View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Apollo

On the computation of distribution-free performance bounds: application to small sample sizes in neuroimaging

Juan M. Górriz[*,a,b,2], Javier Ramirez[b,1], MRC AIMS Consortium, John Suckling[a,3]

[a] *Department of Psychiatry, University of Cambridge, CB20SZ, UK*
[b] *Department of Signal Theory, Telematics and Communications, University of Granada, Granada, 18071 Spain*

Abstract

In this paper we derive practical and novel upper bounds for the resubstitution error estimate by assessing the number of linear decision functions within the problem of pattern recognition in neuroimaging. Linear classifiers and regressors have been considered in many fields, where the number of predictors far exceeds the number of training samples available, to overcome the limitations of high complexity models in terms of computation, interpretability and overfitting. Typically in neuroimaging this is the rule rather than the exception, since the dimensionality of each observation (millions of voxels) in relation to the number of available samples (hundred of scans) implies a high risk of overfitting. Based on classical combinatorial geometry, we estimate the number of hyperplanes or linear decision rules and the corresponding distribution-independent performance bounds, comparing it to those obtained by the use of the VC-dimension concept. Experiments on synthetic and neuroimaging data demonstrate the performance of resubstitution error estimators, which are often overlooked in heterogeneous scenarios where their performance is similar to that obtained by cross-validation methods.

[*] Corresponding author
[1] email: gorriz@ugr.es, jg825@cam.ac.uk
[2] email: javierrp@ugr.es
[3] email: js369@cam.ac.uk

Key words: Resubsitution error estimate, lineal classifiers, upper bounds, neuroimaging, VC dimension

1. Introduction

The pattern recognition problem [40] consists on determining the random class pattern ω, defined on the set of $\{1, \ldots, k\}$, given a set of observations $\mathbf{x} \in \mathbb{R}^d$, without knowing any information about the underlying probability distribution functions (pdf); i.e. $p(\mathbf{x}, w)$. Instead, we exclusively have several realizations of the pair $\{\mathbf{x}_i, \omega_i\}$, for $i = 1, \ldots, l$, that is, the training sample, and a finite set of decision rules $\{\alpha\}$ with cardinality equal to N with a fixed complexity selected by the user; for example, linear classifiers.

Different methods have been proposed in machine learning (ML) for the computation of the *best* classifier, in terms of performance, within a set of decision rules [42, 6, 20]. However, all of them have in common the use of a minimization or maximization strategies of a specific cost function in order to adjust the model. As an example, the Support Vector Machine (SVM) paradigm [43] maximizes the separation margin between classes, whilst the least squares (LS)-based classification [14] minimizes the mean squared error between the estimated and the true outputs. Once the classifier with parameter $\hat{\alpha}$ is designed and selected from the set $\{\alpha\}$, it is expected to provided a low empirical or resubstitution error:

$$P_{emp}(\hat{\alpha}) = \frac{1}{l} \sum_{i=1}^{l} (w_i - F(\hat{\alpha}, \mathbf{x}_i)) \tag{1}$$

where $F(\mathbf{x}, \alpha) : X \to \Omega$ is the selected decision rule with complexity given by α. Nevertheless, the actual performance of the classifier is measured in terms of its generalization ability; that is, the error rate when encountering new patterns $\mathbf{x}_u$ with unknown classes or labels ω_u. Indeed, one of the major problems in pattern recognition is the overfitting problem in high dimensional settings (d) with a small sample size (l), where the designed classifiers are over-adjusted to

the training set, providing an almost vanishing empirical error, but with a poor actual risk or true error rate, which is defined in terms of probability as:

$$P(\hat{\alpha}) = p(F(\hat{\alpha}, \mathbf{x}_u) \neq \omega_u) \tag{2}$$

Unfortunately, in neuroimaging this situation is the rule rather than the exception, since the dimensionality of each observation (millions of voxels) in relation to the number of available samples (hundred of scans) implies a high risk of overfitting [29]. This risk can be also explained in terms of the high probability of the training set to be separable by a given surface in high dimensional spaces [10]. The solution to this problem is multi-fold. In fact, we can overcome this situation by increasing l by resampling methods (i.e. boosting [20] and bagging [6]), or by decreasing d using feature extraction and selection (FES) approaches [18, 16]. In addition, the model complexity is also linked to the concept of overfitting, as it decreases the empirical risk up to a point, the bias trade-off, where the overfitting occurs and the true error rate increases. In general, the estimation of an increasing number of model parameters (complex models) increases the variance of the error estimation, whereas a model with a restricted number of degrees of freedom can have considerably less total uncertainty [4]. To preserve complex models from overfitting, some solutions can be adopted that are well established on cross-validation (CV) methods [31]. In this sense, several authors have studied numerous accuracy estimation methods using complex classifiers [24, 12]. Assessing their validity on real-word datasets with a high number of attributes, the most common method for model selection is ten-fold stratified cross-validation. However, this implies splitting the dataset into folds which could be intractable with small sample sizes [12] and heterogeneous datasets; i.e. classifiers often perform with unacceptably high variability, particularly if l is small.

A possible solution is the use of linear models, that is linear SVMs are regularized, and therefore less prone to be overfitted. Their parameter configuration at the training stage is considered as a relevant measures of variable importance

[7] in feature selection [18], recursive feature elimination [19], lasso regression [38], and elastic networks [46].

This paper considers the use of the resubstitution error estimate when using linear classifiers in small sample sizes and low dimensional scenarios. A novel upper bound on the actual risk is proposed based by the direct application of a theorem in classical combinatorial geometry for linear classifiers, such as the linear SVMs. We provide an analytic expression for the upper bound, tighter than the ubiquitous upper bound on the actual risk based on VC dimension [42], that is applicable for a range of sample sizes l and dimension values d. It is well known that this data-driven error estimate is generally an optimistic estimation of the real error rate $P(\alpha_{emp})$, given the selection rule α_{emp}, and therefore is usually overlooked. However, the major advances in machine learning over the last decades, i.e. [41], are based on this error estimator, as it easily allows the computation of the upper bounds of the actual risk, i.e. $P(\alpha_{emp}) \leq P_{emp}(\alpha_{emp}) + \phi(l, d)$. Moreover, in some heterogeneous-data applications, resubstitution has been demonstrated to be competitive with cross-validation schemes in terms of ranking accuracy, in addition to the enormous savings in computation time afforded by resubstitution [5]. In fact, this is the main difference with other CV-based error estimators that could provide tighter, but more computationally-demanding upper bounds on the actual error rate in general scenarios. This situation forces the reassessment of these bounds in an empirical manner due to the complexity of the underlying problem, with only a limited number of prior studies having tried theoretical modelling [9, 37, 40]. The latter model applied to the neuroimaging paradigm provides a novel insight in the classification of heterogeneous and small sample size data sets, where complex validation procedures fail to reveal hidden patterns in the classification problem. This solution, having been successfully implemented and applied to the analysis of the autistic pattern [17], is theoretically justified in this work and subsequently generalized to other neurological conditions such as Alzheimer and Parkinson Diseases.

2. Methods

In this section we provide the background necessary for the development of the novel upper bounds on the actual risk derived subsequently. It may allow the use of the resubstitution error estimate as a robust estimator, at a high confidence level with probability $1 - \eta$, under some theoretical conditions assumed throughout the paper. In this scenario, other CV-based estimators provide higher variance and consequently may overlook, depending on the sample realization, the variable importance of each dimension, yielding poorer classification results in the pattern recognition problem.

