A connection between pattern classification by machine learning and statistical inference with the General Linear Model

J.M. Górriz, C. Jiménez-Mesa, F. Segovia, J. Ramírez, SiPBA group and J. Suckling

***Abstract*—A connection between the general linear model (GLM) with frequentist statistical testing and machine learning (MLE) inference is derived and illustrated. Initially, the estimation of GLM parameters is expressed as a Linear Regression Model (LRM) of an indicator matrix; that is, in terms of the inverse problem of regressing the observations. Both approaches, i.e. GLM and LRM, apply to different domains, the observation and the label domains, and are linked by a normalization value in the least-squares solution. Subsequently, we derive a more refined predictive statistical test: the linear Support Vector Machine (SVM), that maximizes the class margin of separation within a permutation analysis. This MLE-based inference employs a residual score and associated upper bound to compute a better estimation of the actual (real) error. Experimental results demonstrate how parameter estimations derived from each model result in different classification performance in the equivalent inverse problem. Moreover, using real data, the MLE-based inference including model-free estimators demonstrates an efficient trade-off between type I errors and statistical power.**

***Index Terms*—General Linear Model, Linear Regression Model, Pattern Classification, upper bounds, permutation tests, cross-validation**

I. Introduction

Despite the popularity of machine learning (MLE) as a solution for a wide range of complex problems [28], [14], there remains an open question about its usefulness for between-group statistical inference. Neuroimaging in particular has embraced MLE as a technology to deliver diagnostic and prognostic classification [3], [20] of neurological and psychiatric disorders. Nevertheless, the mainstay of neuroimaging studies are observational and mechanistic, seeking to identify regional between-group differences in brain structure and function. Efforts with MLE in this space are increasing with continuous output variables ([4], with remarks in [31]) rather than the more typical categorical classifications.

Manuscript received Jan, 2021 , J.M. Górriz, C. Jiménez-Mesa, F. Segovia, J. Ramírez, SiPBA group are with the Data Science and Computational Intelligence Institute, University of Granada, Granada, 18071 Spain, e-mail: gorriz@ugr.es, jg825@cam.ac.uk. J.M. Górriz and J. Suckling are with the Dpt. Psychiatry, Herchel Smith Buidling for Brain & Mind Sciences, Forvie Site Robinson Way, University of Cambridge, Cambridge CB2 0SZ, UK. The SiPBA group are R. Romero-García, A. Ortiz, F.J. Martínez, C.G. Puntonet, I.A. Illán, J.E. Eloy, D. Castillo, D. Salas, J.M. Mateos

Part of the data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this manuscript. ADNI investigators include (complete listing available at http://www.loni.ucla.edu/ADNI/Collaboration/ADNI Manuscript Citations.pdf).

Several advances for combining p-value maps have been proposed based on the concept of *prevalence* [18], [32] that go beyond the fixed and mixed (random) effects models [10]. Common to these approaches is the assumption of a mixing of subject classifications at each voxel that is more realistic than those assumed in classic random effect approaches; for example, homogeneity of the binary activation pattern [32], and offers the possibility of a new framework for modern statistics.

The concept of prevalence as a fraction of individuals correctly classified by MLE algorithms in group comparisons is not novel in neuroimaging, and is indeed the main focus of predictive inference. As an example, out-of-sample generalization approaches, such as Cross-Validation (CV), try to estimate on unseen data the accuracy (A_{cc}) of a classifier in a binary classification problem. Although the methods and goals of predictive CV inference are distinct from classical extrapolation procedures [24], they are exploited within frameworks aimed at assessing statistical significance [31]. Bootstrapping, binomial or permutation ("resampling") tests [39] are all examples that have been demonstrated as competitive outside classical statistics, filling otherwise-unmet inferential needs.

In a pattern classification problem we usually assume the existence of classes (H_1) that can be differentiated by classifiers with their performance measured in terms of A_{cc} or *prevalence* on an independent dataset. Then, we accept (improperly in a statistical sense) the alternative hypothesis H_1 using empirical confidence intervals such as standard deviations of the classification A_{cc} from dataset folds. In cases of limited sample sizes, the most popular k-fold CV method [23] is sub-optimal under unstable conditions [12], [13], [37]. In such circumstances, the predictive power of the trained classifiers can be arguable. Moreover, it has been partially demonstrated that when using only a classifier's empirical A_{cc} as a test statistic, the probability of detecting differences between two distributions is lower than that of a *bona fide* statistical test [33], [22].

Beyond empirical techniques for the estimation of performance, MLE is well-established in data-driven statistical learning theory (SLT), which is primarily devoted to problems of estimating dependencies with limited amounts of data [36]. Although CV-MLE approaches were not originally designed to test hypotheses based on prevalence in brain mapping [11], they are theoretically grounded to provide confidence intervals (protected inference) in the classification of image patterns formulated as maps of statistical significance [15]. 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 [15]. This results in improvements to the test's statistical power based on A_{cc}. Thus, assessing with high probability the quality of the fitting function (and its generalization ability) in terms of in- and out-of-sample predictions can be conceptualized, under a hypothesis testing scenario, as the inverse problem of "carefully rejecting H_0"; that is, the problem of rejecting H_1, and thus accepting H_0 (that there is no effect, or it is not significant).

In this paper we show a connection between the classical general linear model (GLM), including random effect models, with the MLE framework for the estimation of model/classifier parameters and subsequent analyses to achieve the level of significance in group comparisons. In this sense, inference based on the parametric T statistic and prevalence-based probability tests are two different paths solving the same problem. We also show a novel method for achieving statistical significance using MLE and permutation tests based on concentration inequalities. This approach assesses the worst case of the actual error and proposes an estimation of the observed distribution of permuted data.

II. Methods: Classical and MLE statistical inferences

A. Background on classical statistics in neuroimaging

The GLM [10] is defined for a single observation level, e.g. in a between-subject comparison, as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{y}$ is the $N \times 1$ observation vector with units of time, voxels, etc., $\boldsymbol{\epsilon}$ is the $N \times 1$ vector of errors that is assumed to be Gaussian distributed, $\mathbf{X}$ is the $N \times M$ matrix containing the explanatory variables or constraints, and $\boldsymbol{\theta}$ is the $M \times 1$ vector of parameters explaining the observations $\mathbf{y}$. Note that: i) for a hierarchical observation model each level requires the prior estimation of the previous levels; and ii) in terms of MLE, $\mathbf{X}$ plays the role of multidimensional labels or regressors acting on the observations $\mathbf{y}$. In the classic GLM, $\boldsymbol{\theta}$ is usually estimated by a Maximum Likelihood (ML) criterion based on the Gaussianity assumption and is given by:

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^t\mathbf{C}_\epsilon^{-1}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{C}_\epsilon^{-1}\mathbf{y} \tag{2}$$

where $\mathbf{C}_\epsilon$ is the covariance matrix of errors. Inferences on this estimate[1]: how large are the components of $\boldsymbol{\theta}$ and the relationship between classical GLM and MLE-based prevalence inferences can be obtained using a linear compound specified by a contrast weight vector $\mathbf{c}$, and writing a T statistic as:

$$T = \frac{\mathbf{c}^t\hat{\boldsymbol{\theta}}}{\sqrt{\mathbf{c}^t Cov(\hat{\boldsymbol{\theta}})\mathbf{c}}} \tag{3}$$

[1]Here, we refer to voxelwise inference since we use a threshold u to classify voxels i as "active" if $T_i \geq u$. Clusterwise inference uses a cluster-forming threshold to define contiguous suprathreshold regions [29].

where $Cov(\hat{\boldsymbol{\theta}}) = (\mathbf{X}^t\mathbf{C}_\epsilon^{-1}\mathbf{X})^{-1}$. This T statistic gives us the probability of observing the ML estimation under H_0 and when it is small enough, e.g. $p < 0.05$, the linear compound is considered significantly different from zero. As an example, given a set of two parameters in $\boldsymbol{\theta} = [\theta_1, \theta_2]^T$, if we select $\mathbf{c} = [1 - 1]$ we are assessing how large is the first parameter with respect to the second; i.e. the difference $\theta_1 - \theta_2$. Thus, if the T statistic suggests a small probability, the difference is statistically significant and observations are generated from different sources.

