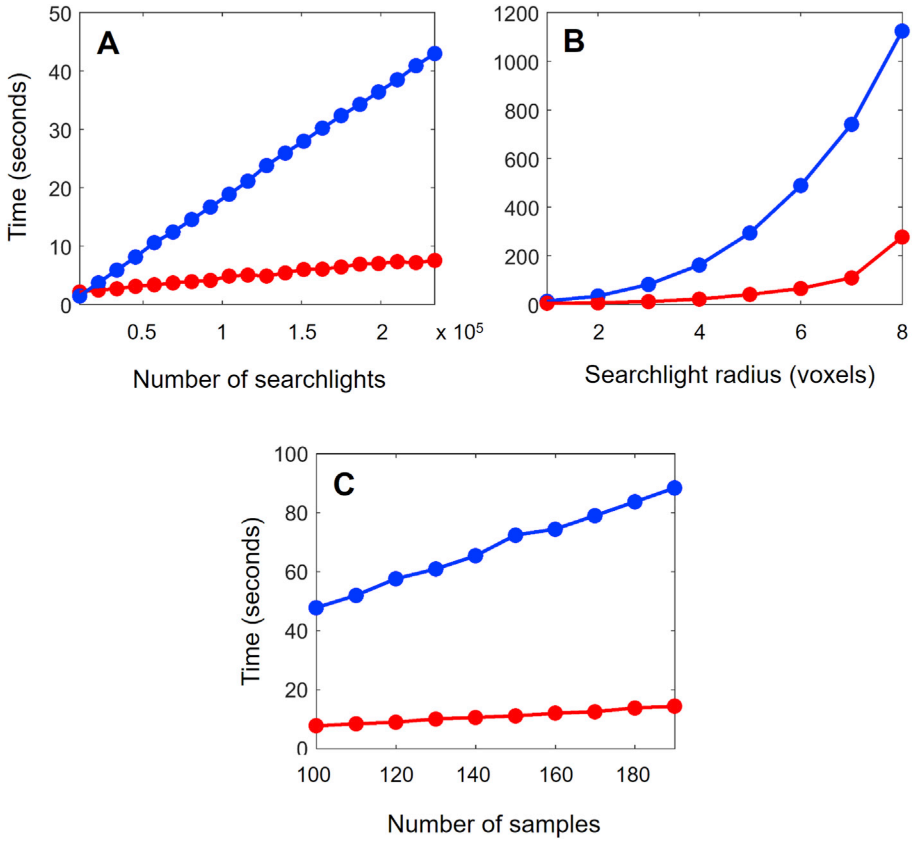

Fig. 1. Runtime analysis for the massive-GNB (red) and searchmight-GNB (blue) while varying three parameters A) number of searchlights, B) mean number of neighbors in the searchlight and C) number of samples (trials).

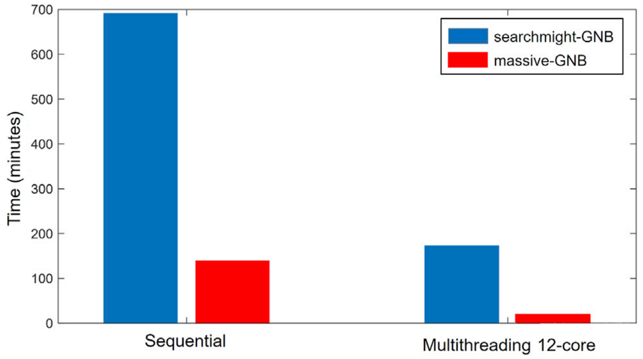

Fig. 2. Results of the parallelization study for searchmight-GNB and for massive-GNB. The columns to the left show the CPU times for the sequential implementation. The columns to the right show the results the distributing the permutation loop across 12 cores using the Matlab parfor function.

4. Discussion and conclusions

An algorithmic framework for GNB was introduced here with the goal of significantly accelerating computations for searchlight-based MVPA, which can be very intensive in computer time. This is a requirement for the widespread application of re-sampling statistical techniques to this type of analysis. This framework is based on sparse matrix operations (which are efficiently implemented in different programming languages), taking advantage of the nature of the neighborhood structure of searchlights. When compared with previous approaches, including the Searchmight toolbox (to our knowledge the fastest publically available GNB method), computational times were notably reduced. The advantage in speed for massive-GNB over libSVM was even more striking (about 34.3 times faster).

