Information Fusion 66 (2021) 198–212

Contents lists available at ScienceDirect

Information Fusion

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

Full length article

Statistical Agnostic Mapping: A framework in neuroimaging based on concentration inequalities

J.M. Gorriz [a,b,c,*], C. Jimenez-Mesa [a], R. Romero-Garcia [b], F. Segovia [a], J. Ramirez [a], D. Castillo-Barnes [a], F.J. Martinez-Murcia [a], A. Ortiz [a], D. Salas-Gonzalez [a], I.A. Illan [a], C.G. Puntonet [a], D. Lopez-Garcia [a], M. Gomez-Rio [c], J. Suckling [b]

[a] DaSCI Institute, University of Granada, Spain
[b] Department of Psychiatry, University of Cambridge, UK
[c] ibs.Granada, Granada, Spain

ARTICLE INFO

Keywords:
Hypothesis testing
Upper bounds
Actual and empirical risks
Finite class lemma
Rademacher averages
Cross-validation

ABSTRACT

In the 1970s a novel branch of statistics emerged focusing its effort on the selection of a function for the pattern recognition problem that would fulfill a relationship between the quality of the approximation and its complexity. This theory is mainly devoted to problems of estimating dependencies in the case of limited sample sizes, and comprise all the empirical out-of sample generalization approaches; e.g. cross validation (CV). In this paper a data-driven approach based on concentration inequalities is designed for testing competing hypothesis or comparing different models. In this sense we derive a Statistical Agnostic (non-parametric) Mapping (SAM) for neuroimages at voxel or regional levels which is able to: (i) relieve the problem of instability with limited sample sizes when estimating the actual risk via CV; and (ii) provide an alternative way of Family-wise-error (FWE) corrected *p*-value maps in inferential statistics for hypothesis testing. Using several neuroimaging datasets (containing large and small effects) and random task group analyses to compute empirical familywise error rates, this novel framework resulted in a model validation method for small samples over dimension ratios, and a less-conservative procedure than FWE *p*-value correction to determine the significance maps from the inferences made using small upper bounds of the actual risk.

1. Introduction

Over the last few decades translational neuroscience has transitioned from qualitative case reports to quantitative, longitudinal and multivariate population studies in the quest for defining patterns of disease pathogenesis, prognostic indicators and treatment response. Neuroscience has provided valuable insights by means of classical statistics, primarily statistical inference based on null-hypothesis (H_0) testing that the brain mapping community has predominantly used for exploratory analyses in whole brain searches [1]. In this context, classical inference emphases in-sample, image-based statistical estimates from previously assumed data models to determine the existence of relevant effects (large or subtle) across a range of designs where the critical *p*-value (significance vs. non-significance) can be complemented by the corresponding effect-size estimates [2].

Recently, several advances for combining *p*-value maps have been proposed based on the concept of *prevalence* [3,4] beyond the fixed and mixed (random) effects models [5]. Common to all these approaches is to assume a voxel-wise model that allows a proportion of conditions or subjects that activated the voxel at some mixing proportion. This assumption that is more realistic than those assumed in classic random effect approaches, e.g. homogeneity in the activation pattern (binary), clearly opens a new application field for modern statistics.

Conversely, out-of sample generalization approaches common in machine learning (ML), such as Cross-Validation (CV), try to estimate on unseen new data the accuracy of the classifier in the (binary) classification problem. Despite the methods and goals of predictive CV inference being distinct from classical extrapolation procedures [6], they are actually exploited within statistical frameworks aimed at assessing statistical significance [7]. Examples include bootstrapping, binomial or permutation ("resampling") tests [8], which have been demonstrated to be competitive outside the comfort zone of classical statistics, filling otherwise-unmet inferential needs.

In the pattern classification problem we usually assume the existence of classes (H_1) that are differentiated by classifiers that are measured by their performance in terms of accuracy (A_{cc}) or *prevalence*

* Corresponding author at: DaSCI Institute, University of Granada, Spain.
E-mail addresses: gorriz@ugr.es, jg825@cam.ac.uk (J.M. Gorriz).

https://doi.org/10.1016/j.inffus.2020.09.008
Received 22 August 2020; Received in revised form 14 September 2020; Accepted 19 September 2020
Available online 26 September 2020

on a independent dataset, and conclude (improperly in a statistical sense) H_1 using empirical confidence intervals. In limited sample sizes the most popular K-fold CV method [9] has been demonstrated to sub-optimally work under unstable conditions [10–12]. In such circumstances, the predictive power of the fitted classifiers can be arguable. Moreover, recent works have partially demonstrate that, when using only a classifier's empirical accuracy as a test statistic, the probability of detecting differences between two distributions is lower than that of a bona fide statistical test [13,14].

Beyond the latter empirical techniques for the estimation of performance, ML is well-framed into a data-driven statistical learning theory (SLT) which is mainly devoted to problems of estimating dependencies with limited amounts of data [15]. Although, CV-ML approaches were not originally designed to test hypotheses based on prevalence in brain mapping [1], they are theoretically grounded to provide confidence intervals in the classification of image patterns (protected inference) that can be seen as maps of statistical *significance*. As shown in the present study, this can be achieved by assessing the upper bounds of the actual error in a binary classification problem (a confidence interval), and by using simple significance tests of a population proportion within it. Definitely, this would result in improvements to the test's statistical power based on accuracy.

The paper is organized as follows. In Section 2, using the agnostic or model-free formulation of the learning problem we derive the analytical upper bound of the real (actual) error from the empirical (measured) error of the model under realistic conditions. Then, we apply these concepts to neuroimaging and define the significance of a region, or the Statistical Agnostic Mapping (SAM), by comparing the upper bound of the actual error of the model in that region to a statistical threshold, that is based on a test for a proportion, as shown in Section 3. Moreover, sample size and the empirical settings regarding the complexity of the selected classifiers (feature extraction and selection, FES) are key to the proposed methodology as they condition the degrees of freedom and the number of separating functions used to derive the *deviation quantity*. Thus, the learning algorithm is preferably fitted with a linear classifier in the *best* possible way to low dimensional data, as shown in Section 4, to estimate the empirical error from the accuracy of classification of the voxels within a region using the full dataset.

In Section 6 we demonstrate the ability of the proposed method to detect *biased* significant regions, within a confidence interval, and to provide several operation modes (depending on the selected upper bound) with a conservative and valid regionwise inference for controlling the FWE. The experiments include a fair comparison with the standard SPM package[1] under several parameter combinations and thresholding approaches. Finally, some discussions are described in Section 7 and conclusions are presented in Section 8.

2. Methods: bounding the actual error with probability 1-δ

2.1. Background on agnostic learning

Assume the agnostic model for the problem of binary pattern classification as proposed in [16] applied to a region of interest (ROI) within an image. Given an independent and identically distributed sample $\mathbf{Z^n} = (\mathbf{Z_1}, \ldots, \mathbf{Z_n})$ (e.g images or ROIs) of d-dimensional predictors (e.g. voxels) and classes (e.g. conditions) pairs $Z_i = (\mathbf{x_i}, y_i)$, where each of them is drawn from the unknown distribution $P \in \mathcal{P}$, the goal is to construct a good approximation to an unknown target function or classifier f^*, using a class of functions $\mathcal{F} : f : \mathbf{X} \to U$, and evaluating their goodness by a predefined expected loss:

$$L_P(f_n) \equiv \mathbb{E}[\ell(y, f_n)|\mathbf{Z^n}] = \int_{\mathbf{X}\times Y} \ell(y, \hat{f}_n) P(d\mathbf{x}, dy) \tag{1}$$

where the loss function $\ell : Y \times U \to [0, 1]$ and f_n is a random element of the hypothesis space U (the output class or condition proposed by the classifier).

To simplify notation, the *function composition*,[2] i.e. $g = \ell \circ f$, defines the class of functions $\mathcal{G} : g_\ell : Y \times \mathbf{X} \to [0, 1]$ with expected loss (probability of error) $P(g_\ell)$ similar to Eq. (1). Thus, the empirical error can be determined by counting the number of "misses" in the sample:

$$P_n(g_\ell) = \frac{1}{n} \sum_{i=1}^{n} g_\ell(\mathbf{Z_i}) \leq 1 \tag{2}$$

A learning algorithm particularly selects g_n given the sample $\mathbf{Z^n}$ via the empirical risk minimization (ERM) $g_n^* = \arg\min_{g\in\mathcal{G}} P_n(g)$ [15], and provides a real error $P(g_n)$ (on the ideal infinite population):

- close to that obtained with the sample, that is, $P(g_n) \simeq P_n(g_n)$
- and to the minimum risk , $L_P^*(\mathcal{G}) = \inf_{g\in\mathcal{G}} P(g) = P(g^*)$

2.2. Upper bound based on concentration inequalities

Unfortunately, $P(g_n) \simeq P_n(g_n)$ is not generally true. More precisely:

$$P(g_n) > P_n(g_n) + \epsilon > P(g^*) + \epsilon; \tag{3}$$

with an arbitrarily $\epsilon > 0$. Under the worst case scenario the uniform deviation can be defined as $\Delta_n(Z^n) = \sup_{g\in\mathcal{G}} |P_n(g) - P(g)|$, for any $g \in \mathcal{G}$. Using the ERM algorithm we readily get the following concentration inequalities:

$$\begin{aligned} P(g_n) &\leq P(g^*) + 2\Delta_n(\mathbf{Z^n}) \\ P(g_n) &\leq P_n(g_n) + \Delta_n(\mathbf{Z^n}) \end{aligned} \tag{4}$$

To our purposes, we prefer to work in terms of prevalence or accuracies, thus the second equation can be rewritten to:

$$\begin{aligned} &Acc(g_n) \geq Acc_n(g_n) - \Delta_n(\mathbf{Z^n}) \\ &\text{worst case: } Acc(g_n) = Acc_n(g_n) - \Delta_n(\mathbf{Z^n}) \end{aligned} \tag{5}$$

where $Acc = 1 - P$ refers to the actual and empirical accuracies.

Bounding $\Delta_n(Z^n)$ can be (but not readily) achieved by using several theorems and lemmas of the SLT [17–21] to finally get (see Appendix A)[3]:

$$P(g_n) \leq P(g^*) + 8\sqrt{\frac{\log(N)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} \tag{6}$$

with probability $1-\delta$, where N is the cardinality of $\mathcal{G}(Z^n)$ or the number of separating functions given the sample realization.[4]

3. Statistical agnostic mapping

The significant areas derived from SAM correspond by construction with those regions having an empirical error $P_n(g_n)$ that, under the worst case scenario, has associated an actual error $P(g_n)$ greater than the random guess accuracy $\pi = 0.5$. Confidence intervals derived from the concentration inequalities allow us to bound the worst case at the "upper" border of the confidence interval, providing a protective inference. Thus, within this confidence interval, a significance test can be used to make an inference about whether the accuracy value for a specific region differs from the null-hypothesis of the random proportion $\pi = 0.5$ (see Appendix A). Therefore, the statistical significance of any region is assessed, in combination with confidence intervals, by evaluating the p-value of any ROI at a given significance level, i.e. $\alpha = 0.05$. A total of $l = 116$ standardized regions [22] were analyzed within a protective interval, avoiding the limitations of significant tests to distinguish statistical from practical importances.