A similar procedure could be established based on a Bayesian estimation and inference to handle complex hierarchical observational models. This framework is based on Expectation Maximization (EM) for parameter estimation along with known priors and *a priori* probability models, with the aim of evaluating the posterior probability (ppm). By thresholding the ppm, relationships between this and the frequentist approach can be established including similarities (statistical power) and differences (specificity) [10].

1) Least Squares of the GLM: The GLM can be estimated without any assumptions about the noise model by simply solving the associated Least Squares (LS) problem. Therefore, if we assume that $\boldsymbol{\epsilon} = 0$ in the GLM, the problem is now to find the "best" set of parameters θ_i that explains each observation y_i by:

$$y_j = \sum_{i=1}^{M} X_{ji}\theta_i; \quad \text{for } j = 1, \ldots, N \tag{4}$$

Thus, we need to solve the linear regression problem given in equation 4 to estimate the parameters θ_i. The most popular estimation method is LS, in which we select the coefficients $\boldsymbol{\theta}$ to minimize the residual sum of squares:

$$RS(\boldsymbol{\theta}) = \sum_{j=1}^{N}(y_j - \sum_{i=1}^{M} X_{ji}\theta_i)^2 \tag{5}$$

The solution to this problem ($\frac{\partial RS(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$), the Markov-Gauss estimate, provides the smallest variance among all linear unbiased estimates and is given by:

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} \tag{6}$$

similar to the GLM solution (equation 2) but assuming $\mathbf{C}_\epsilon = \mathbf{I}$ in the latter model; that is, if the errors are assumed to be independently and identically distributed the ML estimation is equivalent to the LS solution in equation 6.

B. Converting the estimation of $\boldsymbol{\theta}$ into a LS classification problem

In the LS multiclass classification problem, the goal is to design M linear functions $f_i(\mathbf{y}) = \mathbf{w}_i^t\mathbf{y}$, given a set of input patterns $\mathbf{y}_i$ and according to a suitable mean squared error (MSE) criterion with respect to some desired discrete output binary code $\mathbf{x}_i$, i.e. labels. Note that, in general, this setup is found in neuroimage analyses where the design matrix contains discrete values, e.g. experimental conditions in fixed-effects analysis or in random effect modeling between groups. Recently, the residual score or classification error obtained

from several methodologies beyond LS (e.g. by applying the fitted linear hyperplanes to new unseen data) have been deployed to establish a CV A_{cc} test from data with permuted labels [31], [15].

1) The inverse problem: LS for regressing an indicator matrix: Consider the general inverse problem; that is, given a set of observations $\{\mathbf{y}_i\}$, for $i = 1, \ldots, N$, we are interested in explaining a set of "explanatory" binary-coded variables $\mathbf{x}_i$ (labels) by a matrix $\mathbf{W}$ of parameters. This problem, referred to here as the inverse problem in the *label domain*, is also known as the linear regression of an Indicator Matrix or the linear regression model (LRM) [16]. In this model, we regress the explanatory variables instead of regressing the observed responses as in the GLM. This regression can be more accurate depending on the nature of the data to be fitted e.g. for a low number of discrete classes in the specified design matrix $\mathbf{X} = [x_{im}]$.

If we have M classes then $\mathbf{X}$ is a $N \times M$ matrix, where each row $i = 1, \ldots, N$ contains a single $x_{im} = 1$, for $m = 1, \ldots, M$, $\mathbf{Y}$ is the $N \times P$ matrix of column responses $\mathbf{y}_i$ and $\mathbf{W}$ is a $P \times M$ coefficient matrix. Thus, we fit a linear regression model of the form:

$$\mathbf{X} = \mathbf{Y}\mathbf{W} \tag{7}$$

where the P dimension allows the inclusion of several responses (multimodality or multiframe acquisitions) given the same indicator response matrix $\mathbf{X}$. Following the methodology as that leading to equation 6, the best estimation is given by:

$$\hat{\mathbf{W}} = (\mathbf{Y}^t\mathbf{Y})^{-1}\mathbf{Y}^t\mathbf{X} \tag{8}$$

which regresses inputs of observations on to a novel set of labels or constraints:

$$\hat{\mathbf{X}} = \mathbf{Y}\hat{\mathbf{W}} \tag{9}$$

The novel set $\hat{\mathbf{X}}$ can be seen as a guess of the constraints for the set of observation vectors $\mathbf{y}_i$, or an approximation of the posterior probability $p(\text{class} = m|\mathbf{y})$. Thus, it allows us to compute an error model as:

$$\boldsymbol{\epsilon}_{LS} = \mathbf{X} - \hat{\mathbf{X}} \tag{10}$$

2) Connection between θ *and* $\mathbf{w}$*:* For simplicity, and to connect with the GLM as shown in section II-A, let $P = 1$ in the LRM, then $\mathbf{W} = \mathbf{w}$ is a $1 \times M$ row vector and $\mathbf{Y} = \mathbf{y}$ is an $N \times 1$ column vector. A simple relation between the GLM and LRM approximations can be found taking into account that:

$$\mathbf{X} = \mathbf{y}\hat{\mathbf{w}} + \boldsymbol{\epsilon}_{LS} \tag{11}$$

at the LS solution. Thus, multiplying both sides on the right by $\mathbf{w}^t$ we can solve the equation for $\mathbf{y}$ and obtain the corresponding GLM as:

$$\mathbf{y} = (\mathbf{X} - \boldsymbol{\epsilon}_{LS})\tilde{\boldsymbol{\theta}} \tag{12}$$

where we define $\tilde{\boldsymbol{\theta}} = \hat{\mathbf{w}}^t(\hat{\mathbf{w}}\hat{\mathbf{w}}^t)^{-1}$ and the GLM noise model is derived using $\boldsymbol{\epsilon} = -\boldsymbol{\epsilon}_{LS}\tilde{\boldsymbol{\theta}}$. The scalar term of equation 12 can be expressed with the LS solution as:

$$(\hat{\mathbf{w}}\hat{\mathbf{w}}^t)^{-1} = (\mathbf{y}^t\mathbf{y})^2/((\mathbf{X}^t\mathbf{y})^t\mathbf{X}^t\mathbf{y}) = \frac{(\sum_{i=1}^{N} y_i^2)^2}{\sum_{m=1}^{M}\sum_{i,j} y_{im}y_{jm}} \tag{13}$$

where y_{im} denotes observation i belonging to class m. Thus, a LS linear regression of the observations can be described by a GLM regression on the observations (i.e. a linear regression on the explanatory variables), and vice versa.[2]

3) Inference of the inverse GLM based on MLE: The LRM can be seen as a generalization of the GLM for the responses, coding $\mathbf{x}$ as a vector of continuous noisy responses, that is, by constructing vector targets for each class [16]. From equation 11, which is equivalent to the inverse GLM in equation 1, inference on this model based on MLE could proceed as follows. Based on a set of data pairs $(y_i, \mathbf{x_i})$, we estimate the set of parameters $\mathbf{w}$ using a similar expression to equation 8 or other more refined predictive algorithms [6], [25], e.g. SVM. After the fitting process, we assess its significance under the null hypothesis, likewise the T-statistic inference on the GLM, on an independent set Ω_{CV} using a CV A_{cc} test statistic:

$$T_{CV} = \sum_{i \in \Omega_{CV}} ||(\mathbf{x_i} - (y_i\hat{\mathbf{w}})^t||^2 \tag{14}$$

The null distribution is modeled by randomly rearranging labels a large number of times π to create artificial data sets, $(y_i, \mathbf{x}_{\pi_\mathbf{p}})$, for $p = 1, \ldots, O$, i.e. a permutation test, and evaluating the sum of squared residuals T_{CV} with every unseen sample within the permuted and original set. Consequently, the p-value is defined by[3]:

$$p_{value} = \frac{card\{T_{CV}^{\pi} < T_{CV}\} + 1}{O + 1} \tag{15}$$

where $card(.)$ is the cardinality of a set and T_{CV} and T_{CV}^{π} are the CV A_{cc} tests on the original and permuted sets, respectively. These p-values under the null hypothesis are pivotal quantities and, in principle, could be used for multiple testing correction, instead of the statistic image based on accuracy. The distribution of minimum p-values, p_j^{min}, used in the test to deal with the multiple comparison problem is limited by the number of permutations, $\frac{1}{O} \leq p_j^{min}$, and may cause considerable loss of power [38]. Other methods, such as RFT [8] or a correction based on the false-discovery rate (FDR) can be used once the uncorrected p-values have been obtained for each voxel in the image. In the experiments, due to the discreteness of the p-values that is strongly limited for computational reasons ($O = 1000$), we employ a combination of correction methods for multiple testing, e.g. a single-threshold test applied to the map of uncorrected p-values for comparison purposes in the control of FWE rates, or a Bonferroni corrected p-value calculated at each voxel for assessing power.