An exploratory analysis revealed an additional speed gain when massive-GNB was combined with a simple hardware parallelization tools in Matlab. Since the massive-GNB and the Searchmight toolboxes use different programming languages (Matlab and C++ respectively), and the codes were not optimized for parallelization, this comparison of speed between methods should be taken with caution. Additional parallelization strategies should be explored, but are outside the scope of this article.

When massive-GNB was compared with SVM (in an LOC localizer fMRI experiment), classification accuracies tended to be slightly larger for SVM than for massive-GNB analyses across individual searchlights. Although both had large sensitivity in detecting the meta-analytic LOC in the Bayesian analysis, the posterior probabilities for SVM were slightly larger. The somewhat better accuracies found for SVM relative to the massive-GNB across searchlights are in line with previous comparative studies (Ku et al., 2008; Misaki et al., 2010). This superior detection may be related to the greater ability of SVM to adjust the discrimination hyperplane to differences in the covariance matrix between conditions, and/or the existence of covariance between cortical nodes within each searchlight.

In the second-level prevalence test (based on searchlight-wise correction for multiple comparisons), SVM maps also identified somewhat larger informative patches than massive-GNB maps, although they presented almost perfect agreement in a smoothed version. The divergence between the two classifiers was most apparent at the borders of the group-level informative cortical patches. However, at the level of individual participants, the searchlight-wise corrected maps for the two classifiers did not differ significantly in their degree of agreement with the meta-analytic LOC. In fact, the agreement of both methods with the meta-analytic "ground truth" increased greatly under cluster-level compared to searchlight-wise correction. Furthermore, the agreement