[1] https://www.fil.ion.ucl.ac.uk/spm/.

[2] $f(x) = u$; $\ell = \ell(y, u) \Rightarrow \ell(y, f(x)) = \ell_y(f(x)) = (\ell_y \circ f)(x) \equiv g_{\ell_y}(x) = g_\ell(y, x)$.

[3] A similar bound can be achieved for the second row in Eq. (4).

[4] A trivial bound for this *shattering number* can be found: $N \leq 2^n$.

In terms of classical statistics the SAMs are derived as the following. Given a set of regions $j = 1, \ldots, l$ we evaluate for each region the accuracy under the worst case in Eq. (4) by the following hypothesis test:

$$\begin{aligned} H_0 &: Acc^j > \hat{Acc}; \quad \text{region j is significant} \\ H_1 &: Acc^j < \hat{Acc}; \quad \text{region j is not significant} \end{aligned} \tag{7}$$

where $Acc^j = (1 - P_n^j) - \Delta_n(\mathbf{Z}^n)$ is the estimated actual accuracy in the classification of region j with probability $1 - \delta$ (see Eqs. (8) and (9)), and $\hat{Acc}$ is the averaged proportion of subjects correctly classified in all regions within the confidence interval. Note that the term $(1 - P_n^j)$ is the empirical accuracy of region j. Further details regarding the test for a proportion based on prevalence to achieve population inference are given in the Appendix A (see Eq. (18)), although other kind of tests could be applied as well, like those described in [23].

4. Fitting the selected function to current data

In order to minimize the left part of Eq. (6) we could minimize one (or both) of the elements on the right. However, they are dependent on each other in terms of the classifier complexity [15]. One solution could be, as explained in the next section, to prevent the increase of $N \propto \mathcal{O}(n, d)$ given the sample $\mathbf{Z}^n$, by selecting a low classifier order [24], i.e. a linear decision function. However, this comes at the cost of the possibility of a non-negligible empirical error.

4.1. Feature extraction and selection

As an attempt to reduce the ratio d/n (*curse of dimensionality*), the machine learning community has deployed FES methods to enhance the classification performance while preserving the system complexity. This can be achieved by removing irrelevant features from the sample, which can also facilitate interpretation (FS), and by identifying multivariate sets of meaningful features (FE) that best discriminate the classes [25]. The final aim is to provide an almost linearly separable classification problem in the *feature space*.

Several methods have been described based on statistical tests for FS [26], matrix decompositions [27] or even deep learning architectures for FES [28]. Here we perform FE using a popular method in neuroscience: Partial Least Squares (PLS) [27]. PLS methods have demonstrated its utility in describing the relationships between brain activity and experimental design or behavior measures within a multivariate framework (see [10,27,29] and the Appendix A for mathematical details and the interpretation of the PLS-maps as a classical t-test).

4.2. Linear decision functions: a small upper bound

Regularized linear decision functions have been recently applied to neuroimaging for detecting activation patterns, and compared to parametric hypothesis testing, such as univariate t-tests [30–32]. In general, they have limited their analyses to in-sample estimates based on resampling, without demonstrating their out-of-sample performance in terms of confidence intervals.

Recall that the minimization of the left part of inequality (6) can be achieved by decreasing the number of separating functions N given the sample ($\mathbf{Z^n}$). This quantity is decreased by selecting a linear decision function-based classifier in a binary classification problem, as previously described [15,24], etc. The selection of higher order classifiers, e.g. based on multikernel learning [33], is a trade-off between N and the empirical error. After selecting the feature set by FES methods, the concentration inequalities (6) obtained with linear classifiers result in a strong association with a given confidence level whenever the extracted features are significant across ROIs and group comparisons.

Beyond the existing caveats and solutions when using regularization methods in neuroimaging for FS, we adopt the linear support vector machine (SVM) classification algorithm which allows us to tentatively evaluate the worst case of N, that is, $\mathbb{S}_n(\mathcal{G}) \equiv \sup_{\mathbf{X^n} \in \mathbf{Z^n}}(N)$ and to set the following upper bound [15]:

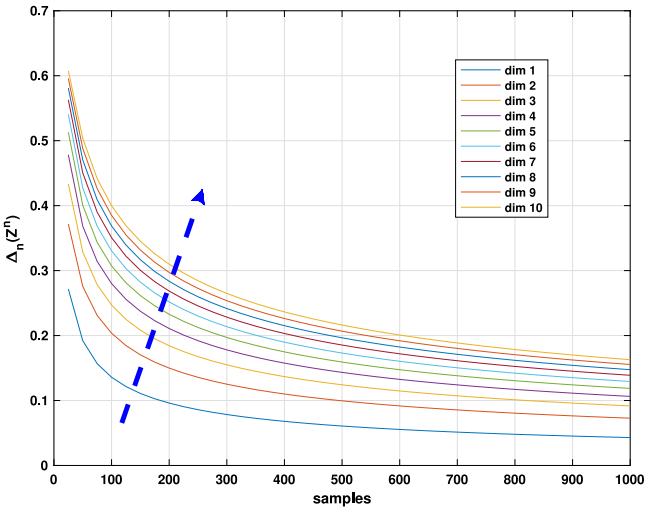


Fig. 1. The upper bound of inequality (9) connecting actual and empirical errors with 95% level of confidence. Note how increasing the feature dimension results in a larger bound (blue-dashed line). However, working in low dimensional scenarios, i.e. $d = 1, 2$, and using medium-sized datasets ($n = 500$), the confidence interval is less than 10%.

$$\Delta_n(\mathbf{Z^n}) \leq \sqrt{\frac{h((\log(2n/h) + 1) - \log(\delta/4))}{n}} \tag{8}$$

with probability $1 - \delta$ and h is the Vapnik–Chervonenkis (VC) dimension, e.g $h = d + 1$ for linear classifiers. In the same manner, several upper bounds could be tested based on several innovative concepts and paradigms, such as those based on data distributions, set shape, Rademacher averages, pseudo-dimension, fat-shattered dimension, etc. [34,35]. We preferred to use, due to its simplicity, the upper bound recently proposed in [11] that is strongly grounded on the geometrical assumption of *in general position* distributed samples and the function-counting theorem of homogeneously linearly separable dichotomies [24]:

$$\Delta_n(\mathbf{Z^n}) \leq \sqrt{\frac{(d-1)\log(n+1) + (2 + \log(1/\delta))}{2n}} \tag{9}$$

where the previous bound is obtained by assessing the number of linear decision functions derived from the latter theorem and then by bounding N as $N = 2\sum_{k=0}^{d-1} \binom{n}{k} \leq 2(n+1)^{d-1}$.

With the help of the inequalities (8) and (9), we can even evaluate the deviation of the empirical error from the actual error at voxel level, although it is preferable, for the aforementioned reasons, to do it region-wise using a fitted linear SVM classifier in the multivariate feature space (see Fig. 1). In this sense, the motivation for a multivariate framework in assessing the areas of relevance is analogous to other proposed techniques for addressing the multiple comparison problem in functional imaging, e.g. Random Field Theory for neuroimaging analysis [36], random/mixed/conjunction analyses in multiple p-value maps [3] or the classical p-value corrections for multiple comparison after null-hypothesis testing. In general, only those voxels (or ROIs) showing a tight association, i.e. high performance in terms of accuracy, should be considered as relevant maps or patterns in that particular condition with probability $1 - \delta$.

5. Summary of the procedure

The following summary of the procedure has been implemented in SAM (middle and right column in Fig. 2):

- **Step 1: Data Preparation and parcellation:** Design the group comparison (design matrix) and select the regions (ROI) to be analyzed across subjects.
  for each ROI do:
- **Step 2: Training feature set**
  - Apply a FES stage, e.g. based on PLS, to the ROI and obtain $\mathbf{Z^n}$ (the feature space)
  - Fit linear SVM by ERM to obtain g_n^* (resubstitution estimate of the actual error)
- **Step 3: Assessment of concentration inequalities:**
  - Compute empirical error (or accuracy) Eqs. (4) and (5).
  - Determine the actual accuracy Acc^j under the worst case with probability $1-\delta$

  end for
- **Step 4: Statistical assessment of the accuracies:** Calculate the z-test statistic for each actual accuracy in $\{Acc^j\}$ Eq. (18) for testing significance.

In Section 6 and in the supplementary material, we will show how the combination of the aforementioned protective intervals and significance tests may be used to derive a SAM in different group comparisons, such as Alzheimer's disease (AD) vs normal controls (NC), Parkinson's disease (PD) vs NC and on a well-known example of single-subject activation map in fMRI, and how they relate with the classical approach based on null-hypothesis testing, i.e. two sample t-test with corrected-p value. Unlike, previous approaches, the proposed model-free method is less specific but more robust against sample size, artifacts and nuisance effects. See the complete diagram of the proposed method in Fig. 2.

6. Experiments

The aim of this section is to present a novel methodology in neuroimaging based on analytical concentration inequalities,[5] and to experimentally compare them to the accepted framework used by the neuroscience community based on the SPM analysis [37,38]. Thus, we will assess several experiments collected from well-known databases that include imaging data from patients with a variety of conditions/pathologies. Nevertheless, we will avoid somewhat related theoretical discussions about the comparison of both branches of statistics, referring the readers to the introduction section in this paper and the vast extant literature addressing these issues [1,6,7,39].

All the datasets were preprocessed using standardized neuroimaging methods and protocols implemented by the SPM software (registration in MNI space by spatial normalization and segmented to differentiate brain tissues, e.g. Gray matter (GM)) [37]. Then, we performed PLS-based feature extraction ($d = 1$) and fitted a linear SVM following the methodology presented in the previous sections.

6.1. A structural MRI (sMRI) study: the ADNI database

Data used in preparation of this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI database contains 1.5 T and 3.0 T t1w MRI scans for AD, Mild Cognitive Impairment (MCI), and cognitively NC which are acquired at multiple time points. Here we only included 1.5T sMRI corresponding to the three different groups of subjects. The original database contained more than 1000 T1-weighted MRI images, comprising 229 NC, 401 MCI (252 stable MCI and 149 progressive MCI) and 188 AD, although for the proposed study, only the first medical examination of each subject is considered, resulting in 818 GM images. Following the recommendation of the National Institute on Aging and the Alzheimer's Association (NIA-AA) for the use of imaging biomarkers [40], we considered the group comparison NC vs. AD for establishing a clear framework for comparing statistical paradigms (SPM and SAM), since the MCI class is strictly based on clinical criteria, without including any other biomarker [41]. Demographic data of subjects in the database is summarized in Table 1.