2.1. A background on uniform convergence of means

One of the major advances in machine learning theory has been the application of the law of large numbers, in terms of the third Hoeffding inequality [21], to the minimization of the empirical risk [41]. Given a finite set of rules denoted by $\{\alpha_i\}$, for $i = 1, \ldots, N$, the probability that the actual risk $P(\alpha_i)$ is greater than the empirical risk $P_{emp}(\alpha_i)$ by a small value ϵ, for a fixed rule α_i, is bounded by:

$$P\{P(\alpha_i) - P_{emp}(\alpha_i) \geq \epsilon)\} \leq e^{-2l\epsilon^2} \tag{3}$$

Considering the two-sided convergence of this inequality we can easily rewrite equation 3 as:

$$P\{|P(\alpha_i) - P_{emp}(\alpha_i) \geq \epsilon)|\} \leq 2e^{-2l\epsilon^2} \tag{4}$$

The higher the difference between actual and empirical risks, the lower is the bound to this probability. Thus, we may take the supremum in the set of decision rules and bound its probability as:

$$\begin{aligned} &P\{\sup_i |P(\alpha_i) - P_{emp}(\alpha_i)| \geq \epsilon)\} \leq \\ &\textstyle\sum_{i=1}^{N} P\{|P(\alpha_i) - P_{emp}(\alpha_i)| \geq \epsilon)\} \leq 2Ne^{-2l\epsilon^2} \end{aligned} \tag{5}$$

This inequality holds for all the decision rules α_i including α_{emp}. If we require that this probability does not exceed a threshold η, then we can establish a bound for the actual risk:

$$P\{|P(\alpha_{emp}) - P_{emp}(\alpha_{emp})| \leq \epsilon\} \geq (1 - \eta) \tag{6}$$

where $\epsilon = \sqrt{\frac{ln2N - ln(\eta/2)}{2l}}$.

2.2. On the upper bound of the actual risk for linear classifers

As aforementioned, linear classifiers provide considerably less total uncertainty in the determination of model parameters. In this sense, with a restricted number of degrees of freedom we can easily measure the overall number of classifiers contained in the class of decision rules. A homogeneous linear threshold (HLT) function $F(\alpha, x) : \mathbb{R}^d \rightarrow \{-1, 0, 1\}$ is defined in terms of the weight vector α as:

$$F(\alpha, \mathbf{x}) = \begin{cases} -1 & , \quad \alpha \cdot \mathbf{x} < 0 \\ 0 & , \quad \alpha \cdot \mathbf{x} = 0 \\ 1 & , \quad \alpha \cdot \mathbf{x} \geq 0 \end{cases} \tag{7}$$

naturally dividing the $\mathbb{R}^d$ space into two subspaces or dichotomies $\{\mathbf{X}^+, \mathbf{X}^-\}$ by the hyperplane $\{\mathbf{x} : \alpha \cdot \mathbf{x} = 0\}$.

Definition 1: A set of l vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_l\}$ is in general position (i.g.p) in $\mathbb{R}^d$ space if every subset of d or fewer vectors is linearly independent.

The necessary and sufficient condition for this property is that the probability be zero that any point will fall on any given $d-1$ dimensional subspace (measure zero), making this property more than feasible in real problems. A generalization of the Function-Counting theorem [10] can be given considering that the number p of points that fall on the separation hyperplane is zero, and rewritten as a Lemma for our purposes as:

Lemma 1: Given a training set $\{\mathbf{x}_i, \omega_i\}$, for $i = 1, \ldots, l$, distributed i.g.p. in $\mathbb{R}^d$ and α the class of HLT functions, if the set of roots of any $F(\mathbf{x}, \alpha)$ on the training set is $\emptyset$, then the number of functions in $\{\alpha\}$ is:

$$N(l, d) = 2 \sum_{k=0}^{d-1} \binom{l-1}{k} \tag{8}$$

Under the conditions of this lemma the number of functions $F(\mathbf{x}, \alpha)$ corresponds to the number of different ways of dichotomizing l points i.g.p. in the $\mathbb{R}^d$ space. Thus, we only need to demonstrate the Function-Counting Theorem expressed in equation 8, which has been analyzed by several authors in the past [45, 22, 10] in different manners. See Appendix 1 for a demonstration of this lemma.