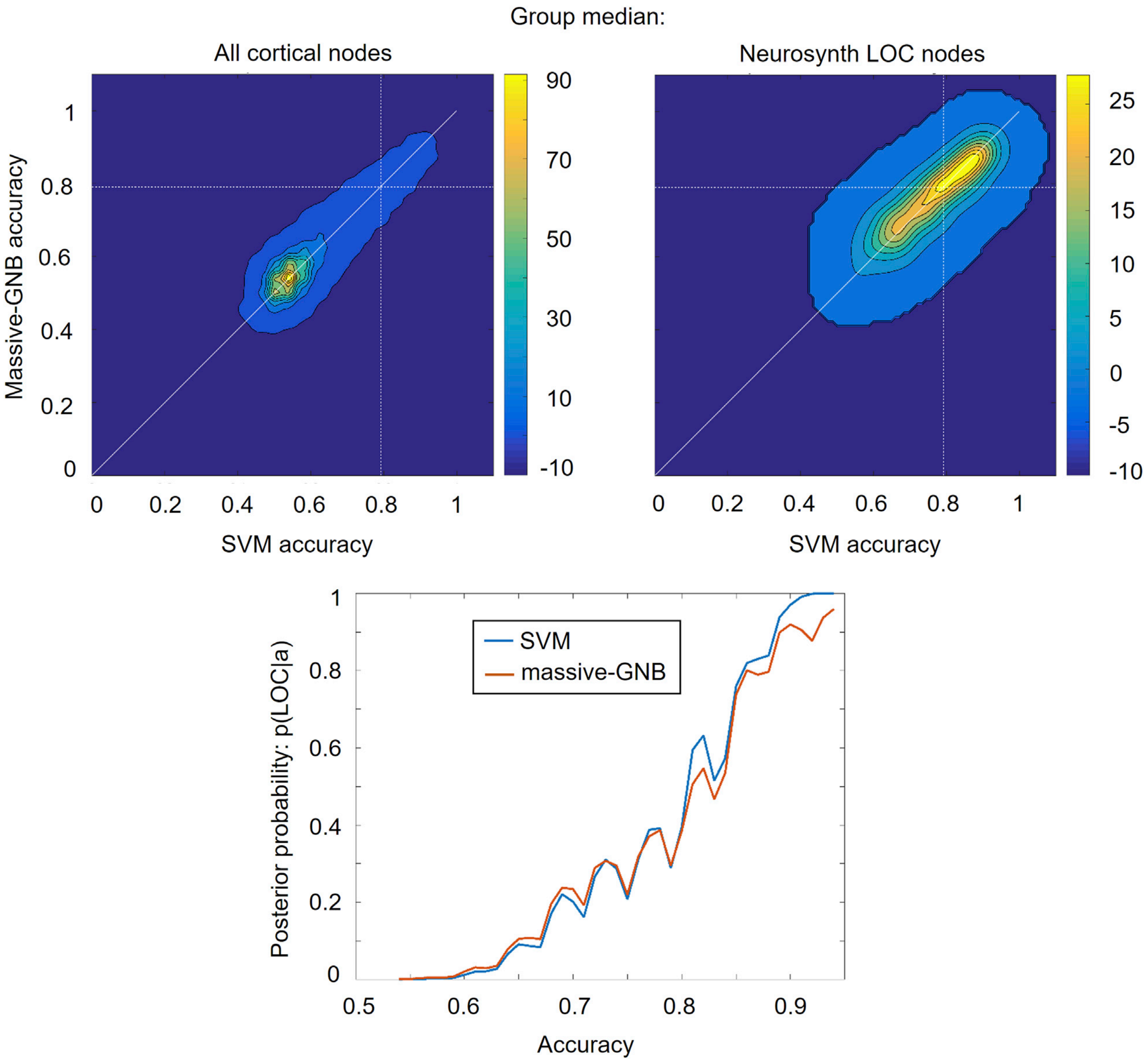

Fig. 3. Bivariate distributions (kernel density) of group median accuracies. In the upper left panel, the contour plot of the distribution of massive-GNB accuracy as a function of SVM accuracy for all searchlights (left hemisphere only). In the upper right panel, the equivalent plot for the subset of searchlights whose centers fell within the LOC as defined by the Neurosynth meta-analysis. The dotted lines indicate the thresholds for above chance accuracy (searchlight-wise corrected for multiple comparisons). Equal accuracy corresponds to solid white lines. In the lower panel the posterior probability of a searchlight being in LOC given its accuracy for the two classifiers.

between maps for the two classifiers was almost perfect under cluster-level correction. The massive-GNB performed best in detecting informative clusters (instead of individual searchlights), both in the individual participants and the group level. This indicates that the enhanced speed of massive-GNB is of practical import for identifying informative clusters.

Perhaps the extent of clusters would be underestimated due to poorer GNB performance at patch borders. However, this is a problem with any cluster identification method, which all depend strongly on how thresholds are (arbitrarily) fixed (Smith and Nichols, 2009). A recent study found that GNB second-level maps were very consistent across different experiments and subjects (Raizada and Lee, 2013), probably due to the ability of GNB to “smooth” the individual accuracy maps facilitating between subjects alignment. This would facilitate the reliable detection of core areas of informative clusters across subjects.

Our results reinforce the idea that classifiers for searchlight MVPA should be measured against a common “ground truth” (when it is available), instead of merely comparing their accuracies. Recently Zhang et al (2017) have shown that SVM can produce false positives in fMRI data when classifying object categories, even in white matter ROIs, which confirms that we cannot trust any classifier to be right always. A caveat is that “ground truth” used here is based on the properties of LOC, which is associated to very robust fMRI effects. More subtle differences in activation pattern in other experiments may show a greater advantage in sensitivity and specificity for the SVM respect to GNB. However, our results indicate that for certain goals and experiments the GNB can be practically equivalent to SVM.