In the latter test, also known as P-test [31], we assume that we have a good procedure for estimating $\mathbf{w}$. However, CV is a standard procedure for estimating the actual error of any classifier, which is found to be unstable in limited samples sizes [37], [13]. We could improve this estimation by including a term to cope with the possibility that the fitting process is

[2] given a GLM on the observations, we can define a LRM on the explanatory variables as $\tilde{\mathbf{w}} = \theta^T(\theta\theta^T)^{-1}$, at the ML solution, with an error $\boldsymbol{\epsilon}_{LS} = -\boldsymbol{\epsilon}\tilde{\mathbf{w}}$

[3] the correction factor +1 in the numerator and denominator is justified by the inclusion of the original sample set in the test

not as good as expected, and thus the resulting estimate is not a good predictor. In this sense, other alternatives [1], [26] could be tested by the assessment of the worst case based on concentration inequalities and the resubstitution estimate as:

$$T_{Res} = \sum_{i \in \Omega} ||(\mathbf{x_i} - (y_i \hat{\mathbf{w}})^t||^2 + \Delta^2(N, P) \qquad (16)$$

where $\Delta(N, P)$ is model-free upper bound of the actual risk [13], [36], [15], [5] with a probability at least $1 - \alpha$ and Ω is the dataset at hand.

III. Experimental results

In the first of the experiments, and to clearly state the problem and the solutions, we consider a simple group comparison with only a single (second) level analysis (a Bayesian approach of this problem is equivalent to the GLM based inference on this single level), using a binary design matrix, e.g. it models the population-specific effects. This is the well-known case-control design often used as the basis for a diagnostic test; e.g. Alzheimer patients vs unaffected controls. We adjusted the GLM and the equivalent problems using LRM and SVM by regressing observed variables using a simple explanatory matrix $\mathbf{X}$ and a Gaussian model for the noise to obtain two parameters θ_1, θ_2, as follows:

$$\mathbf{Y}|_{N\times 1} = \mathbf{X}|_{N\times 2}\boldsymbol{\theta}|_{2\times 1} + \boldsymbol{\epsilon}|_{N\times 1}$$

where, as an example,

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ \dots & \dots \end{pmatrix}$$

is a matrix of explanatory variables containing 1s and 0s indicating the class membership of the observation using a two-element binary code. A more general hierarchical model with a non-binary design matrix (including regressors, covariates, etc.) could be processed the same way by fitting the set of parameters step by step by *pattern regression*, however we are interested in assessing the connection between $\boldsymbol{\theta}$ and $\mathbf{w}$ in this paper for a binary (design matrix) pattern classification problem. The objective is two-fold: i) the estimation of model parameters using both methodologies and domains, linking them by the theoretical connection in equation 12; and, ii) to assess how completely they explain observations and labels in both domains. The second objective can be tackled by showing the estimations and the group of observations in both domains, and by quantitatively evaluating the classification error in the equivalent label domain, given the expected ideal values for model parameters.

In the last part of this section, we show the inference analysis derived from the two methodologies in each domain. We regressed on the observations and on the labels to construct and assess the spatially extended statistical processes, generating maps of significance, using the MRI ADNI dataset [15]. In doing so we compared Statistical Parametric Maps (SPMs, a two-sample T-statistic similar to equation 3), where significance is first individually assessed at each voxel, and then combined using three configurations: first, with a cluster-defining threshold of $P = 0.001$ (uncorrected for multiple comparisons) alone; second, then adding a cluster extent threshold (CET) = 10 voxels; and third, a Family Wise Error correction at $P = 0.05$ on the clusters based on random field theory (RFT) [8]. In addition, the P-tests described in section II-B3 were also conducted.

A. Simulated Data with Noisy Observations (DG1)

A N-dimensional Gaussian noise vector $\mathbf{v}$ was randomly drawn with zero mean and an $N \times N$ covariance matrix with 2-norm equal to 1. A vector of observations was then constructed by adding the noise to a binary vector (a column in the explanatory matrix of indicators); i.e. $\mathbf{y} = \mathbf{X}_k + \mathbf{v}$ for $k \in \{1, 2\}$. The design matrix was then obtained by $X = [\mathbf{X}_k \bar{\mathbf{X}}_k]$, where $\bar{(.)}$ denotes logical negation.

Once the observations were artificially drawn (see figure 1), with increasing sample size we regressed both explanatory variables (LRM by LS and SVM) and observations (GLM) to obtain a set of two parameters for each model $\boldsymbol{\theta} = [\theta_1, \theta_2]$, $\mathbf{w} = [w_1, w_2]$. All these methods can be employed to estimate the regressed observed variables using equations 1 and 12, given the explanatory matrix and the estimated parameters, as shown in figure 2. In this figure we also plot the distribution of the T-statistic over 1000 simulations (top), a sample of this distribution that shows the variability of the estimation using the GLM around the ideal value of $\theta_2 - \theta_1 = 1$ (middle) and the estimated observation by each of the models (bottom).

Connected with the previous one-sample GLM estimations, we plot in figure 3 the estimated parameters explaining the observations using all the methods along with the observations they model. Note the large variability of the GLM estimation with increasing sample size. In figure 4 we show the inverse problem; how the methods estimate the $\mathbf{w}$ from the point of view of the label regression. In this case, it is readily seen that the one sample GLM model provides a sub-optimal estimation at different sample sizes; i.e. the red curve lies above the blue curve. As expected, the use of these parameters in the dual classification problem results in a larger empirical error as shown in figure 5.

From these results we can conclude that the link between the two approaches resides in the differing nature of the regression procedure. In both domains there is an implicit classification task once the parameters, that better explain the corresponding observations, are derived. These parameters are fitted taking into account only the empirical data available (including a noise model, if present). Therefore, w_m for a given model m, can be used to regress the observations to obtain a novel data set in the label space (new regressed labels), which can be associated with the states (or classifications) of the explanatory matrix. This classification task provides an empirical error (figures 5 and **??**). Other methods could be used to obtain such parameters in a (non-)linear fashion. As an example, we compared the decision boundary obtained by LRM with that with SVM (figures 5 and **??**), which illustrates the differences between methods in terms of generalization ability.

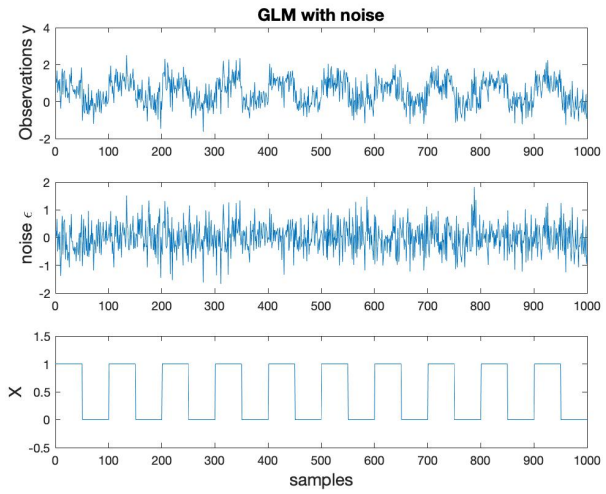

Fig. 1: Simulated data with noisy observations (DG1) example

B. Empirical data: a case-control design of the ADNI Dataset

The data used 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 T1-weighted structural MRI scans from patients with Alzheimer's disease (AD), Mild Cognitive Impairment (MCI), and cognitively normal controls (NC) acquired at multiple time points. Here we only included 1.5T structural MRI. The original database contains more than 1000 T1-weighted MRI images, although for the proposed study only the first MRI examination of each subject was included, resulting in 417 structural MRIs. Following the recommendation of the National Institute on Aging and the Alzheimer's Association (NIA-AA) for the use of imaging biomarkers [21], we considered the group comparison NC vs. AD for establishing a clear framework for comparing statistical paradigms (SPM and T_{CV}), since the MCI class is strictly based on clinical criteria, without including any other biomarker information [27]. Demographic data is summarized in Table I. The dataset was preprocessed using standardised neuroimaging methods and protocols implemented in the SPM software (www.fil.ion.ucl.ac.uk/spm/), including registration in MNI space by spatial normalization and segmentation to differentiate grey and white matter and other brain tissues [9]. Here, we used the grey matter (GM) estimates.