3. The one-sided free-distribution upper bound:

Once the cardinality of the finite set of decision rules has been determined $N(l, d)$, a novel upper bound is derived in terms of the one-sided uniform convergence of the means, which means tighter bounds [15]. In particular, we are interested in assessing for a given significance level η and fixed classifier α_i:

$$P\{\sup_i(P(\alpha_i) - P_{emp}(\alpha_i)) > \epsilon\} < \eta \tag{9}$$

where $P(\alpha_i) = P(F(\mathbf{x}, \alpha_i)) \neq \omega_{real}$. Of course, with a sample size $l \to \infty$, the law of large numbers expressed in terms of the third Hoeffding inequality [21] for any functional α_i establishes that:

$$\lim_{l\to\infty} P\{\sup_i(P(\alpha_i) - P_{emp}(\alpha_i)) > \epsilon\} = 0 \tag{10}$$

and the uniform convergence in equation 9 is achieved. In the other case $l < \infty$, the aforementioned inequality can be used to establish the bound of the actual risk as:

$$P\{\Gamma_i > \epsilon\} \leq \sum_{i=1}^{N(l,d)} P\{\gamma_i > \epsilon\} < \eta = N(l, d)\exp(-2\epsilon^2 l) \tag{11}$$

where $\Gamma_i = \sup_i(\gamma_i)$, $\gamma_i = P(\alpha_i) - P_{emp}(\alpha_i)$ and $N(l, d)$ is the finite number of functional dependencies previously determined. Since the inequality is valid for all decision functions $F(x, \alpha_i)$, the actual risk obtained by α_{emp} is bounded with probability $1 - \eta$ by:

$$\gamma_{emp} \leq \sqrt{\frac{1}{2l}\log\left(\frac{N(l, d)}{\eta}\right)} \tag{12}$$

where a similar bound, considering the additional $Q(l, d)$ functions in equation 18, could also be derived.

t

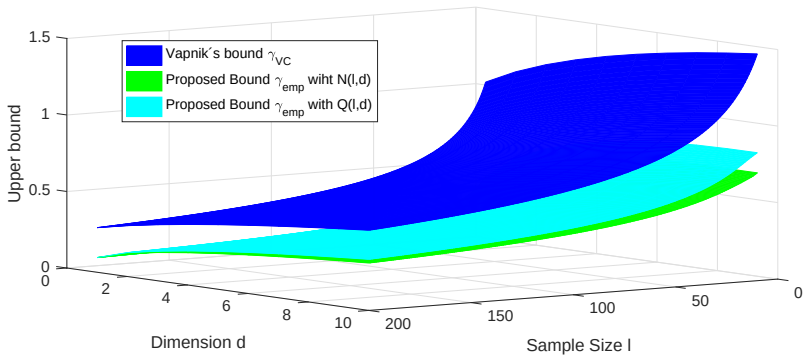


Figure 1: Bounds for the linear classifier using one-sided and two-sided uniform convergence in the Hoeffding inequality at a 95% confidence level.

In general, these bounds could be further improved by considering the relative deviations [41, 15] under scenarios where $P(\alpha_i)$ tends to the extremes $0, 1$; that is, far from our problem. However, we took a step further by limiting the set of $F(\alpha, \mathbf{x})$ to the class of linear classifiers, allowing use of the analytical expression shown in equation 12 for the upper bound (similar to the VC-dimension) that does not depend on the empirical risk, unlike previous approaches [15]. As an example, if $d \simeq l$ then $N(l, d) \simeq 2^l$, the number of functions is such that it separates the sample size in all possible ways (non-falsifiable learning machine), the minimum of the empirical risk is zero, and the upper bound of $P(\alpha_{emp})$ is trivial (> 0.5), independent of the sample size l, at a high confidence level. On the contrary, if $d << l$ the actual risk reaches its maximum value close to the empirical risk, i.e. for $d = \{1, 2, 3\}$ and $l = 120$, the maximum deviation of the frequencies are obtained with probability $1 - \eta$ $(= 0.95)$ as:

$$\gamma_{emp} \leq \{0.1398, 0.1879, 0.2286\} \tag{13}$$

Nevertheless, for complex statistical classifiers even with the $P_{emp}(\alpha_{emp})$ close to 0, the deviation of the frequencies will be very large as the number of decision functions or dependencies N increases, i.e. $N(l, (d+r)!d!r!)$ dichotomies

by an rth-order rational variety [10].