Table 1
Demographics details of the datasets (ADNI, PPMI, VV and fMRI), with group means with their standard deviation.

	Status	Number	Age	Gender (M/F)	MMSE
MRI					
ADNI	NC	229	75.97 ± 5.0	119/110	29.00 ± 1.0
	MCI	401	74.85 ± 7.4	258/143	27.01 ± 1.8
	AD	188	75.36 ± 7.5	99/89	23.28 ± 2.0
SPECT					
PPMI	NC	194	53.02 ± 2.27	129/65	–
	PD	168	53.14 ± 2.37	103/65	–
SPECT					
VV	NC	108	69.05 ± 14.53	54/54	–
	PS	100	68.62 ± 13.41	53/47	–
fMRI					
Auditory	Res	41	–	1	–
	List	43	–	1	–

6.1.1. Static classification results

First, the proposed methodology try to fit in an optimal way a linear SVM classifier in the feature space obtained after a FES approach (PLS). With the aim of applying a regression-type analysis to the dataset, we parcellated the brain volume into 116 standardized regions [22] and then, obtained an optimistic estimation of the actual error $P(g_n)$ as shown in solid blue line in Fig. 3. This estimation is corrected by the use of upper bounds drawing a novel set of accuracy values (proportions) and a confidence interval, depending on the selected theoretical method, i.e. Vapnik's bound. The lower accuracies in this plot corresponds to the worst cases as considered by the selected concentration inequalities.

It is worth mentioning that the results, shown in Fig. 3, are obtained with the first PLS component extracted by this regression analysis ($d = 1$). This PLS score for each subject can be conceptualized as the representation of the subject into a multi-dimensional reference system as described in [10] (see supplementary material for the analysis in higher dimensions).

6.1.2. Statistical agnostic maps

In Fig. 3 we heuristically identified all those relevant regions for the characterization of AD based on absolute values using the complete ADNI dataset. Therefore, a definition of relevancy in terms of hypothesis testing within confidence intervals is required. Following the method presented in the Appendix A, we provide an automatic (and statistical) method for selecting ROIs in which a regionally specific activation is identified. As depicted in Fig. 4, the main result is the SAM obtained with the same p-value as the one of confidence intervals using concentration inequalities.

For further comparison with the SAM proposed in this paper, significance maps were obtained with SPM (voxel-wise inference) using a standard two-sample t-test with FWE p-value = 0.05 (and null extent threshold -voxels-).[6] For SPM we first conducted a first-level analysis to derive the GLM for the dataset under assessment (a design matrix for group comparisons) and then, in the 2nd-level analysis, the contrast

[5] please visit https://github.com/SiPBA to download a preliminary version of the software in Matlab.

[6] For a full analysis based on cluster and voxelwise inferences using the complete dataset please see the supplementary material.

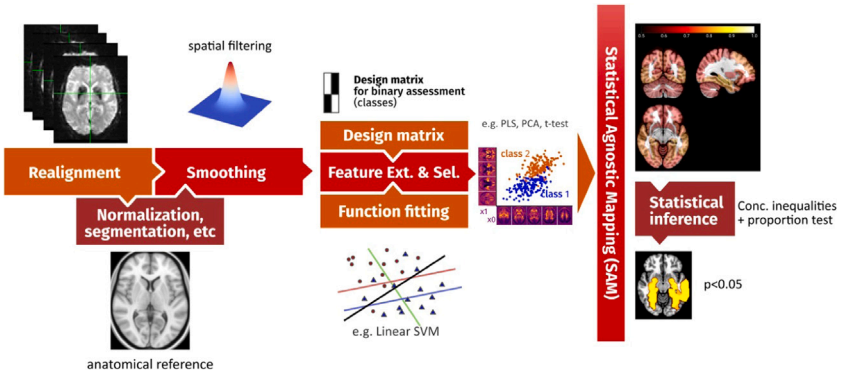


Fig. 2. Complete diagram of the proposed methodology including typical preprocessing steps in SPM for different modalities (left column of blocks), classification fitting and FES for actual risk estimation (middle column) and inference to derive the SAM (right column).

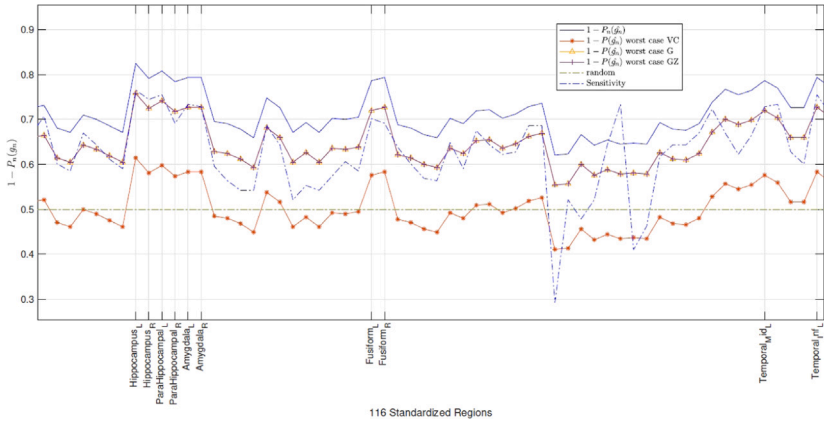


Fig. 3. Accuracy values and upper bounds in standardized ROIs (only significant regions from #30 to #90 are shown) for three methods based on concentration inequalities in (4) and (6). We highlight several regions, relevant in the biological definition of AD, i.e. Hippocampus, Temporal, Amygdala and Parahippocampal regions, corresponding to peaks of these curves. Moreover, observe how the VC approach is more pessimistic than the one based on [11]. The confidence interval is drawn in the space between the solid blue line and the colored lines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

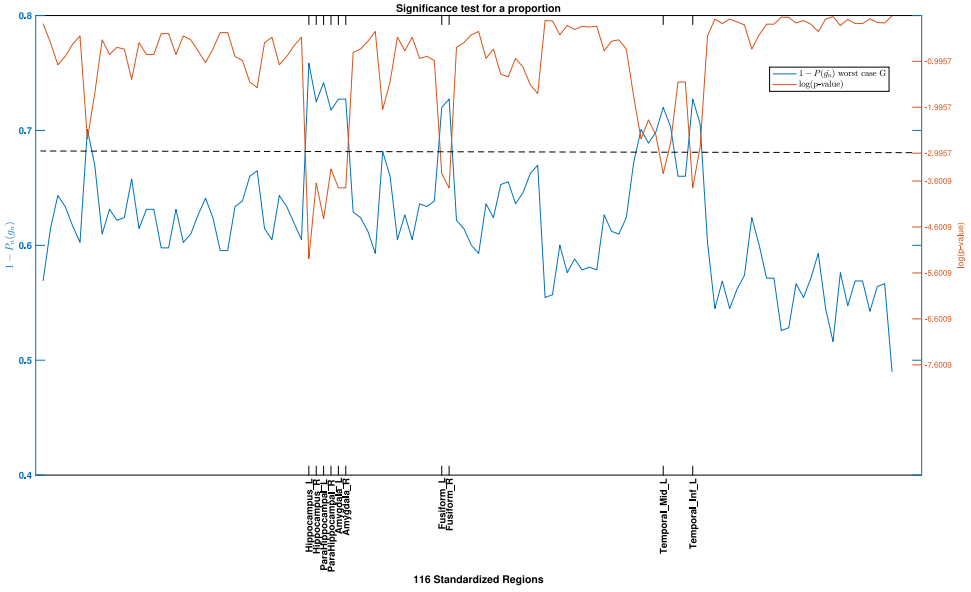


Fig. 4. Accuracy values in the worst case using the method in [11] and the set of probabilities (log(p-values)) within the confidence interval. The ROIs ($p < 0.05$) are detected out of 116 standardized regions using a significance test for a proportion π (see Appendix A). Note that we show the probability of observation (in the right "y axis") of the set of accuracy values under H_0, i.e. random distribution.

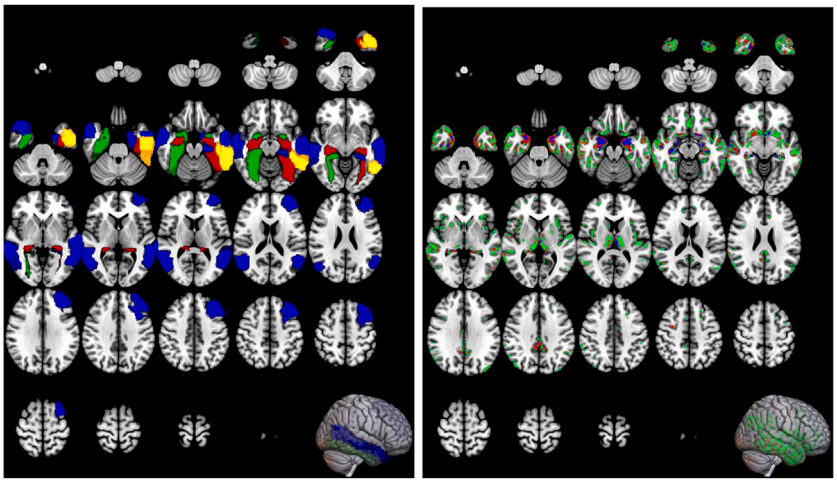


Fig. 5. Statistical comparison of brain volumes using SAM (left) and SPM (right) in the ADNI database. Green area corresponds to the whole dataset while the rest of colors (red, blue, yellow) are linked to data subsets, which are plotted in increasing n (opacity of representations is preserved for clarity reasons). The ROIs selected for increase $n = 50, 100, 200, 417$, satisfy $S_j \subset S_{j+1}$ except for $n = 50$ where an additional region "Frontal Mid L" is selected. It is worth mentioning that all the ROIs extracted in different sample-size configurations were included in the confidence interval and with probability slightly higher than the significance level ($\alpha = 0.05$). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

images were fed into a GLM for implementing the statistical test. A direct comparison with the SPM approach is shown in Figs. 5 and 6, in terms of the sample-size analysis and the relevant regions determined by both methods. Key to this comparison is the different working operations, i.e. SAM includes the spatial structure of data at the first FES stage, whilst SPM do it at the final stage, by means of RFT. For this reason, SPM is more specific (voxel-wise) but widespread comparing to SAM. The number of identified ROIs conforming the SPM increases as the number of sample increases, unlike the proposed approach, which provides the same volumetric differences for $n = 200, 417$. It is worth mentioning that from the perspective of SLT, due the small ratio n/d in all these experiments proposed in this paper (and in the extant literature), we are dealing with the "small sample size problem". In terms of classical statistics (SPM) this derives in a challenging scenario that constrains the generalization of the results from small datasets to new unseen samples.