The huge gain in speed obtained with our massive-GNB approach, enables carrying out the vast amount of calculations needed to solve the challenges for searchlight analysis outlined in the introduction without compromising validity (especially at the cluster-level or second-level analyses). Permutation tests and different forms of cross-validation, are some of the procedures that can now be performed rapidly on personal PCs without the need for distributed processors (although their use lead to an even greater gain in speed). The high demands of these methods on

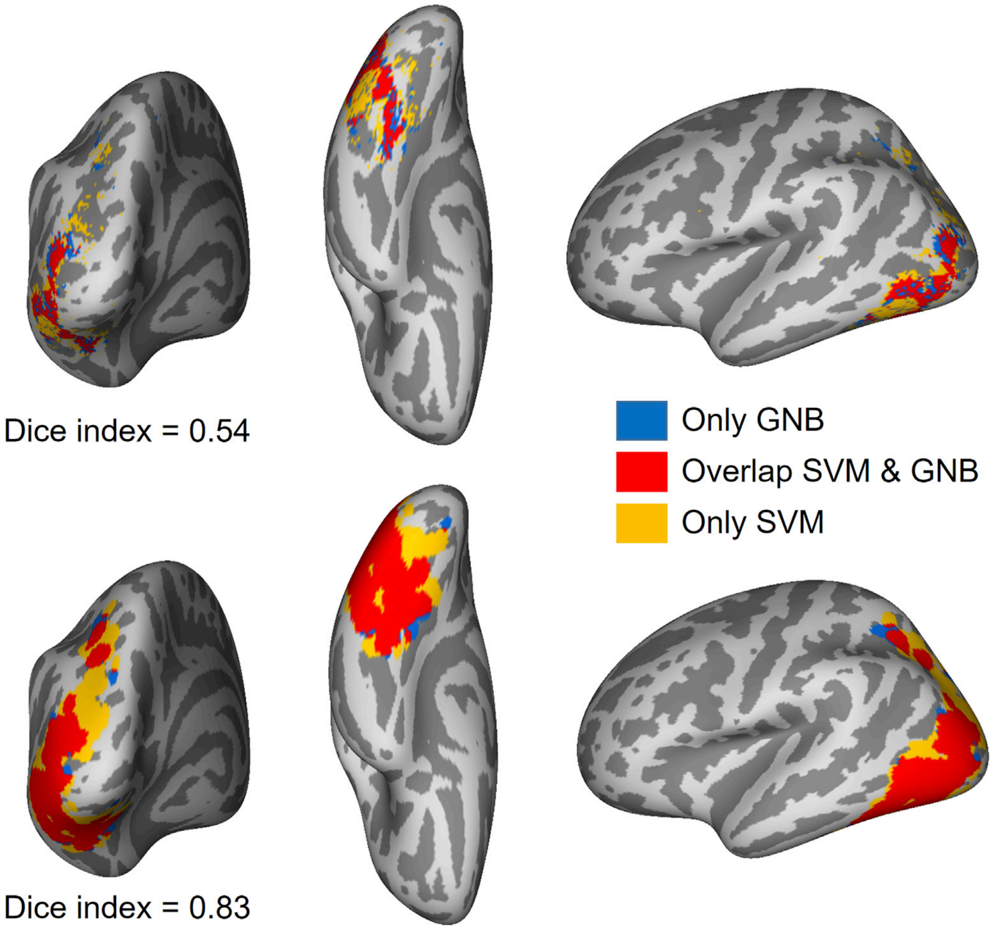

Fig. 4. Prevalence group analysis. Searchlights that classify objects vs. scrambled objects above chance are mapped for GNB and SVM. Different colors indicate areas in which only GNB, only SVM, and both classifiers were significant. In the top row a traditional map in which only the centers of significant searchlights are displayed. Below a smoothed map in which the full extent of each significant searchlights is filled (clusters < 200 mm^2 were omitted).