TABLE I: Demographic details of the MRI ADNI dataset, with group means and their standard deviations

Status	Number	Age	Gender (M/F)	MMSE
NC	229	75.97±5.0	119/110	29.00±1.0
AD	188	75.36±7.5	99/89	23.28±2.0

1) Assessing the statistical power: We fitted the set of parameters using linear SVM and evaluated the T_{CV} statistic on the original dataset; see figure 6. As shown in this figure, the resubstitution estimate is more optimistic in the A_{cc} distribution than the K-fold based estimate. Note that this analysis is independent of the selected fold as we performed $\sim 10^6$ folds, one per voxel. However, both are optimistic since the mean of the distribution is not clearly located around 0.5 (it is already shifted to the right, beyond the effect due to truly significant regions). The effect is even larger when the groups are slightly imbalanced, simulating the case of over-powered datasets, as shown in the bottom of the figure 6. However, note how the corrected bound [13] clearly shifts the A_{cc} obtained by resubstitution to the left, resulting in a better (more conservative) estimation of the statistic across the whole volume.

Based on the T_{CV} and T_{Res} values from the original dataset, and those obtained using a permutation analysis ($O = 1000$) for a selection of structures (e.g. left hippocampus, a brain structure with a well-established role in the progression of AD), we compared the SPMs processed with the inference approaches described in section II-B3. Note that the large number of voxels that composes an image limits the permutation analysis to specific structures. Results from the left hippocampus are depicted in figure 7. The permutation analysis reveals how the power of the T_{CV} approach is affected in this featured region, where a true effect might be found in almost the entire structure. The statistical power of the T_{Res} is preserved through the permutation procedure (2058 detected voxels vs 1024 voxels out of 2237, figure 7). It is also worth mentioning the CDF of the errors derived in this particular region and the corresponding distribution of p-values, recalling that the dataset included patients with advanced AD, and thus the selected structure should be significantly affected by the disease.

To extend the analysis to the whole volume, we approximately simulated the null distribution outside the left hippocampal region in two steps. First, we computed the set of p-values in the left hippocampus (around $2 \cdot 10^3$ voxels) following equation 15 and determined the averaged T threshold, T_{th}, that approximately provided an appropriate significance level, e.g. 0.05. Then, assuming that for any $T < T_{th}$ the probability of an observation is p-value< 0.05, we thresholded the remainder of the image to obtain the significant voxels showing an effect. This approach clearly requires multiple-comparison correction as several dependent or independent statistical tests are being performed simultaneously at a given significance level. Therefore, we made the significance level more conservative with $\alpha = 0.001$ to reduce the presence of false positives (FP) in the permutation analyses shown in section II-B3 and then compared it with the inference made using the three SPM configurations across the whole volume. In figure 8 we show the detection ability together with the control of type I errors in the T_{Res} approach (map in red font). Note how the permutation test affects the detection ability of the classical k-fold CV approach (map in green font), and how the uncorrected voxelwise approaches (in blue font, bottom left and middle) inflates the number of FPs.

2) Controlling type I error: It is important to evaluate the ability of the inference methods for controlling the FP rates [15], [7]. As an example, in [15] the ability of upper-bound based inference to control type I errors with increasing sample size was demonstrated using a global test for a proportion within the whole volume. Moreover, in [7] clusterwise in-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3101662, IEEE Journal of Biomedical and Health Informatics

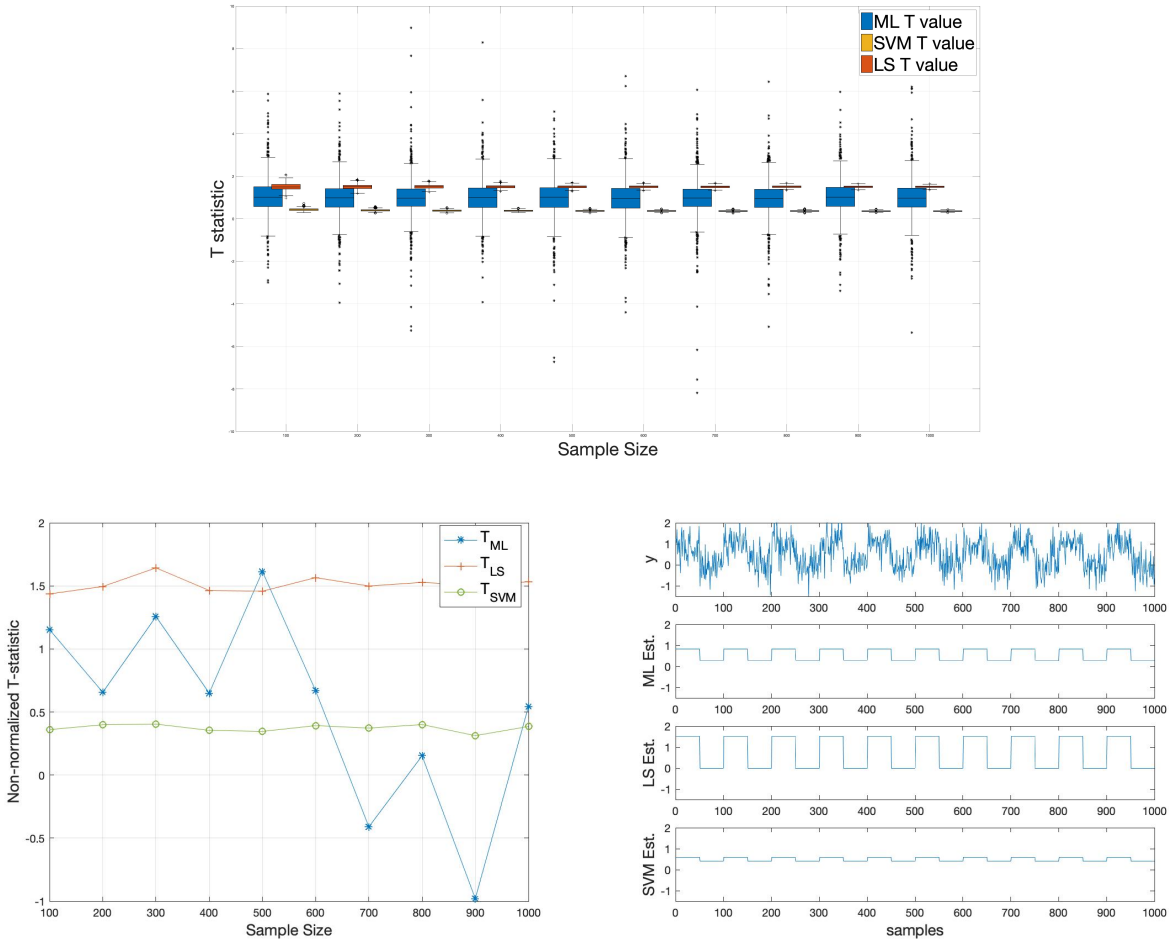

Fig. 2: Estimated Observations and T-statistic distribution. Note that in the GLM model we use the covariance matrix of the noise to evaluate equation 2; that is, in the estimation of θ. We show the comparison between non-normalized statistics of all the estimations, i.e. suppressing the covariance term in the GLM, in a random (R=1000) simulation. This clearly demonstrates that only on average does the ML statistic converge to the ideal value $\theta_2 - \theta_1 = 1$ unlike the single sample of this distribution shown in the bottom left.