Fig. 6 shows that main regions identified by SPM are included in the ROIs deployed by SAM-based approach. In addition, the number of "activated" voxels in SPM is associated with sample size and these voxels are widespread across several anatomical regions. The number of voxels in ROIs obtained by SAM is almost independent on the sample size, except for the extreme case $n = 50$, and given the magnitude of the effect being sought in the NCvsAD comparison.

6.1.3. Is SAM dependent on the a-priori specified atlas?

The SAM approach proposed with ML also relies on an a priori atlas [22], that divides the brain volume into an arbitrary set of ROI, as shown in Section 3. In this section we investigate the robustness of SAM against the number of predictors (ROI dimensions). The standardized regions considered in this paper have different sizes, from hundreds to thousands of voxels. They closely align with anatomical structures traditionally used in the medical literature that, unlike voxel-wise approaches, have demonstrated their value for supporting the diagnosis of neurological diseases. To evaluate the impact of the parcellation scheme on SAMs performance, we additionally parcellated each ROI within the former atlas using a simple cluster approach (k-means approach) on the voxels coordinates (x, y, z) that resulted in 232 ($K = 2$), 348 ($K = 3$) and 464 ($K = 4$) regions. Fig. 7 shows an example of the new generated atlas for K = 4 on the left (464 regions), and the SAMs obtained with the new configurations on the right, using the same experimental setup as detailed in Section 6.1.2.

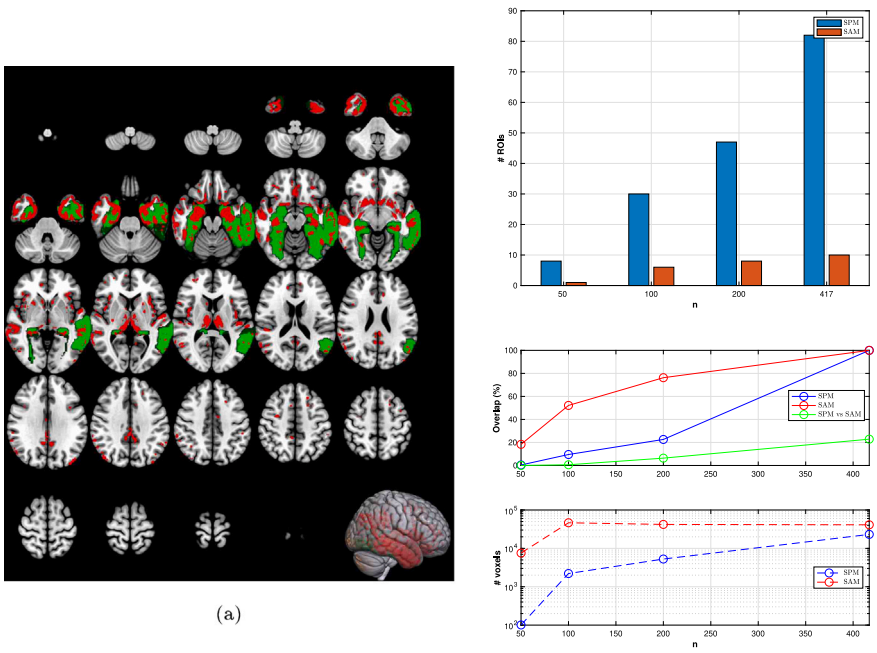
In summary, similar to voxel-wise SPM, this method improves its sensitivity when the number of ROIs increases, but it could increase the number of False Positive (FP) regions, as shown in Fig. 7(B). However, in general, the regions were very similar across all different configurations. although another interesting question arises in relation to the control of FPs of the SAM approach, that will be discussed in the next section.

6.1.4. Controlling the FWE rate in a small-effect dataset

Moreover, we should analyze the ability of the proposed method for controlling the FWE rates for voxel, clusterwise or regionwise inference as shown in [42]. To this purpose two groups of subjects ($N = 50, 100, 228$) are randomly drawn from a relatively large ($N = 228$) group of NCs from the ADNI dataset, where the null hypothesis of no group difference in brain activation should be true.

A total of 48k random group analyses were performed, following the same steps as in the previous section (parcellation, FES, function fitting, etc.) to compute the empirical false-positive rates of SAM and SPM (in cluster and voxel-wise inferences). Regarding the SPM study, two types of two-sample t-test-based inferences were performed using FWE and uncorrected p-value. Each statistic map was first thresholded using a CDT of $P = 0.001$ (uncorrected for multiple comparisons). The degree of FP to compute the FWE rate was finally estimated as the number of significant voxels within any of the 116 atlas regions, meaning a voxelwise inference. Furthermore, a conservative clusterwise inference is applied by using uncorrected p-values, where the surviving clusters are then compared with a cluster extent threshold-based criteria in regions (at least 25% of activated voxels in any of the 116 regions of the atlas). The estimated FWE rates are simply the number of analyses with a significant results divided by the number of analyses (1.000).

Fig. 8 illustrates similar results as those described in [42] that were obtained using a conservative voxelwise inference and invalid clusterwise inference for the two-sample t test used in these simulations. SAM provided a conservative operation mode when the Vapnik's bound (see Eq. (8)) was applied to the empirical data (often falling below the significance level, e.g. 5%). It also performs a very realistic and competitive approach when the estimation of the upper bound in Eq. (9) is used, as shown in [11]. In particular, FWE rates for clusterwise inference far exceed their nominal level using SPM (using a CDT of $P = 0.001$ uncorrected for multiple comparisons), despite the surviving clusters were then compared with an cluster-extent threshold based criterion of 25% of activated voxels in that region for a fair comparison with SAM. On the other hand, parametric voxelwise (SPM) and regionwise (SAM, e.g. Eq. (8)) based inferences are valid but over



(a) (b)

Fig. 6. (a) SPM (red) over SAM (green) using the complete ADNI dataset ($n = 417$). (b) overlap analysis vs sample size. Observe how the SPM activation map linearly increases with n and is located on more than 80 standardized regions with the whole dataset (although part of these isolated activation voxels could be removed from the map using the extent threshold). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

conservative, often falling below the predefined levels of significance, $\alpha = 0.001; 0.01; 0.05; 0.1$ (see Fig. 8). However, estimated FWE rates based on corrections described in [11] are close to the predefined levels, and are almost independent on the number of subjects that were randomly drawn in the simulation. Finally, the true positive (TP) rate[7] (i.e. the null hypothesis of no group differences in brain activation should be false) may also be assessed on balanced groups of subjects ($N = 50, 100, 150, 300$) that are randomly drawn from a relatively large ($N = 229 + 188$) group of NC and AD from the ADNI dataset (making a total of 64K analyses). At a given significance level, the TP rate should be almost constant for different sample sizes (i.e. the method provides the same number of significant regions in each experiment). As clearly shown in the bottom right panel of Fig. 8, the SAM methodology provides almost constant TP rate with different sample sizes, unlike SPM TP rate in cluster and voxel inferences, which clearly increases with sample size.

7. Discussion

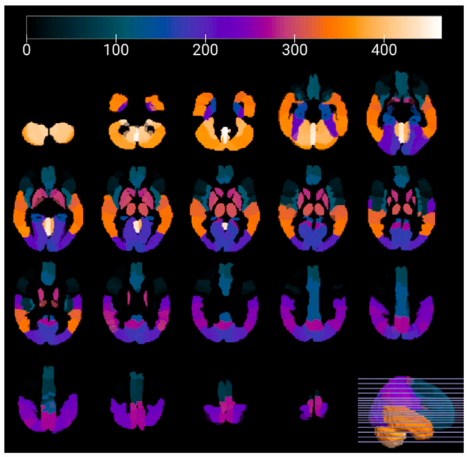
As shown in the latter section, in general the SAM is a very robust method, in terms of sample size, to find relevant standardized areas, and a stable framework which contains those regions defined as relevant by the SPM, with sufficiently large sample size. It is worth mentioning that SAM employs the concept of prevalence [4] to derive the activation maps, since it is the result of the classification performance in terms of accuracies or proportions in the feature space. The experiments carried out in different experimental frameworks and datasets have demonstrated the ability of this multivariate approach for establishing a novel model-free method for the assessment of significant changes across brain volumes.

The behavior of the analyzed methods depends on the size of effect we are interested in. In the seek of subtle effects, such as the ones found in AD or Autistic patterns, and provided that hypothesis tests cannot separate important, but subtle, and actually trivial effects [6], our SAM focus on standardized ROIs to avoid the presence of false positives in the sought maps. In this sense, SPM is more specific and can detect, within these regions sought by SAM, which substructures are responsible for the discrimination between classes. Nevertheless, this voxel-wise analysis could be carried out as well using this framework, e.g. by assessing the PLS-maps derived at the FES stage as shown in [10]. However, we additionally found that using the SPM univariate approach (i) small effect sizes in a heterogeneous population with a limited sample size fail to be detected whilst with larger samples sizes their detection overshoots, and (ii) in large sample sizes it can yield highly significant p values even when effect sizes are so small that they become trivial in practical terms, e.g. the SPECT experiments, similar to the findings in [43].

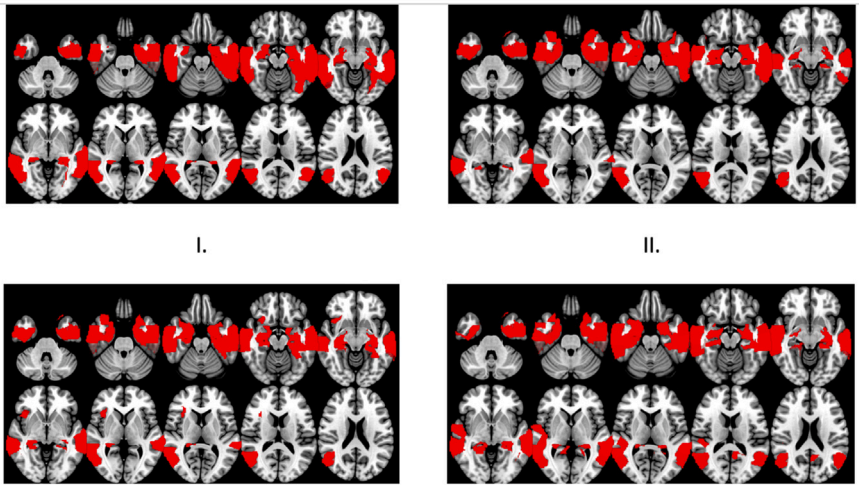
On the other hand, when large effects are bound to be found, SAM is a suitable method in their detection since, with a few amounts of samples, it provides similar results than the ones obtained with complete databases, i.e. auditory fMRI or DatSCAN PD imaging experiments. This is in line with the main idea derived from [39] that when an effect is found in small datasets is more than likely to be extrapolated in large samples. On the contrary, only in small datasets with small – but meaningful – effects that are missed, missing data, sampling bias, etc. we found the absence of replication, i.e. across data collecting sites [6]. All these statistical features in the analysis of neuroimaging data are experimentally described in the datasets analyzed throughout this paper.