4. Connection to Vapnik's Bound

It is worth exploring the relation of the proposed bound with that commonly used in machine learning, namely:

$$\gamma_{VC} \leq \sqrt{\frac{h(\log(2l/h)+1) - \log(\eta/4)}{l}} \tag{14}$$

where h is the VC dimension equal to $d+1$ for linear functions [42]. This expression is obtained considering the two-sided uniform convergence of equation 9, and is therefore expected to be greater than that proposed in equation 14 making it inadequate for the purposes of working with low dimensional patterns extracted by FE algorithms, and suggesting that the proposed bound is a valuable solution in machine learning to avoid overfitting. Indeed, the two bounds, together with that obtained by the inclusion of $p = 1, \ldots, d-1$ roots of the class of HLT functions α, are plotted in figure 1. The expected behavior is clearly observed and the analitical relationship is shown in Appendix 2.

The bounds obtained by Vapnik [42] are derived under the assumption of i.i.d. samples for the construction of the random entropy $H = E\{N(\mathbf{x}_1, \ldots, \mathbf{x}_l)\}$, or the growth function $G = \ln\left(\sup_{\mathbf{x}_1,\ldots,\mathbf{x}_l} N(\mathbf{x}_1, \ldots, \mathbf{x}_l)\right)$, where the random variable $N(\mathbf{x}_1, \ldots, \mathbf{x}_l)$ is the number of subdivisions of the sample which can be accomplished by the rules $F(\mathbf{x}, \alpha)$. Consequently, the bounds are derived on the rate of uniform convergence from the inequalities of the theory of bounds, as shown in equation 14. The following lemma establishes the connection in terms of Vapnik's theory.

Lemma 2: The analytical i.g.p. sample-based *growth function* $N(l, d)$ majorizes the classical growth function $G(l)$ and, at the same time, is majorized by:

$$\sup_{\mathbf{x}_1,\ldots,\mathbf{x}_l} N(\mathbf{x}_1, \ldots, \mathbf{x}_l) < N(l, d) < 1.5\frac{l^{d-1}}{(d-1)!} < 1.5\frac{l^h}{h!}, \quad \text{for } d < l \tag{15}$$

where h is the largest number of points that can be separated in all possible ways using functions of the given class (VC dimension), i.e. for linear classifiers

$h = d + 1$. Note that the right hand side of the last inequality is often used to derive the looser Vapnik's upper bound in equation 14. The demonstration of this lemma is shown in Appendix 3.

The statistical independence of samples implies their linear independency, but the reverse is not necessarily true. To reduce the number of dichotomies a further, and so the bound, the concept of in general-position (i.g.p.) distributed samples is considered in equation 12. This assumption is reasonable with real data, as shown in figure 2, and is already included in the i.i.d assumption [41]. The novel assumption is especially relevant in brain MR imaging where the research problems are usually solved by locally modeling the MRI signals due to the intensity non-uniformity, partial volume effect, scanner specificity, and the intrinsic properties of tissue types, such as T1 and T2 relaxation times and proton density, which vary across an individual brain [39].

5. Machine learning in Computer-Aided Diagnosis Systems

The application of machine learning methods to neuroimaging data needs to overcome the small sample size problem found in these applications, that is, $l << d$. One of the machine learning tricks for solving this problem is to reduce d by FES approaches prior to Computer-Aided Diagnosis (CAD) classification. FE methods radically reduce the input dimension d by projecting the data into feature spaces where the relevant information in the reduced set of features is preserved [16, 27]. In this scenario, the proposed theoretical upper bound is useful as it effectively connects the empirical and actual risks for linear classifiers within a small confidence interval. Therefore, the resubstitution error estimate using these simple classifiers is an accurate measure of the importance of the variable in the feature space. Moreover, under the *unstable* condition of the inducer [24], which can be met in heterogenous datasets including sub groups or classes that avoid CAD systems to generalize well from training to test sets, the resubstitution estimate could be the only suitable choice.