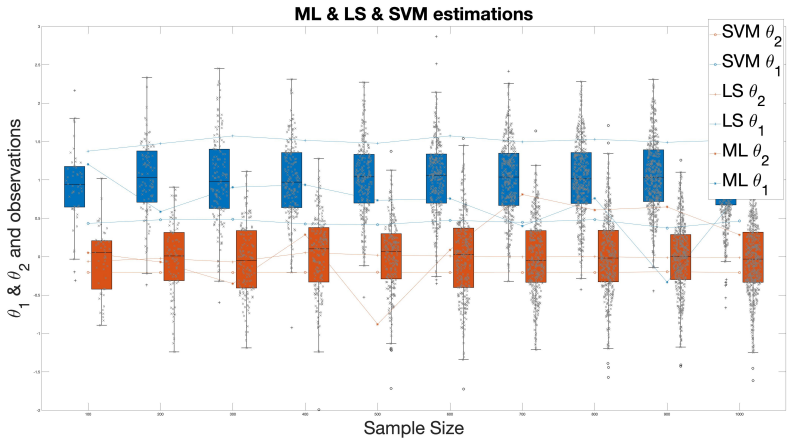

Fig. 3: Distribution of observations (y) and estimations of θ for GLM, LRM and SVM in DG1

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3101662, IEEE Journal of Biomedical and Health Informatics

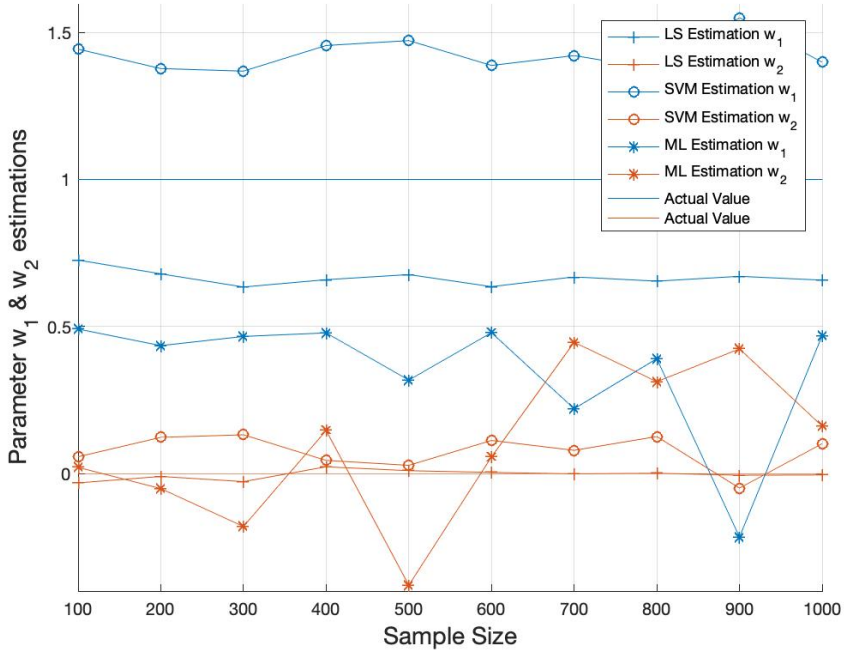

Fig. 4: Estimations of the parameter $\mathbf{w}$ regressing the observations with increasing sample size in DG1

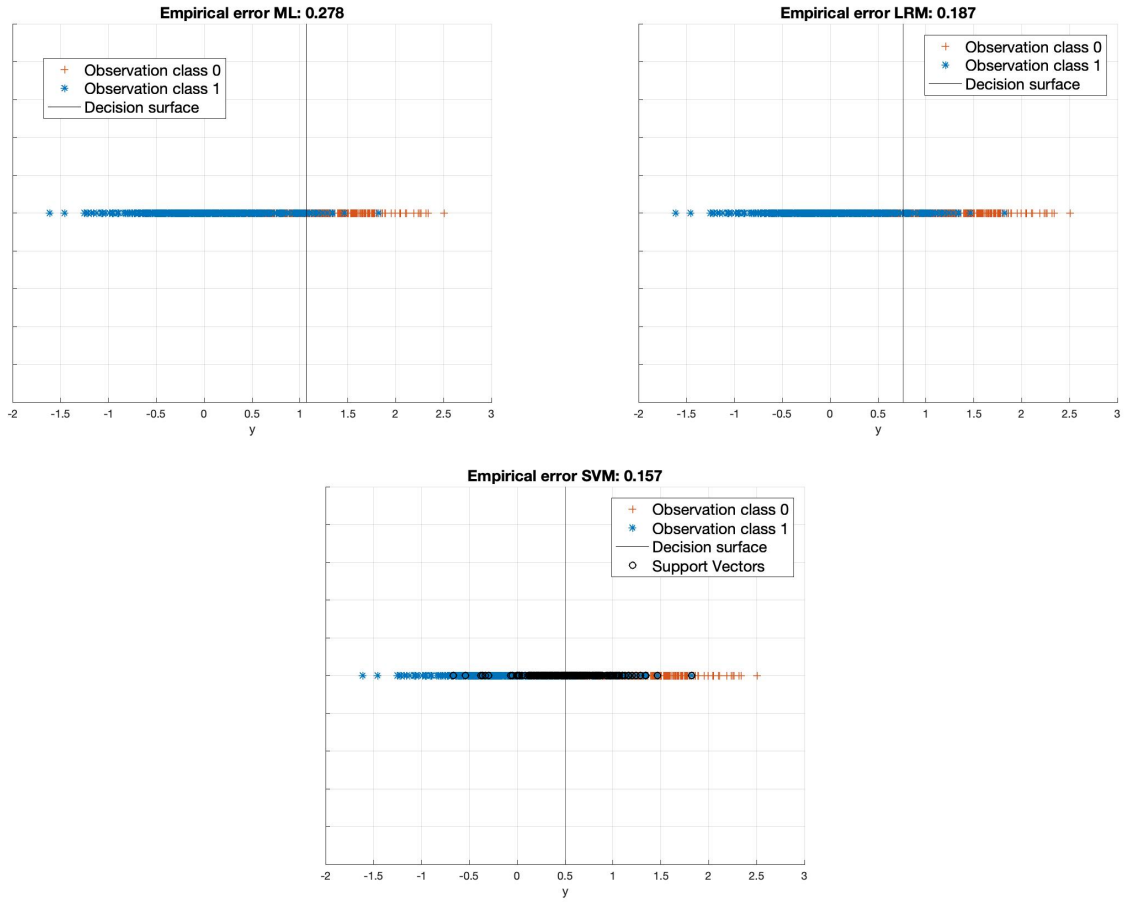

Fig. 5: Classification boundaries and empirical errors given the observations (y) in GLM, LRM and SVM ($N = 1000$, DG1).

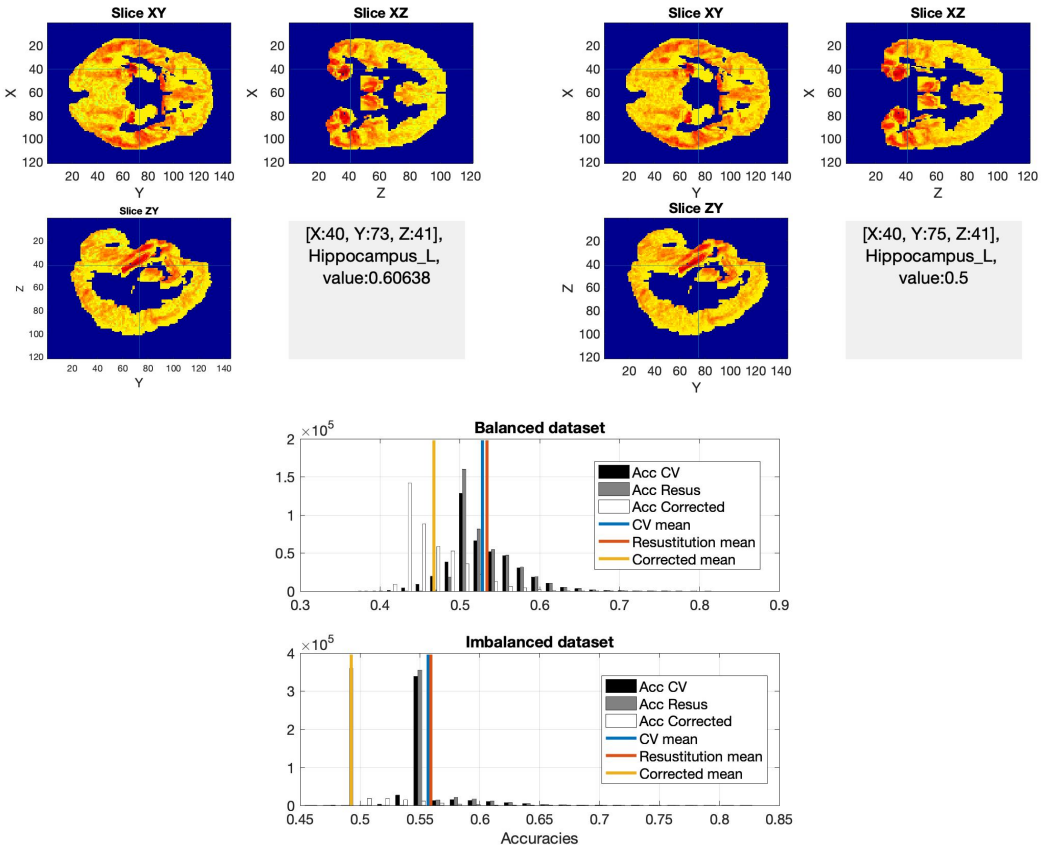

Fig. 6: Bottom: Distribution of voxelwise accuracies of the empirical dataset in two cases: balanced (188 vs 188) and imbalanced groups (188 vs 229), using $k = 10$-fold, resubstitution and concentration inequalities [15]. Up: 3D distribution of the accuracies using k-fold CV and the corrected Accuracy by upper-bounding.