At this stage of development, SAM is only aimed at providing an alternative method for the univariate SPM based on two-sample t-test in groups comparison with predefined classes. The method could be adapted to manage other kind of contrasts or factorial analyses by the multiple developments of ML and FES methods in the last years [44]. This could include encoding and decoding applications based on supervised learning to link brain images with stimuli [45] or Error-Correcting Output Codes(ECOC) that were designed to solve the multi-class classification problem [46].

[7] The estimated TP rates are simply the probability of a region to be activated, e.g in voxelwise SPM $P_{TP} = \frac{\#TP}{116}$.



(a)



# ROIs\ROIs	Insula_R'	'Hippocampus_L'	'Hippocampus_R'	'ParaHippocampal_L'	ParaHippocampal_R'	'Amygdala_L'	'Amygdala_R'	'Lingual_R'	'Fusiform_L'
116		x	x	x		x	x		x
232		x	x	x	x	x	x		x
348	x	x	x	x	x	x	x		x
464		x	x	x	x	x	x	x	x

# ROIs\ROIs	Fusiform_R'	'Precuneus_L'	'Temporal_Sup_R'	'Temporal_Pole_Sup_L'	Temporal_Pole_Sup_R'	'Temporal_Mid_L'	'Temporal_Mid_R'	Temporal_Inf_l	'Temporal_Inf_R'
116						x	x	x	x
232	x			x	x	x	x	x	x
348	x			x	x	x	x	x	x
464	x	x	x	x	x	x	x	x	x

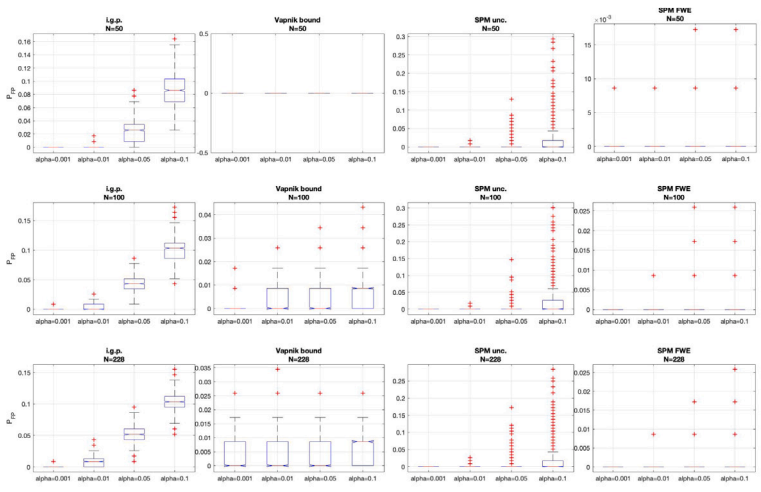
(b)

Fig. 7. (a) Example of atlas derived using k-means ($K = 4$) clustering from the standardized 116-region atlas. (b) SAMs obtained for different configurations in the number of clusters, I: baseline; II: $K = 2$; III: $K = 3$; IV: $K = 4$; and the corresponding relevant areas at the bottom table.

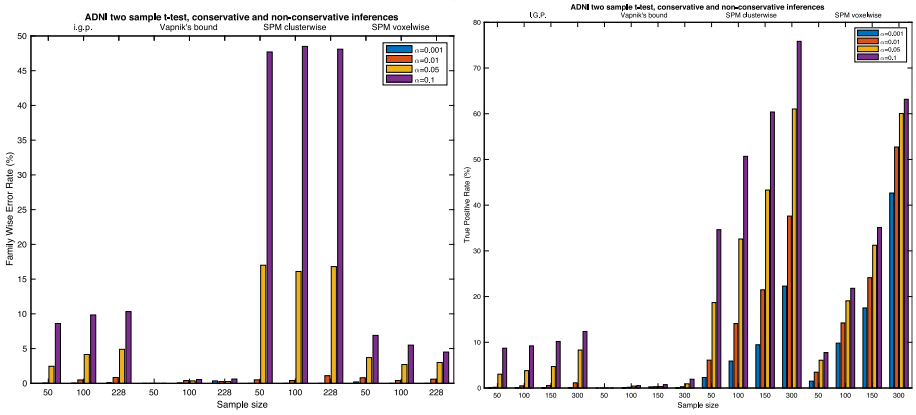
Finally, we have seen the usefulness of the confidence intervals derived for the STL based on concentration inequalities to achieve a confidence framework beyond sharp null-hypothesis testing. Key to this methodology in the field of SLT is that it is based on in-sample estimates (a similar procedure in exploratory analysis using hypothesis testing), unlike the out-sample estimates in CV procedures, which usually subdivide the (small) datasets for an estimation of the actual error. In this way, an analytical bound depending on sample size (n) and number of predictors (d) defines a "worst-case" operation point. Nevertheless, the experiments showed the application of a systematic hypothesis test for the selection of significant empirical errors which conforms the highlighted regions in the SAM. Only in this case, a model is assumed in the set of accuracies, but it has been demonstrated to be in accordance with the nature of the one-dimensional data and sufficiently accurate for our purposes.

8. Conclusion

In this paper we present a data-driven approach, mainly devoted to classification problems with limited sample sizes, to derive statistical model-free (agnostic) mappings. Although the latter is *not designed for testing competing hypothesis or comparing different models* in neuroimaging, we derive the SAM assuming the existence of classes (H_1), at voxel or multi-voxel level. The analysis of the "worst case' considers the upper bounds of the actual risk, under suitable theoretical conditions (see methods and Appendix A) and a selection of regions with a highly-corrected empirical risk, according with a test for significance on a population proportion. As a conclusion, the SAM relieved the problem of instability in limited sample sizes, when determining maps of relevance in several neurological conditions, such as AD, PD or auditory tasks, and resulted in a very competitive and complementary



(a)



(b)

Fig. 8. Results for two-sample t test and ad-hoc clusterwise/voxel inference in regions, showing estimated FWE rates and TP rates for four different activity paradigms (FWE SPM, uncor. SPM, Vapnik SAM, and i.g.p. SAM). Up: Distribution of FPs found in experiments; Bottom: on the left FWE rate, on the right experimental TP rate. These results were generated using the ADNI data and $\{50; 100; 228\}$ subjects in each group analysis for FWE rate and ($N = 50, 100, 150, 300$) for TP rate. Note: SPM unc: cluster-wise inference, SPM-FWE: voxel-wise inference.

method with the SPM framework, which is mainly accepted by the neuroimaging community. Moreover, the latter usually employs several strategies for reducing the false positive rates in multiple comparisons, such as the (FWE) corrected p-value maps in inferential statistics null-hypothesis testing, and RFT to tackle with the spatial structure of the maps. However, this approach is found to be very conservative in our experiments and in the extant literature to control the FWE rate. In a nutshell, the novel framework based on SLT provides similar activation maps than the ones obtained by the voxel-wise SPM, *but defined on ROIs*, under a rigorous development in scenarios with a small sample/dimension ratio and large, small and trivial effect sizes, as shown in the experimental part.

Acknowledgments

This work was partly supported by the MINECO/ FEDER, Spain under the RTI2018-098913-B100, CV20-45250 and A-TIC-080-UGR18 projects and by the Ministerio de Universidades, Spain under the FPU Predoctoral Grant FPU 18/04902. We would like to thank the reviewers for their thoughtful comments and efforts towards improving our manuscript. J.M. Gorriz would like to thank Dr. Maxim Raginsky for his elegant abstract notation which is borrowed in this paper.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI), USA (National Institutes of Health Grant U01 AG024904) and DOD ADNI, Spain (Department of Defense award number W81- XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of

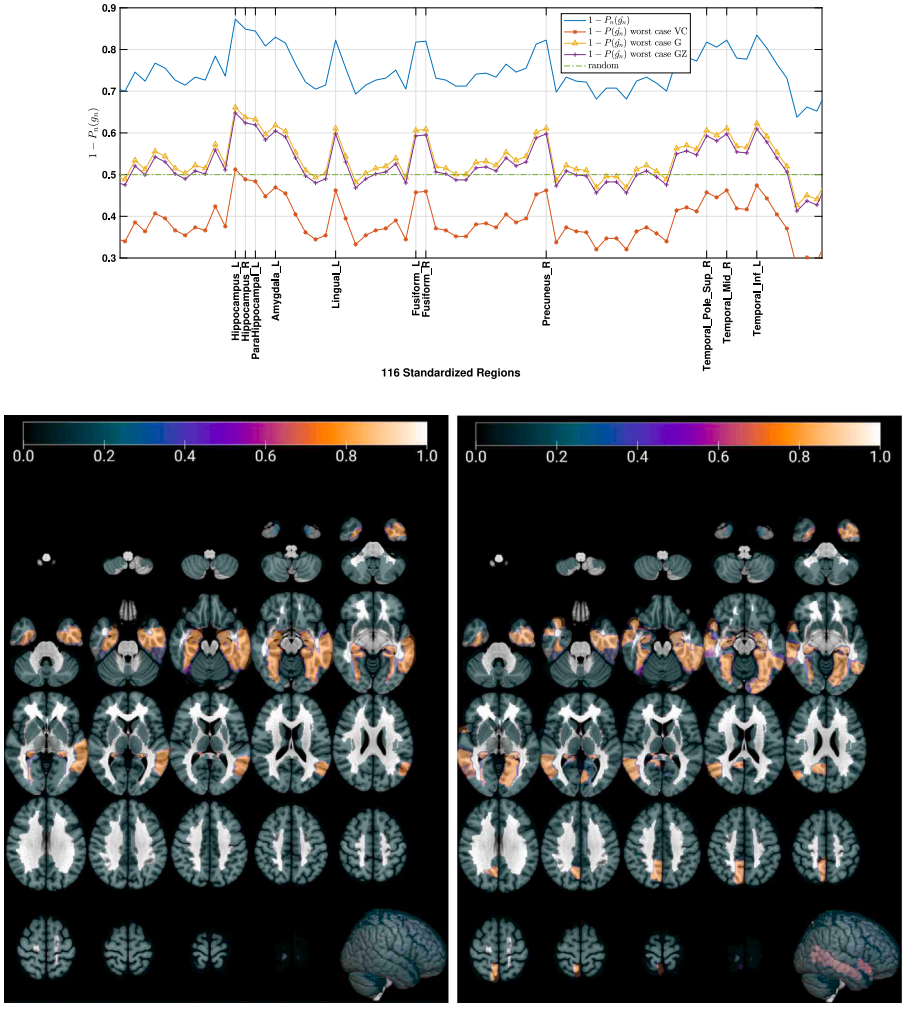


Fig. 9. The same analysis as the one proposed in Fig. 3 but with $d = N_{comp} = 8$. Note that the ROIs showing highest accuracy values (bottom on the right) are similar to the ones selected in the latter experiment (bottom on the left) but with an increase in the confidence interval of the approximation. Examples are Hippocampal, Hippocampus, Amygdala and Temporal regions (see Section 6.1).