t

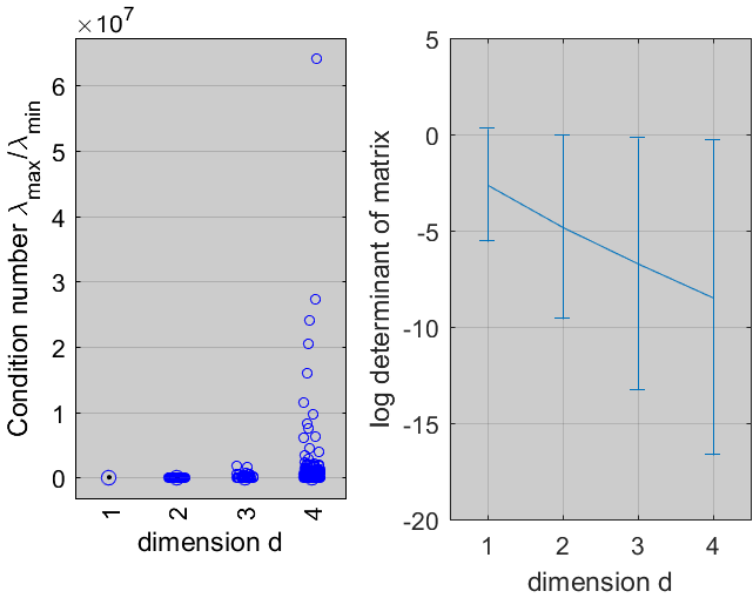


Figure 2: The i.g.p. assumption on real data (Partial least squares features of the dataset in [26]) is fulfilled by analyzing the log determinant and the condition number of the $\binom{l}{d}$ $d \times d$ matrices obtained from the $l \times d$ data matrix.

5.1. Feature Extraction Methods

One of the classical feature extraction methods, the Partial least Squares (PLS) method, is considered at this step to manage the small sample size problem. PLS has been widely used in neuroimaging for performing image analysis and classification [28, 35]. It is a statistical method which models relationships among sets of observed variables by means of latent variables [35]. It includes regression analysis and classification tasks and is intended as a dimension reduction technique [18]. The starting point for PLS is the very simple assumption that the observed data is generated by a system or process which is driven by a smaller number of latent (not directly observed or measured) variables. PLS finds the relationship between the input data $\mathbf{X}_{l\times d}$ and the set of labels $\mathbf{Y}_{l\times 1}$ as linear combinations of the score matrices via the matrices of loadings assuming an error matrix ($\mathbf{X} = \mathbf{X}_s\mathbf{X}_l^T + \mathbf{E}$). The main idea of this dimension reduction is to truncate the number of components to the first k components ($k < d$), thus the $d \times k$ loading matrix $\mathbf{X}_l$ contains the transformation of the d original features to the new k-dimensional space, that is, $\tilde{\mathbf{X}}_{l\times k} = \mathbf{X}_{l\times d}\mathbf{W}_{d\times k}$.

5.2. Ova Multi-label Classification with a linear SVM classifier

In Error-Correcting Output Codes (ECOC) we combine binary dichotomizers to solve a multi-label classification problem [13]. In the one-versus-all (ova) ECOC strategy, all the classes are considered by each dichotomizer as a member of one or both possible partitions of classes that define each binary problem. This results in K binary problems for a given K-class problem. In this paper we applied this strategy to the neuroimaging example, analyzed in the following section, to develop a MRI-based CAD system within the methodological framework proposed in [32], that includes feature selection using ANOVA, PLS-based feature extraction and a classification stage based on linear classifiers such as SVM. The latter classifiers have been predicated on the minimization of the VC dimension and successfully shown to be a robust solution in classification learning [42] that minimizes the separation margin between the binary-labeled training data by constructing an HLT decision function $F(\alpha, \mathbf{x})$ whose norm is

minimum [42]:

$$||\alpha||^2 + C\sum_{i=1}^{l}\xi_i \tag{16}$$

subject to

$$\omega_i(\alpha \cdot \mathbf{x}_i) \geq 1 - \xi_i; \quad \xi_i \geq 0; \quad i = 1, \ldots, l$$

where ξ_i are slack variables, C is a constant that allows a trade-off between training error and model complexity, and the decision rule is defined as $F(\mathbf{x}, \alpha)$ in previous sections. The solution is computed using $\alpha = \sum_{i=1}^{l} a_i y_i \mathbf{x_i}$, where the multipliers $0 \leq a_i \leq C$ were derived using the sequential minimal optimization [30] of a dual Lagrangian problem in equation 16.