ferences were demonstrated to be under-conservative as they inflate false positives when the analysis is performed on resting-state fMRI data. In this paper, we focus our analysis on specific standardized MRI structures for computational reasons [19], e.g. the left Heschl gyrus region, instead of doing that on whole-volume searches as already analyzed in [15], [7] and are interested in comparing the statistic images derived from the devised methods; i.e. the number of activated voxels and FWE rates that arise from them, rather than a specific inference to control the FWE rate.

In this analysis two groups of subjects ($N = 114$) were randomly drawn from a relatively large ($N = 228$) pool of NC, and the corresponding p-values, e.g. the ones defined in equation 15, were computed accordingly. Thus, the null hypothesis of no group difference in brain activation is true by construction. The proportion of analyses that give rise to any significant results; that is, the number of FPs detected, should be approximately equal to the significance level.

First, we estimated the voxelwise activation rate provided by the uncorrected SPM within standardized areas [34] in this randomization analysis. In this case, each voxel statistic is tested individually and the activation rate for each region is simply the overall number of suprathrehold voxels divided by the number of analysis (1000) times the number of voxels within the region (Nv). Nevertheless, the *ensemble* of such partial results (the omnibus test) using specific inferences provides the estimated FPs given in [15], [7]. The estimated FWE rates are the number of analyses with any significant group activation divided by the number of analyses. From this figure we selected two dummy regions, left Heschl gyrus and left hippocampus, since they are small and provide extreme values for the between-group difference.

A total of $O = 1k$ random group draws were undertaken to obtain the statistic images for each configuration (SPM unc, SPM unc with CET=10 and P-tests) and then, the empirical FP rates were computed on the selected regions using the same Omnibus test. The estimated FP rates are simply the number of significant results divided by the total number of permutations. To establish a fair comparison between parametric (SPM unc, SPM unc $CET = 10$) and non-parametric maps (CV P-tests), we employed the same inference method for these configurations. In particular, we employ a single-threshold test [19] applied to the p-value images of each method and select a threshold α on the frequency that the minimum p-value distribution across the region p_i^{min}, for $i = 1, \ldots, O$ derived from randomization is less or equal the minimum p-value of the test image (see section II-B3). The FP rate is estimated accordingly as the proportion of randomisation values less or equal to the number of occurrences divided by O.

Figure 9 illustrates similar results as those described in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3101662, IEEE Journal of Biomedical and Health Informatics

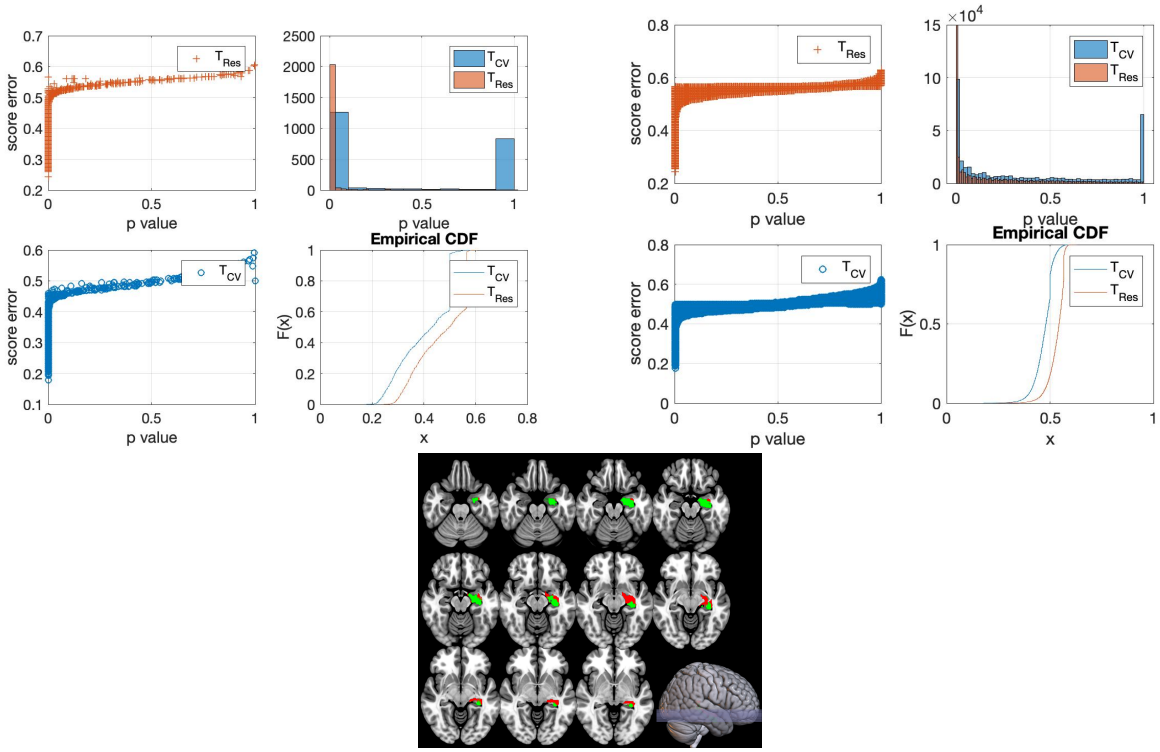

Fig. 7: Permutation analysis of the hippocampus. Note that $O = 1000$ and the upper bound [13] was obtained with a probability at least 0.05 and the similarity of the histograms for the p values derived from the analysis on the hippocampus (left, $O = 1000$) and on whole volume (right, $O = 263$). In both cases the number of regions detected by T_{res} (red map) was larger than that detected by T_{CV} (green map).

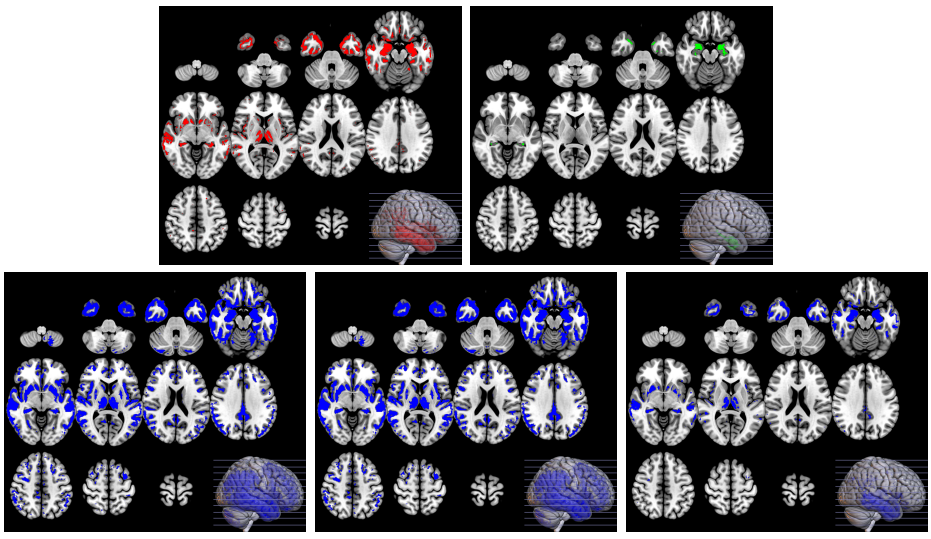

Fig. 8: Parametric and non-parametric statistical maps. Note the trade-off in detection and control of the FWE of the T_{res} approach (red map) compared with T_{CV} (green map) and the three SPM configurations (blue maps): on the left, a cluster-defining threshold of $P = 0.001$ (uncorrected for multiple comparisons), in the middle, adding a cluster extent threshold = 10 voxels, and on the right a Family Wise Error correction at 0.05 on the clusters.