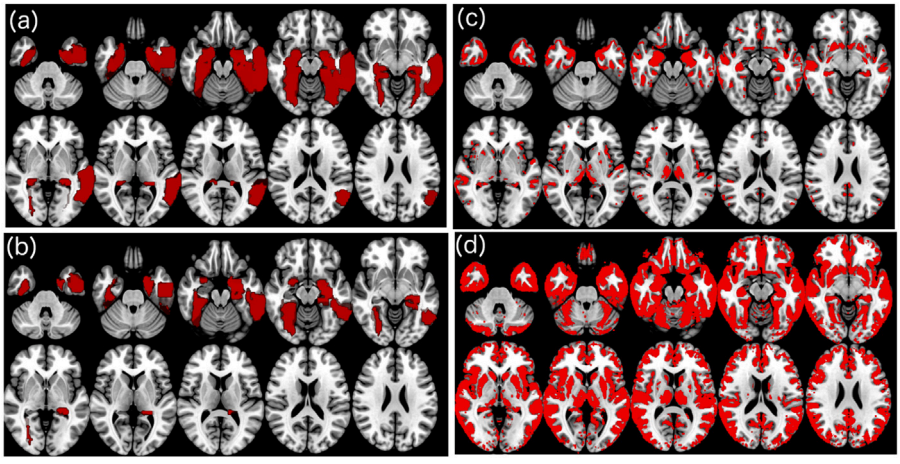


Fig. 10. ADNI SAM and SPM for conservative and non-conservative inferences: (a) i.g.p (b) Vapnik's (c) voxelwise (d) clusterwise.

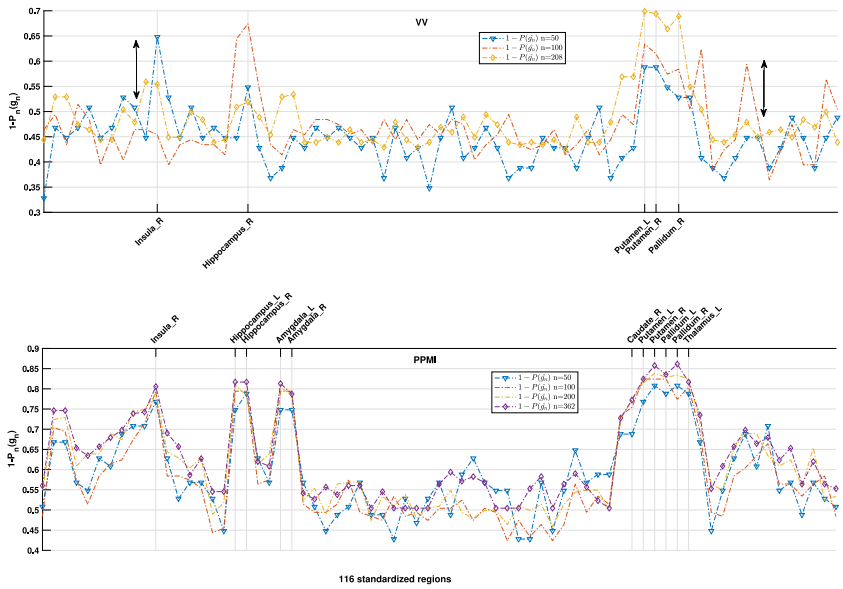


Fig. 11. Effect sizes and significant ROI selection using the proposed methodology. Observe in top figure (VV) the black arrows highlighting the observed trivial effects, outside the specific region, in studies with low statistical power, i.e. $n = 50$. In addition, we also remark the reduction of "true effect" in the top figure compared to the same effect depicted in the bottom figure (PPMI), due to the presence of a more complex PD-plus pattern in the VV dataset.

Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Appendix A

A.1. Upper bounding the worst case: a summary

A shown in Eq. (4) the consistency of the ERM algorithm is mainly dependent on the evaluation of the two-sided uniform deviation of the error probabilities in the worst case. An upper bound with probability at least $1-\delta$ for this quantity can be obtained by invoking a result in [19], since $\Delta(\mathbf{Z^n})$ has the bounded differences property by $1/n$:

$$\mathbb{P}\left(\Delta(\mathbf{Z^n}) \geq \mathbb{E}[\Delta(\mathbf{Z^n})] + t\right) \leq e^{-2nt^2} \equiv \delta; \quad \text{for any} \quad t > 0 \tag{10}$$

This is known as the generalized Hoeffding inequation. Then.,[8]

$$\Delta(\mathbf{Z^n}) \leq \mathbb{E}[\Delta(\mathbf{Z^n})] + \sqrt{\frac{\log(1/\delta)}{2n}} \tag{11}$$

with probability $1-\delta$.

Moreover, the expected value of the deviation $\mathbb{E}[\Delta(\mathbf{Z^n})]$ can be absolutely bounded by the so-called Rademacher average [47] as follows. First, the uniform deviation is bounded by its expected value w.r.t the set of random error functions g, using the "symmetrization" trick proposed in [15] and the convexity property of the norm function:

$$\Delta_n(\mathbf{Z^n}) = \sup_{g\in\mathcal{G}} |P_n(g) - P(g)| \leq \mathbb{E}_{g'}[\sup_{g\in\mathcal{G}} |P_n(g) - P_n(g')|] \tag{12}$$

where g' is randomly selected from $\mathcal{G}$ and $\mathbb{E}_{g'}[P_n(g')] = P(g)$. Taking whole expectations on the both sides we get:

$$\mathbb{E}[\Delta_n(\mathbf{Z^n})] \leq \mathbb{E}[\sup_{g\in\mathcal{G}} |P_n(g) - P_n(g')|] \tag{13}$$

By using the triangle inequality and the definition of empirical error we finally obtain:

$$\mathbb{E}[\sup_{g\in\mathcal{G}} |P_n(g) - P_n(g')|] \leq 2\mathbb{E}[\sup_{g\in\mathcal{G}} |\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{Z_i})|] \tag{14}$$

where the right part of inequality is equally distributed as the Rademacher average $\mathcal{R}(\mathcal{G}(\mathbb{Z}^n)) \equiv \mathbb{E}_\sigma \sup_{g\in\mathcal{G}} |\frac{1}{n}\sum_{i=1}^{n} \sigma_i g_i|$, where σ_i are independent random variables in $\{\pm 1\}$ with equal probability. Finally, using Massart's finite class lemma [18] we can bound the left part of the latter inequality as:

$$\mathbb{E}\mathcal{R}(\mathcal{G}(\mathbb{Z}^n)) \leq 2\sqrt{\frac{\log(N)}{n}} \tag{15}$$

Consequently, introducing Eqs. (11), (13), (14) and (15) in Eq. (4) we finally prove Eq. (6).

A.2. The partial least squares algorithm

The PLS algorithm extracts the relevant patterns within ROIs across brains by a regression between the $n\times d$ multivariate data matrix $\mathbf{X}$ and the $n\times 1$ label vector $\mathbf{Y}$. In short, we maximize:

$$\boldsymbol{\omega}_o = \max_{\omega} (cov(\mathbf{X}\boldsymbol{\omega}, \mathbf{Y}))^2; \quad \text{s.t. } \|\omega\| = 1 \tag{16}$$

where the score vectors $\mathbf{s} = \mathbf{X}\boldsymbol{\omega}$ are iteratively extracted and used to deflate the input matrix $\mathbf{X}$ by subtracting their rank-one approximations based on $\mathbf{s}$ [29]. The *deflation process* is accomplished by the computation of the vector of loadings $\mathbf{p}$ as a coefficient of regressing $\mathbf{X}$ on $\mathbf{s}$:

$$\mathbf{p} = \frac{\mathbf{X}^T\mathbf{s}}{\mathbf{s}^T\mathbf{s}} = \mathbf{X}^T\hat{\mathbf{s}} \tag{17}$$

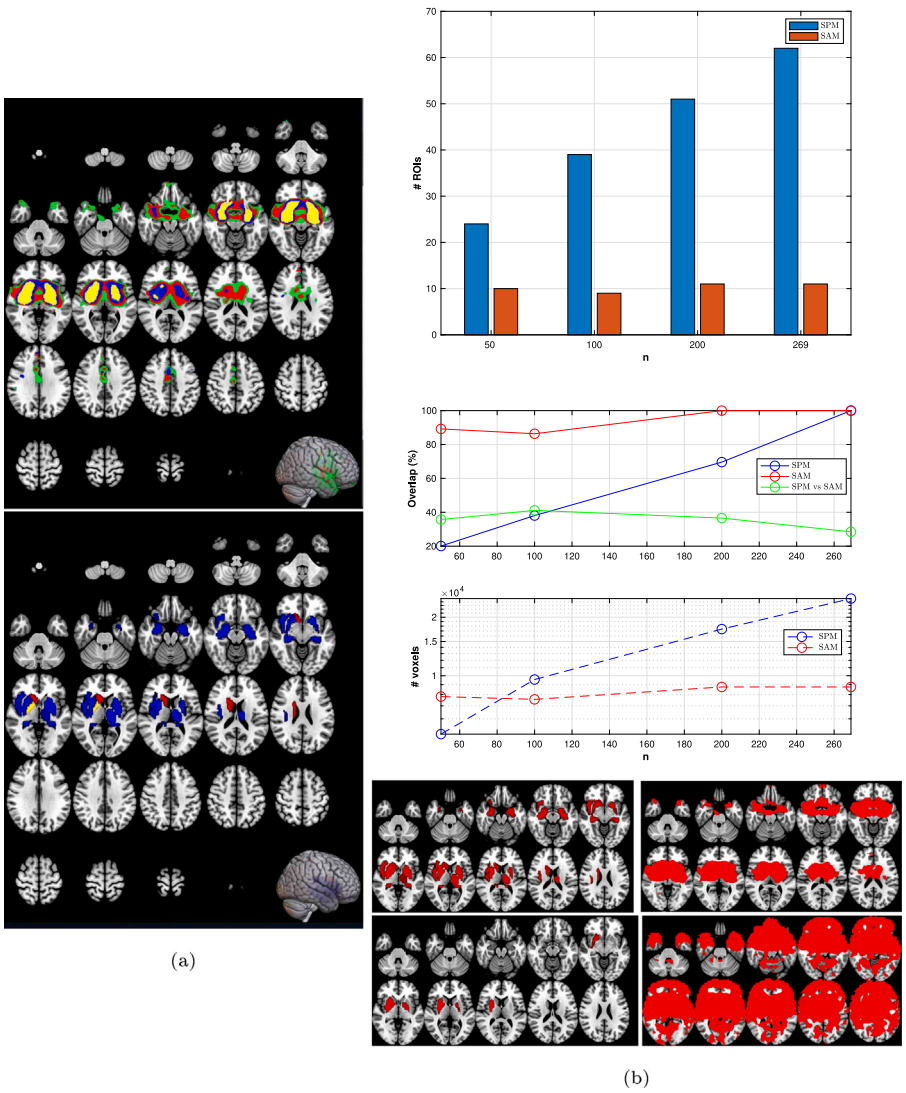
As shown in [11] the size of the input data d is crucial to the assessment of the relationship volume data and group membership within the evaluated ROIs, where some statistical properties of the involved processes, such as the stationarity or the ergodicity in the correlation, must be assumed. The PLS-maps derived can be seen as a multivariate two-sample test weighted by the scores of each sample with unknown distribution, except for a normalization term that depends on the pooled standard deviation [11], thus its statistical significance can be assessed in a similar manner of a t-test [27].