6. Experimental Analysis

Experiments were run on an Intel(R) Core(TM) i-5-5200U CPU at 2.20 GHz with 8GB RAM. Several experiments were carried out to analyze the performance of the resubstitution error estimate in combination with FE methods and bagging ensemble learning approaches. In particular, ensembles of linear SVMs with increasing dimensionality of the input vector were used to train the most representative method of SVM: the SMO implementation developed in [30]. The loss of the resulting predictors was estimated through a ten-fold stratified CV process, a leave-one-out CV and the resubstitution analysis, where the feature extraction methods were applied in the CV loop to avoid overfitting in this neuroimaging example. The synthetic comparisons were performed with increasing task complexity (increasing dimension) to highlight and model a real pattern classification problem found in heterogeneous datasets [26, 44]. These simulations allow an effective estimate of the actual risk and therefore check the theoretical bounds proposed in the previous sections. Given that, we apply our resubstitution estimate to the selected and extracted features derived from a CAD system in neuroimaging using MRI scans.

The dual purpose of the novel upper bounds developed in the previous sections is described in this experimental section. First, given a fixed number of

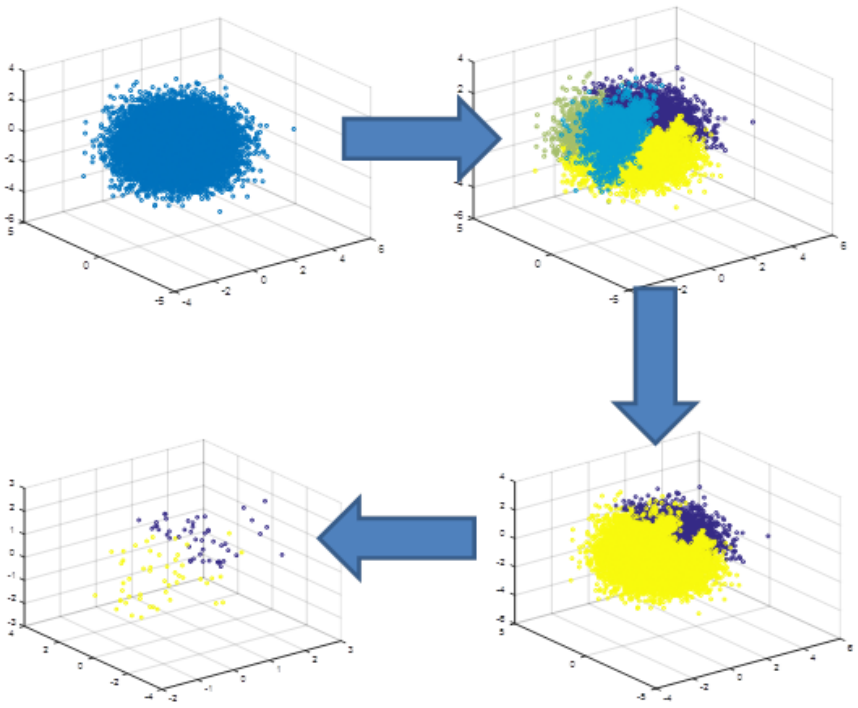
samples (l) and predictors (d) a strong connection is naturally established between actual and empirical risks, depending on the number of decision functions or dependencies N (an additional optimization of the learning parameters of the classifier could strengthen this relation [17]). In this sense, this methodology could be applied to other learning scenarios in neuroimaging, such as the deep learning (DL), where architectures are employed for FE, e.g. stack-autoencoders as a decomposition method [2] that learns low-dimensional representations of an image. Consequently, the low-dimensional features resulting from the output layer of the encoder (Z-layer) could be analyzed with the upper bounds devised in this paper. Second, the upper bounds can be used as a method for determining the optimal sample size l of a machine-learning-based study that best avoids type II (false negative) errors.