[15], [7] for global analyses. First, in the figure above we show the activation rate for each structure by evaluating the voxel statistic individually. Results are in agreement with the expected significance levels. On the contrary, the Omnibus test on p-value images reports more FPs than expected in all SPM configurations, except for the one based on the over-conservative clusterwise inference, e.g. voxelwise SPM FWE corrected. Although no precise control of FPs is assured, we found our CV-P tests controls FWE below the significance level, whilst the same inference on the SPMs had FWE ranging $20\% - 70\%$. In other words, the methods based on $P = \alpha$ $CET = 10$ voxels has a FWE-corrected P value of $0.2 - 0.7$.

The P-tests (T_{res} and T_{CV}) maps based on the single-threshold test provide a better control of the type I error than those based on SPM, whilst it is worth mentioning that uncorrected SPM-based inferences are clearly dependent on the selected structure, e.g. in the left hippocampus is close to being valid, unlike the results found in previous global analyses [7]. A simple single-threshold inference on the minimum p-value distribution derived from randomization relieves this issue. We also show how the permutation approach based on the upper bounds provides a similar estimated FWE rate as that based on the k-fold CV P-test, but with larger statistical power as shown in the preceding section.

IV. Discussion

In the context of classification for statistical inference, there are two primary strategies, either: i) performing k-fold cross-validation and assessing A_{cc} in several averaged folds; or, ii) proposing a cross-validation based statistic (P-test) using an estimation of the actual error of the classifier on a new set of samples (equation 8). In both cases, if this residual square (error) is small a good classification is achieved and constitutes evidence against the H_0. In the second approach, to simulate the null distribution researchers employ a technique (section II-B3) that is also used in frequentist inference: the permutation test. A set of label permutations, π_p for $p = 1, \ldots, O$, is generated and then applied to the dataset, using the same observations, $\mathbf{y}$, and permuted constraints, x_{π_p}, estimating the parameters w_{pi_p} and computing a set of residuals for all the permutations. The p-value is derived by dividing the number of times we randomly obtain a *residual score* less than the one we obtain with the original value over the number of permutations; i.e. $p - value = p(RS_\pi < RS)$. This methodology is called a CV-P test [31], where LRM could be replaced by SVM or another predictive algorithm.

Several limitations are found using only LRM for estimating the posterior probability. Linear regression is only operative in binary classification, e.g the regression could be negative or even greater than zero [16]. Indeed, and as shown in the example, in this case there is a strong correspondence between GLM and LRM for a single level analysis in group comparisons. Thus, complex classifiers and other loss functions are needed for relieving bad estimations on the set of parameters. Beyond that, the selected predictive algorithms build their P-tests on the CV strategy that could be a biased estimator of the actual error in heterogeneous datasets, such as those encountered in neuroimaging [37].

Frequentist and Bayesian inferences depend on specified models when proposing a T statistic and fitting parameters of the GLM. This is partly solved again by the use of permutation analysis in the estimation of the null distribution, but what about the T statistic definition? This is also described in terms of the error covariance matrix, which must be estimated on empirical data in limited sample sizes. In the synthetic examples we assumed a known covariance matrix in the formulation of the GLM. Despite that, the T-statistic following on from the best guess fluctuated around the ideal value and resulted in low classification rates. How is frequentist or Bayesian analysis actually undertaken? Again, there are model selection and parameter fitting stages to achieve where, in complex scenarios with a limited sample size, heuristics are the common solutions [40]. Indeed, in the high dimensional case or under the assumption of complex models, the performance and operation of these approaches are arguable [31]. Where the estimation of parameters is computationally costly, the tendency is to use heuristics for solving such issues. For example, in the FSL tools based on Bayesian inference, such as BET (Brain extraction tool), TBSS (tract-based spatial statistics), FLIRT (FMRIB's linear image registration tool), PRELUDE/FUGUE (phase unwarping and MRI unwarping), and MELODIC ICA, the use of heuristics is common practice and the estimation of the full posterior distribution of model parameters is biased.

In summary, limited samples sizes and the selection/estimation of any specific model is still an issue in neuroimaging, made more difficult when the model and the interaction between model parameters becomes too complex for an accurate posterior probability estimation, or a feasible numerical computation of the Bayes rule. Given the connection between the two observation models - GLM and LRM - in this paper we propose a statistical inference that leverages an agnostic theory about the estimation of dependencies, established in the pattern classification problem with limited amounts of data [36], [17].

In this sense, given the connection between the two paradigms of statistical inference, we are supported by MLE algorithms to provide new statistical tests, e.g. the P-tests, to highlight differences in patterns of imaging-derived measures between groups. The P-test based on the upper bound correction provides the same type I error control as the k-fold CV approach, and a trade-off in statistical power between clusterwise inferences (invalid for global analyses) and over-conservative voxelwise parametric and k-fold CV P-test inferences, as shown in the experiments.

V. Conclusions

In this paper we propose the application of permutation tests and agnostic theory to the set of regressed outputs by the definition of the residual score or A_{cc} test. The latter framework is a consequence of the connection between the classification problem and statistical inference based on the GLM. Then, we employed permutation tests and a better estimation of actual error based on concentration inequalities to provide a trade-off between the type I error and statistical

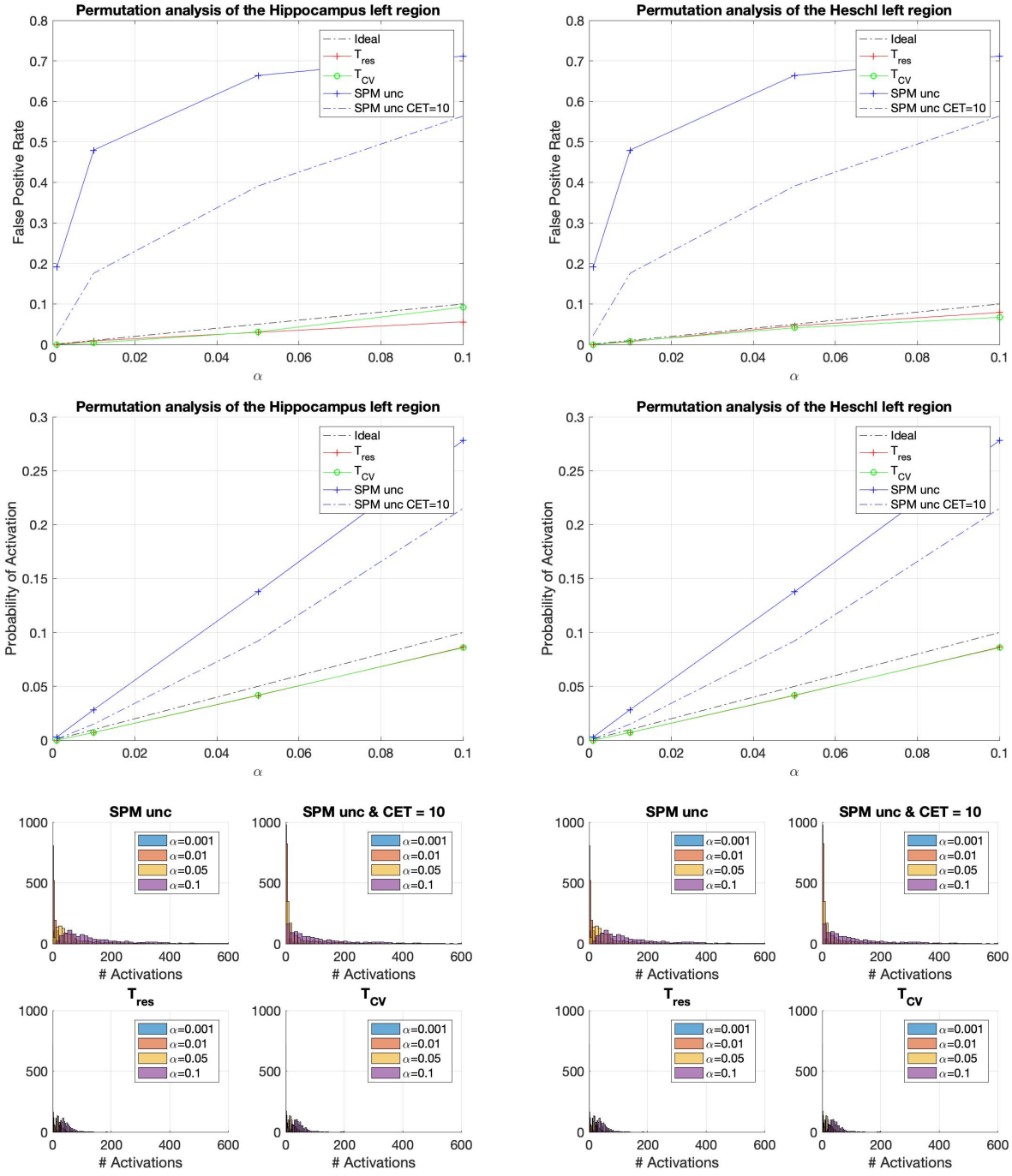

Fig. 9: Estimated activation and FP rates and histograms of the activated voxels for the selected structures at given significance levels α. On the left, the Left hippocampus analysis: we show at the top the FP rates derived from the Omnibus test. In the middle, the probability of activation for individual analyses (type I error control) and the histogram of activated voxels (counts vs. number of activated voxels per test). On the right, left Heschl gyrus analysis. Note: no FPs were detected using over-conservative voxelwise SPM inference.