A.3. Significance test for a proportion

Let denote $\hat{\pi}$ the sampling distribution of empirical errors $P_n^i(g_n)$, for $i = 1,\ldots,l$, then the null hypothesis test about the population proportion within the confidence interval has the form:

$$H_0 : \pi = \pi_0 \quad ; \quad H_1 : \pi > \pi_0$$

[8] Given x a random variable, if $P(x > \epsilon) \leq \eta$ then $P(x < \epsilon) = 1 - \eta$.



(a)

(b)

Fig. 12. (a) SPM (up) and SAM (bottom) using the PPMI dataset for $n = 50, 100, 200$ and 269 (red, blue yellow, green). (b) Up and middle: overlap analysis vs sample size. Observe how the SPM activation map linearly increases with n and is located on more than 60 standardized regions with the whole dataset. Typical effects, such as PVE, in this kind of low-resolution image modality results in rejecting the null-hypothesis although FWE corrected p-values were considered in the inference test. On the contrary, SLT is less specific but more stable in the rejection of the null-hypothesis. In addition, the ROIs obtained are overlapped more than 80%, using a wide range of small sample sizes. Bottom: A comparison between all the inferences, i.g.p., Vapnik's, voxel and clusterwise inferences. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where π_0 denotes a particular proportion value between 0 and 1, i.e. 0.5. The test-statistic in a population proportion is:

$$z = \frac{\pi \hat{-} \pi_0}{\sigma_0} \tag{18}$$

where $\sigma_0 = \sqrt{(\pi_0(1-\pi_0))/l}$. For large samples, i.e. for $\pi_0 = 0.5$ at least $l = 20$, if H_0 is true, the sampling distribution of the z test statistic is the standard normal distribution.

Appendix B. Supplementary material

B.1. ADNI in a higher dimension and inference

A similar analysis was carried out with increasing number of components, i.e. $N_{comp} = 2, 3, \ldots$, however the upper bounds are increased accordingly as shown in Fig. 9. This highlights the benefits of working in a low dimensional scenario, $d = 1$, although the use of new features in the analysis allows us to detect other regions, such as "Temporal Pole Sup" and "Temporal Mid" regions.

The analysis described in Section 6.1.2 included only the more competitive approaches for SPM and SAM, that is voxelwise and i.g.p. based approaches. In the following Fig. 10 we show all the configurations in SAM and SPM, using the complete ADNI dataset for evaluation purposes, following the experimental setup previously described in the latter section.

B.2. A SPECT study: the PPMI database

Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org PPMI is a public–private partnership funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including all partners listed on www.ppmi-info.org/fundingpartners.

Informed consents to clinical testing and neuroimaging prior to participation of the PPMI cohort were obtained, approved by the institutional review boards (IRB) of all participating institutions. The PPMI obtained written informed consent from all study participants before

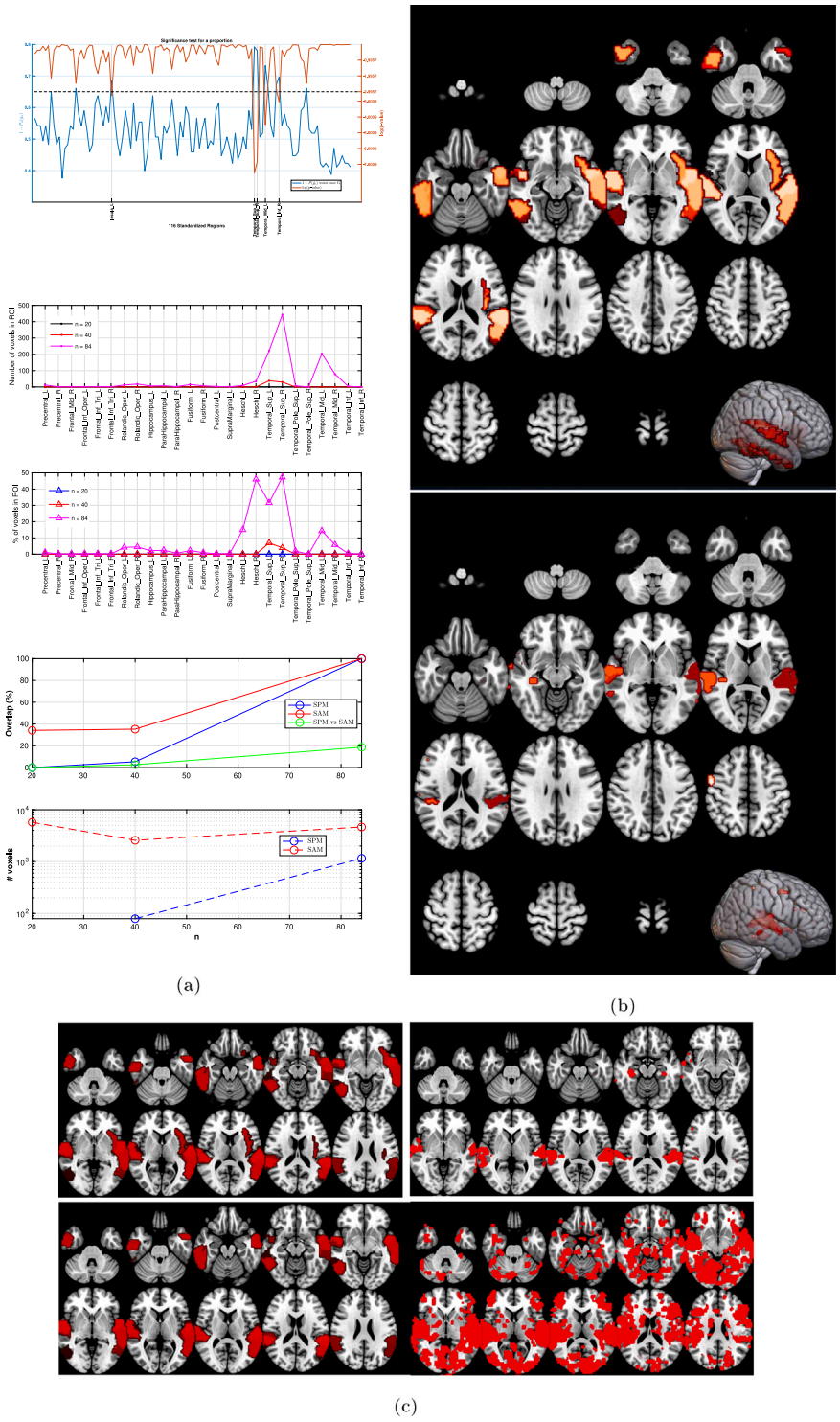


Fig. 13. (a) Significant tests for a population proportion in the fMRI experiment (up), number of activated voxels in ROIs using SPM (middle) and overlap analysis between SPM and SAM (bottom) (b) Activation maps in the auditory experiment for the whole dataset using the SAM (up) and the SPM (bottom) (c) A comparison between all the inferences, i.g.p., Vapnik's, voxel and clusterwise inferences.

enrolled in the Initiative. None of the participants were taking any PD medication when they enrolled in the PPMI.

The inclusion criteria adopted in the PPMI cohort study are available in http://www.ppmi-info.org/wp-content/uploads/2014/06/PPMI-Amendment-8-Protocol.pdf. This diagnostic procedure also includes a confirmation step based on imaging biomarkers. A selection of $N = 269$ DaTSCAN images from this database were used in the preparation of the article. Specifically, the baseline acquisition from 158 subjects suffering from PD and 111 NC was used. In addition, a similar SPECT image database from a the "Virgen de la Victoria" (VV) Hospital (Malaga, Spain) was used to validate and generalize our findings to a dataset that contains a more complex pattern in the Parkinsonian Syndrome (PS) class derived from a clinical diagnosis criteria [48] (see Table 1).

B.2.1. Effect size in static classification

Following the methodology presented in Section 6.1 we will show (i) the robustness of the proposed methodology in limited sample sizes regarding effect size and (ii) a quantitative interpretation of effect size appealing to image classification in diagnostics. As already commented in [49], studies with low statistical power require large effects to be observed by hypothesis testing with a pre-specified p-value threshold (typically 0.05). In DatSCAN imaging of PD the true effect size is known to be considerably large on specific regions, e.g. striatum. On the contrary, large effects observed in studies with reduced sizes do not assure that the true effect is large, or even that it exist at all. These studies are usually related to poorly mechanistically grounded hypothesis [39] or a bad specification of clinical analysis plans to conform the set of observations, i.e. dataset [49].

These issues are can be observed in Fig. 11, where accuracy values are shown for increasing $n = 50, 100, \ldots$. Effect sizes are large when they can discriminate between subjects that do and do not show an effect [39]. Large (but trivial under our hypothesis PDvsHC) effects observed for n = 50,100 samples reduce as the sample size increases in the VV dataset, unlike in the PPMI dataset. In the latter dataset, the proposed methodology provides almost the same accuracy values, which are, in general, shifted up w.r.t the former database, for a wide range of samples sizes of randomly selected subjects. Anyway, our method reports effect sizes (in terms of accuracy values) and confidence intervals alongside exact p values, thus improving the strength of inference.

B.2.2. Statistical agnostic maps

Compared with the subtle effects in the ADNI dataset, the magnitude of the effect in this study is relatively large. Thus, maps of significance derived from both approaches should be similar each other in the specific regions. However, this image technology has associated important challenges, such as low resolution empowering partial volume effects (PVE) [50] and lack of structural information in the images to perform an accurate spatial normalization and co-registration, [48]. These issues could reveal the limitations of voxel-wise approaches using sharp null hypothesis tests, which may find small effects that are practically unimportant. All these questions are found in Fig. 12 where we show how SAM are stable several sample sizes and included in the regions detected by the SPM approach. Moreover, we see how the number of voxels in the classical approach is dramatically rising with increasing n, due to the fact that large studies are more likely to find a significant difference for a persistent trivial effect that is not really meaningfully different from the null [51].