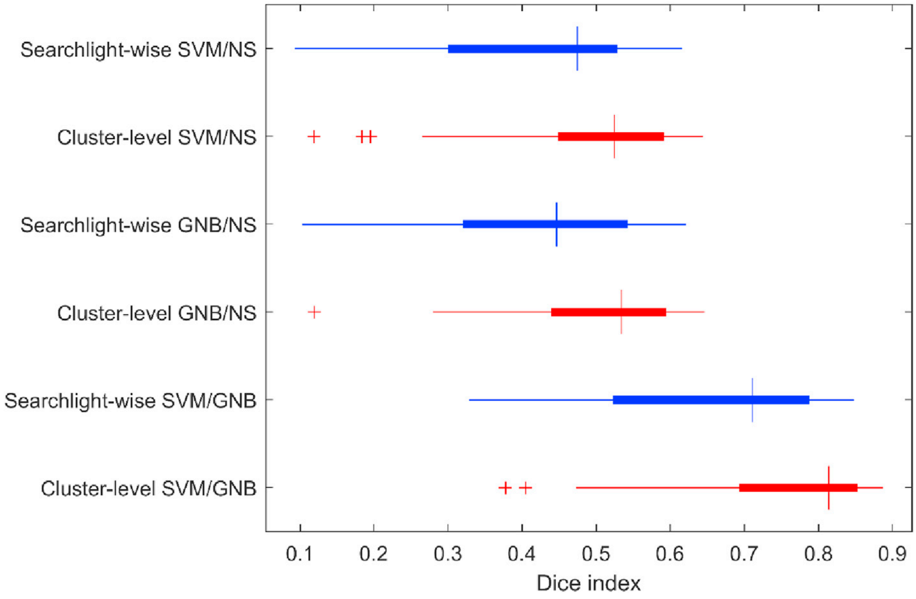

Fig. 5. Boxplot of Dice indexes across participants. Blue boxes correspond to maps with searchlight-wise thresholds, red boxes correspond to cluster-level thresholds. NS indicates that the (Neurosynth) meta-analytic LOC map was involved. Box extremes indicate the 25 and 75 percentiles, whereas vertical lines depict the median, and horizontal lines the range. This analysis was restricted to the anatomical mask shown in Fig. S6.

computation time may discourage their more widespread use. Although this result is of practical importance for fMRI data obtained at 1.5 and 3 T field strength, it is even more important for the huge datasets obtained at ultra-high field strengths. Finally, the methods developed in this article are publically available at: https://github.com/mlsttin/massive_gaussian_naive_bayes.

Acknowledgements

We thank Dezhong Yao, Jinnan Gong, Cheng Luo for their participation in collecting the LOC localizer data, as well as Thomas Emerling for the use of the imagined motion data, and also Elia Formisano and Federico de Martino for comments on the manuscript.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.neuroimage.2017.09.001.

Funding

This work was supported by the Foundation for Science and Innovation, Cuba (Grant code: CNeuro-12-15).

Competing interests

The authors have declared that no competing interests exist.

References

Allefeld, C., Haynes, J.D., 2014. Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. Neuroimage 89, 345–357.

Allefeld, C., Görgen, K., Haynes, J.D., 2016. Valid population inference for information-based imaging: from the second-level t-test to prevalence inference. Neuroimage 141, 378–392.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer-Verlag New York, Inc, Secaucus, NJ, USA ©2006.

Chang, C.C., Lin, C.J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intelligent Syst. Technol. (TIST) 2 (3), 27.

Chumbley, J.R., Friston, K.J., 2009. False discovery rate revisited: FDR and topological inference using Gaussian random fields. Neuroimage 44 (1), 62–70.

Coutanche, M.N., 2013. Distinguishing multi-voxel patterns and mean activation: why, how, and what does it tell us? Cognitive, Affect. Behav. Neurosci. 13 (3), 667–673.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302.

Eklund, A., Dufort, P., Villani, M., LaConte, S., 2014. BROCCOLI: software for fast fMRI analysis on many-core CPUs and GPUs. Front. Neuroinf. 8, 24.

Emmerling, T.C., Zimmermann, J., Sorger, B., Frost, M.A., Goebel, R., 2016. Decoding the direction of imagined visual motion using 7T ultra-high field fMRI. Neuroimage 125, 61–73.

Etzel, J.A., Zacks, J.M., Braver, T.S., 2013. Searchlight analysis: promise, pitfalls, and potential. Neuroimage 78, 261–269.

Golomb, J.D., Kanwisher, N., 2011. Higher level visual cortex represents retinotopic, not spatiotopic, object location. Cereb. Cortex 22 (12), 2794–2810.

Grill-Spector, K., Kourtzi, Z., Kanwisher, N., 2001. The lateral occipital complex and its role in object recognition. Vis. Res. 41 (10), 1409–1422.