power. Previous results have demonstrated the ability of such estimator to provide maps of significance [15] where a random simulation on controls resulted in a nominal rate of FPs.

In particular, we see equivalence in the estimation of the observation and (explanatory) label domains, thus any test performed in the label space using an A_{cc} test is similar to those used in neuroimaging over the last decade. Moreover, prevalence (the scores in equations 14 and 16) is a valid measure for statistical inference without using any model as first assumptions. Our approach computes this score using all the data available, instead of using a k-fold strategy, and with the resulting set of accuracies we estimate the true value based on the upper bounds with probability at least $1 - \alpha$. Then, a permutation analysis is derived using this measure to simulate the distribution of the null hypothesis and finally, a test can be formulated as a classic statistical inference. Putative design tasks and random experiments on empirical datasets to assess type I error and statistical power, respectively, confirm the nominal performance of the methodology, and demonstrate its potential.

ACKNOWLEDGMENTS

This work was partly supported by the Ministerio de Ciencia e Innovación (España)/ FEDER under the RTI2018-098913-B100 project, by the Consejería de Economía, Innovación, Ciencia y Empleo (Junta de Andalucía) and FEDER under CV20-45250, A-TIC-080-UGR18 and P20-00525 projects, and by the Ministerio de Universidades under the FPU Predoctoral Grant FPU 18/04902.

REFERENCES

[1] A. Antós, et al. Data-dependent margin-based generalization bounds for classification. Journal of Machine Learning Research 3 (2002) 73?98
[2] C.J.C Burges. A tutorial on support vector machines for pattern recognition Data Mining and Knowledge Discovery, 2 (2) (1998), pp. 121-167
[3] D.Bzdok. Classical Statistics and Statistical Learning in Imaging Neuroscience. Front. Neurosci., 06 October 2017
[4] J. R.Cohen, et al. Decoding continuous behavioral variables from neuroimaging data. Front. Neurosci. 5. 2011.
[5] T.M. Cover. Geometrical and Statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Transactions on Electronic Computers. EC-14: 326?334 (1965)
[6] F. DeMartino, et al. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns NeuroImage, 43 (1) (2008), pp. 44-58
[7] A.Eklund, et al. Cluster failure: Inflated false positives for fMRI. Proceedings of the National Academy of Sciences Jul 2016, 113 (28) 7900-7905.
[8] R.S.J. Frackowiak, et al. Human Brain Function (Second Edition). Chap. 44. Introduction to Random Field Theory. ISBN 978-0-12-264841-0 Academic Press. 867-879, 2004.
[9] K.J.Friston, et al. Statistical Parametric Maps in functional imaging: A general linear approach Hum. Brain Mapp. 2:189-210 (1995)
[10] K.J.Friston, et al. Classical and Bayesian inference in neuroimaging: theory NeuroImage, 16 (2) (2002), pp. 465-483
[11] K. J.Friston, Sample size and the fallacies of classical inference. NeuroImage 81 (2013) 503?504
[12] J.M.Górriz, et al. A Machine Learning Approach to Reveal the NeuroPhenotypes of Autisms. International journal of neural systems, 1850058. 2019.
[13] J.M.Górriz, et al. On the computation of distribution-free performance bounds: Application to small sample sizes in neuroimaging. Pattern Recognition 93, 1-13, 2019.
[14] J.M.Gorriz, et al. Artificial intelligence within the interplay between natural and artificial computation: Advances in data science, trends and applications. Neurocomputing Volume 410, 14 October 237-270 2020.
[15] J.M.Gorriz, et al. Statistical Agnostic Mapping: A framework in neuroimaging based on concentration inequalities. Information Fusion Volume 66, February 2021, Pages 198-212
[16] T. Hastie, et al. The elements of statistical learning theory. Data Mining inference and prediction. Ed Springer. isbn 0-387-95284-5. 2001
[17] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. Information and Computation Volume 100, Issue 1, September 1992, Pages 78-150
[18] R. Heller, et al. 2007. Conjunction group analysis: An alternative to mixed/random effect analysis. NeuroImage 37, 1178-1185.
[19] A.P. Holmes. Nonparametric Analysis of Statistic Images from Functional Mapping Experiments. Journal of Cerebral Blood Flow and Metabolism. 16:7-22
[20] I.A. Illan, et al. Automatic assistance to Parkinson's disease diagnosis in DaTSCAN SPECT imaging. Medical Physics. 2012
[21] C.C.Jack,Jr. ,et al. NIA-AA Research Framework: Toward a biological definition of Alzheimer?s disease. Alzheimers Dement. 2018 Apr; 14(4): 535?562.
[22] I. Kim, et al. Classification accuracy as a proxy for two sample testing Annals of Stat., 2020
[23] R.A.Kohavi. Study of CV and bootstrap for accuracy estimation and model selection. Proc. of the 14th international joint conference on AI - Vol. 2 pp 1137-1143 (1995)
[24] M.A Lindquist, et al. Ironing out the statistical wrinkles in "ten ironic rules". Neuroimage. 2013 Nov 1;81:499-502.
[25] F.J.Martinez, et al. Studying the Manifold Structure of Alzheimer's Disease: A Deep Learning Approach Using Convolutional Autoencoders. IEEE J Biomed Health Inform. 2019 Jun 17.
[26] P.Massart. Some applications of concentration inequalities to statistics. Annales de la Faculté des Sciences de Toulouse, 2000.
[27] 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 Association Workgroup. Alzheimers Dement. 2011;7:263?9.
[28] J.Mouro-Miranda, et al. Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data. NeuroImage, 28, 980?995. (2005).
[29] T.E.Nichols. Multiple testing corrections, nonparametric methods, and random field theory. NeuroImage 62 (2012) 811?815
[30] M.Ojala, et al. Permutation tests for studying classifier performance. Journal of Machine Learning Research. 2010; 11:1833?1863.
[31] P.T. Reiss, et al. Cross-validation and hypothesis testing in neuroimaging: an irenic comment on the exchange between Friston and Lindquist et al. Neuroimage. 2015 August 1; 116: 248?254
[32] J.D. Rosenblatt, et al. Revisiting multi-subject random effects in fMRI: Advocating prevalence estimation. NeuroImage 84 (2014): 113-121.
[33] J.D. Rosenblatt, et al. Better-than-chance classification for signal detection. Biostatistics (2016).
[34] 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 2002; 15: 273-289.
[35] V. Vapnik, A.Y. Chervonenkis On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and Its Applications, 16:264?280, 1971.
[36] V. Vapnik. Estimation dependencies based on Empirical Data. Springer-Verlach. 1982 ISBN 0-387-90733-5
[37] G. Varoquaux. Cross-validation failure: Small sample sizes lead to large error bars. NeuroImage 180 (2018) 68?77.
[38] A.M.Winkler. Permutation inference for the general linear model. NeuroImage Volume 92, 15 May 2014, Pages 381-397.
[39] A.M.Winkler, et al. Non?parametric combination and related permutation tests for neuroimaging. Human brain mapping 37.4 (2016): 1486-1511.
[40] M.K. Woolrich, et al. Bayesian analysis of neuroimaging data in FSL. NeuroImage 45 (2009) S173?S186.