B.3. A functional MRI study

Data used in the preparation of this article was obtained from the SPM database related to an epoch auditory fMRI activation data.[9] This database is one of several databases included in the SPM site[10] for personal education and evaluation purposes, and shows the ability of the SPM methodology for detecting auditory stimulation maps. Specifically, the experiment associated with the data was conducted by the FIL methods group and was designed for exploring equipment and techniques related to fMRI.

The database consists of BOLD/EPI images obtained from a single subject. They were acquired on a modified 2T Siemens MAGNETOM Vision system. The number of acquisitions was 96 and each one consisted of 64 contiguous slices ($64\times64\times64\times 3\times3\times3$ mm^3 voxels). Acquisition took 6.05s, with the scan to scan repeat time (TR) set arbitrarily to 7 s. The acquisitions were made in blocks of 6, giving 16 42 s blocks. The condition associated with each block alternated between rest and auditory stimulation, starting with rest. Auditory stimulation was bi-syllabic words presented binaurally, at a rate of 60 per minute. As the SPM site recommends the first few scans are discarded to avoid T1 effects in the initial scans of an fMRI time series. Then, 84 acquisitions were finally used after discarding the first complete cycle (12 scans). The images were preprocessed (realigned, coregistered using a sMRI, normalized and smoothed) for collecting two different conditions, rest and listening. Then, a GLM specification followed by model estimation and a t-test-based inference (FWE p-value = 0.05) resulted in the activation maps for this auditory-evoked potential experiment.

B.3.1. Detecting auditory stimulation maps

In the last sections we have seen the potentiality of the proposed approach for ROI detection in several binary classification paradigms, i.e. diagnosis, given the usefulness of machine learning. Images collected from the aforementioned experiment are used to identify areas performing a specific information processing function, such as the primary auditory cortex.

The areas identified by the proposed approach are mainly those corresponding with the temporal lobe, as shown in Fig. 13. A mosaic and the 3D representation of the activated cortical areas are shown in the same Fig. 13(b), together with the activation pattern sought by the SPM methodology. The comparison analysis of both approaches is displayed in Fig. 13(a). In the upper figure we see the significance test for a proportion (n = 84) that was applied to this auditory fMRI experiment. The SAM is mainly located on regions where we found the activation voxels in SPM. In the middle we represent the number of voxels in ROIs (for different sample sizes) and the ratio w.r.t the total number of voxels in that region. Finally, in the bottom we compared both approaches using the overlap-analysis type measures, as described in the last sections. To sum up, we found: (i) the same ROIs in both approaches, (ii) SPM required sufficiently large sample size to provide significant ROIs, i.e. for n = 20 no significant areas were sought, and (iii) both approaches converge with increasing sample size to the same number of activated voxels.

References

[1] K. Friston, Sample size and the fallacies of classical inference, Neuroimage 81 (2013) 503–504.

[2] D. Bzdok, Classical statistics and statistical learning in imaging neuroscience, Front. Neurosci. (2017) http://dx.doi.org/10.3389/fnins.2017.00543.

[3] R. Heller, et al., Conjunction group analysis: An alternative to mixed/random effect analysis, Neuroimage 37 (2007) 1178–1185.

[4] J.D. Rosenblatt, et al., Revisiting multi-subject random effects in fMRI: Advocating prevalence estimation, Neuroimage 84 (2014) 113–121.

[5] K.J. Friston, et al., Classical and Bayesian inference in neuroimaging: theory, Neuroimage 16 (2) (2002) 465–483.

[6] M.A. Lindquist, et al., Ironing out the statistical wrinkles in ten ironic rules, Neuroimage 81 (2013) 499–502.

[7] P.T. Reiss, et al., Cross-validation and hypothesis testing in neuroimaging: an irenic comment on the exchange between Friston and Lindquist, Neuroimage 116 (2015) 248–254.

[8] A.M. Winkler, et al., Non-parametric combination and related permutation tests for neuroimaging, Hum. Brain Mapp. 37 (4) (2016) 1486–1511.

[9] R. Kohavi, A study of CV and bootstrap for accuracy estimation and model selection, in: Proc. of the 14th International Joint Conference on AI, vol. 2, 1995, pp. 1137–1143.

[10] J.M. Górriz, et al., A machine learning approach to reveal the neurophenotypes of autisms, Int. J. Neural Syst. (2019) 1850058.

[11] J.M. Górriz, et al., On the computation of distribution-free performance bounds: Application to small sample sizes in neuroimaging, Pattern Recognit. 93 (2019) 1–13.

[12] G. Varoquaux, Cross-validation failure: Small sample sizes lead to large error bars, Neuroimage 180 (2018) 68–77.

[13] J.D. Rosenblatt, et al., Better-than-chance classification for signal detection, Biostatistics (2016).

[14] I. Kim, et al., Classification accuracy as a proxy for two sample testing, Ann. Stat. (2020).

[15] V. Vapnik, Estimation Dependencies Based on Empirical Data, Springer-Verlach, ISBN: 0-387-90733-5, 1982.

[9] www.fil.ion.ucl.ac.uk/spm/data/auditory/.

[10] www.fil.ion.ucl.ac.uk/spm/.

[16] D. Haussler, Decision theoretic generalizations of the PAC model for neural net and other learning applications, Inform. and Comput. 100 (1) (1992) 78–150.
[17] V. Vapnik, et al., On the uniform convergence of relative frequencies of events to their probabilities, Theory Probab. Appl. 16 (1971) 264–280.
[18] P. Massart, Some applications of concentration inequalities to statistics, Ann. Fac. Sci. Toulouse (2000).
[19] C. McDiarmid, On the method of bounded differences, Surv. Combin. 141 (1989) 148–188.
[20] N. Sauer, On the density of families of sets, J. Combin. Theory Ser. A 13 (1972) 145–147.
[21] S. Shelah, A combinatorial problem: stability and order for models and theories in infinity languages, Pacific J. Math. 41 (1972) 247–261.
[22] N. Tzourio-Mazoyer, et al., Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the MNI MRI single subject brain, Neuroimage 15 (2002) 273–289.
[23] C. Allefeld, et al., Valid population inference for information-based imaging: From the second-level t-test to prevalence inference, Neuroimage 141 (2016) 378–392.
[24] T.M. Cover, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, IEEE Trans. Electron. Comput. EC-14 (1965) 326–334.
[25] J.M. Rondina, Scors - a method based on stability for feature selection and mapping in neuroimaging, IEEE Trans. Med. Imaging 33 (1) (2014) 85–98.
[26] F. De Martino, et al., Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns, Neuroimage 43 (1) (2008) 44–58.
[27] A.R. McIntosh, et al., Spatial pattern analysis of functional brain images using partial least squares, Neuroimage 3 (3 Pt 1) (1996) 143–157.
[28] F.J. Martinez-Murcia, et al., Studying the manifold structure of Alzheimer's disease: A deep learning approach using convolutional autoencoders, IEEE J. Biomed. Health Inform. (2019).
[29] R. Rosipal, et al., Overview and Recent Advances in Partial Least Squares, Springer, Berlin, Heidelberg, 2006, pp. 34–51.
[30] J. Mouro-Miranda, et al., Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data, Neuroimage 28 (2005) 980–995.
[31] V. Gómez-Verdejo, et al., Sign-consistency based variable importance for machine learning in brain imaging, Neuroinformatics 17 (4) (2019) 593–609.
[32] B.S. Khundrakpam, et al., Prediction of brain maturity based on cortical thickness at different spatial resolutions, Neuroimage 111 (2015) 350–359.
[33] A. Rakotomamonjy, et al., Simplemkl, J. Mach. Learn. Res. 9 (Nov) (2008) 2491–2521.
[34] A. Antós, et al., Data-dependent margin-based generalization bounds for classification, J. Mach. Learn. Res. 3 (2002) 73–98.
[35] M. Vidyasagar, Learning and Generalisation with Applications To Neural Networks, Springer, ISBN: 978-1-84996-867-6, 2003.
[36] R.S.J. Frackowiak, et al., Human brain function, in: Introduction To Random Field Theory, second ed., Academic Press, ISBN: 978-0-12-264841-0, 2004, pp. 867–879 (Chapter 44).
[37] K. Friston, et al., Statistical parametric maps in functional imaging: A general linear approach, Hum. Brain Mapp. 2 (1995) 189–210.
[38] J. Ashburner, et al., Unified segmentation, Neuroimage 26 (3) (2005) 839–851.
[39] K. Friston, Ten ironic rules for non-statistical reviewers, Neuroimage 61 (2012) 1300–1310.
[40] C.C. Jack Jr., et al., NIA-AA Research framework: Toward a biological definition of Alzheimer's disease, Alzheimers Dement 14 (4) (2018) 535–562.
[41] G.M. McKhann, et al., The diagnosis of dementia due to Alzheimer's disease: recommendations from the national institute on aging and the Alzheimer's assocation workgroup, Alzheimers Dement. 7 (2011) 263–269.
[42] A. Eklund, et al., Cluster failure: Inflated false positives for fMRI, Proc. Natl. Acad. Sci. 113 (28) (2016) 7900–7905, http://dx.doi.org/10.1073/pnas.1602413113.
[43] D.L. Lorca-Puls, et al., The impact of sample size on the reproducibility of voxel-based lesion-deficit mappings, Neuropsychologia 115 (2018) (2018) 101–111.
[44] J.M. Gorriz, et al., Artificial intelligence within the interplay between natural and artificial computation: Advances in data science, trends and applications, Neurocomputing 410 (14) (2020) 237–270.
[45] A. Abraham, et al., Machine learning for neuroimaging with scikit-learn, Front. Neuroinform. 8 (2014) 14.
[46] J. Gareth, et al., The error coding method and picts, J. Comput. Graph. Statist. 7 (3) (1998) 377–387.
[47] S. Shalev-Shwartz, et al., Understanding Machine Learning – from Theory To Algorithms, Cambridge University Press, ISBN: 9781107057135, 2014.
[48] I.A. Illan, et al., Automatic assistance to Parkinson's disease diagnosis in DaTSCAN SPECT imaging, Med. Phys. (2012).
[49] K.S. Button, et al., Confidence and precision increase with high statistical power, Nat. Rev. Neurosci. 14 (2013) 585.
[50] H. Zaidi, et al., Quantitative Analysis in Nuclear Medicine Imaging, Springer Science Business Media, Inc., ISBN: 10: 0-387-23854-9.
[51] J.P.A. Ioannidis, Why most published research findings are false, PLoS Med. 2 (8 (e124)) (2005) 696–701.