Hayasaka, S., Nichols, T.E., 2003. Validating cluster size inference: random field and permutation methods. Neuroimage 20 (4), 2343–2356.

Haynes, J.D., 2015. A primer on pattern-based approaches to fMRI: principles, pitfalls, and perspectives. Neuron 87 (2), 257–270.

Hebart, M.N., Görgen, K., Haynes, J.D., 2015. The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. Front. Neuroinf. 8 (88).

Jamalabadi, H., Alizadeh, S., Schönauer, M., Leibold, C., Gais, S., 2016. Classification based hypothesis testing in neuroscience: below-chance level classification rates and overlooked statistical properties of linear parametric classifiers. Hum. Brain Mapp. 37 (5), 1842–1855.

Jimura, K., Poldrack, R.A., 2012. Analyses of regional-average activation and multivoxel pattern information tell complementary stories. Neuropsychologia 50 (4), 544–552.

Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. 103, 3863–3868.

Krzanowski, W.J., 1988. Principles of Multivariate Analysis: a User's Perspective. Clarendon.

Ku, S.P., Gretton, A., Macke, J., Logothetis, N.K., 2008. Comparison of pattern recognition methods in classifying high-resolution BOLD signals obtained at high magnetic field in monkeys. Magn. Reson. Imaging 26 (7), 1007–1014.

MacEvoy, S.P., Yang, Z., 2012. Joint neuronal tuning for object form and position in the human lateral occipital complex. Neuroimage 63 (4), 1901–1908.

Misaki, M., Kim, Y., Bandettini, P.A., Kriegeskorte, N., 2010. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. Neuroimage 53 (1), 103–118.

Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15 (1), 1–25.

Noirhomme, Q., Lesenfants, D., Gomez, F., Soddu, A., Schrouff, J., Garraux, G., Laureys, S., 2014. Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions. Neuroimage Clin. 4, 687–694.

Ojala, M., Garriga, G.C., 2010. Permutation tests for studying classifier performance. J. Mach. Learn. Res. 11 (Jun), 1833–1863.

Oosterhof, N.N., Wiestler, T., Downing, P.E., Diedrichsen, J., 2011. A comparison of volume-based and surface-based multi-voxel pattern analysis. Neuroimage 56 (2), 593–600.

Oosterhof, N.N., Connolly, A.C., Haxby, J.V., 2016. CoSMoMVPA: multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU Octave. Front. Neuroinf. 10.

Pereira, F., Botvinick, M., 2011. Information mapping with pattern classifiers: a comparative study. Neuroimage 56, 476–496.

Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classi!ers and fMRI: a tutorial overview. Neuroimage 45, 199–209.

Raizada, R.D.S., Lee, Y.S., 2013. Smoothness without smoothing: why gaussian naive bayes is not naive for multi-subject searchlight studies. PLoS One 8 (7).

Saxe, R., Brett, M., Kanwisher, N., 2006. Divide and conquer: a defense of functional localizers. Neuroimage 30, 1088–1096.

Smith, S.M., Nichols, T.E., 2009. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage 44 (1), 83–98.

Snodgrass, J.G., Vanderwart, M., 1980. A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. J. Exp. Psychol. Hum. Learn. Mem. 6 (2), 174.

Spiridon, M., Fischl, B., Kanwisher, N., 2006. Location and spatial profile of category-specific regions in human extrastriate cortex. Hum. Brain Mapp. 27 (1), 77–89.

Stelzer, J., Chen, Y., Turner, R., 2013. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. Neuroimage 65, 69–82.

Valente, G., Castellanos, A.L., Vanacore, G., Formisano, E., 2014. Multivariate linear regression of high-dimensional fMRI data with multiple target variables. Hum. Brain Mapp. 35 (5), 2163–2177.

Wang, X., Hutchinson, R., Mitchell, T.M., 2004. Training fMRI classifiers to discriminate cognitive state across multiple human subjects. In: Proceedings of the 18th Annual Conference on Neural Information Processing Systems, pp. 709–716.

Zhang, Z., Jiang, Y., Sun, Y., Zhang, H., 2017. Potential for false positive results from multi-voxel pattern analysis on functional imaging data. Technol. Health Care, (Preprint) 1–8.