This methodology was partially tested on several brain Imaging datasets of autistic individuals corroborating predictions of the "extreme male brain" theory of autism [3] in sexual dimorphic areas, by the evaluation of several statistical tests, such as the spatial overlap analysis on reference maps obtained from different statistical hypothesis testing approaches [17]. As expected, the proposed learning machine revealed how autism was modulated by biological sex using a low-dimensional feature space extracted from VBM, whereas other standard CV methodologies were not able to effectively address the subclass problem in Autism.

6.1. A simulated-heterogeneous dataset with a controlled-actual risk

The developed experiment consisted of generating a large number of samples ($l_T = 20000 \simeq \infty$) drawn from a standard normal distribution $\mathcal{N}(0, 1)$, in an increasing dimensional space $\mathbb{R}^d$, $d = \{1, 2, \ldots, 8\}$. The puporse of this synthetic dataset was to simulate the heterogeneity pattern in Autism MR imaging, i.e. Male Control vs. Male Autism, or in Alzheimer's Disease (AD) structural/functional imaging compared with mild cognitive impairment (MCI). Thus, beyond the classical binary classification problem between two populations drawn from known distributions, an agglomerative hierarchical cluster tree

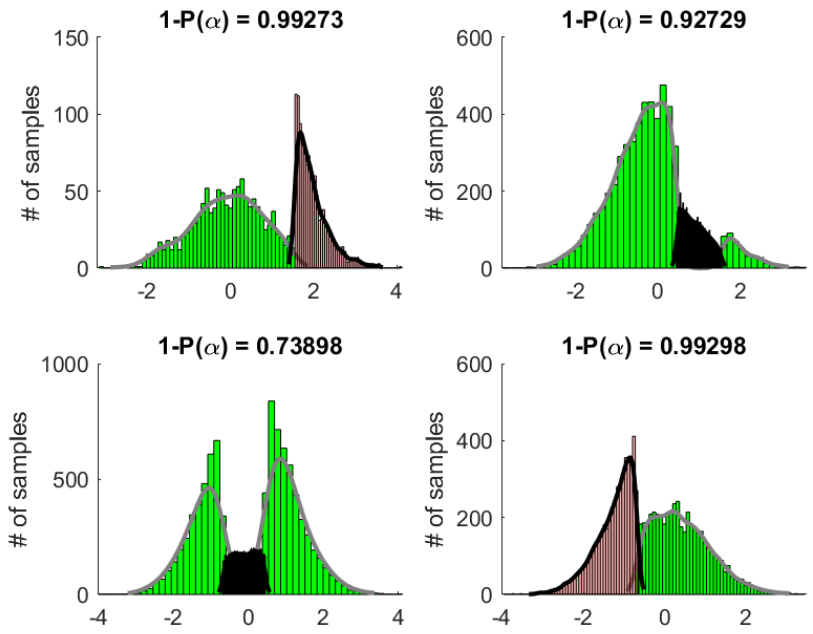
Figure 3: Schematic representation of the data synthesis in three dimensions. The number of clusters of the tree was set to four and the Ward's linkage was used for the optimization procedure. The downsampling procedure of the large dataset results in a reduced set of 100 samples in this example.



was built within the multidimensional generated data. In particular, we minimized the variance within each cluster using the inner squared distance and the Ward's linkage [23]. As a result, we obtained a number of C overlapping clusters, i.e. $C = 4$, as shown in figure 3 (step 1).

One particular cluster models the control class (labeled as 0) whilst the others characterizes a heterogeneous class containing three subgroups (labeled as 1), i.e. MA. Then, for each binary "one versus all" (ova) classification problem, we randomly downsampled without replacement from data $l_S = \{100, 200, \ldots, 500\} << l_T$ realizations and cross-validated the selected linear SVM model in the result-

Figure 4: Example of each ova experiment in one dimension. Observe how multi-modal distributions arise (green) using the modelling procedure described in section 6.1, especially in the "3 out of 4" ova experiment (bottom on the left).

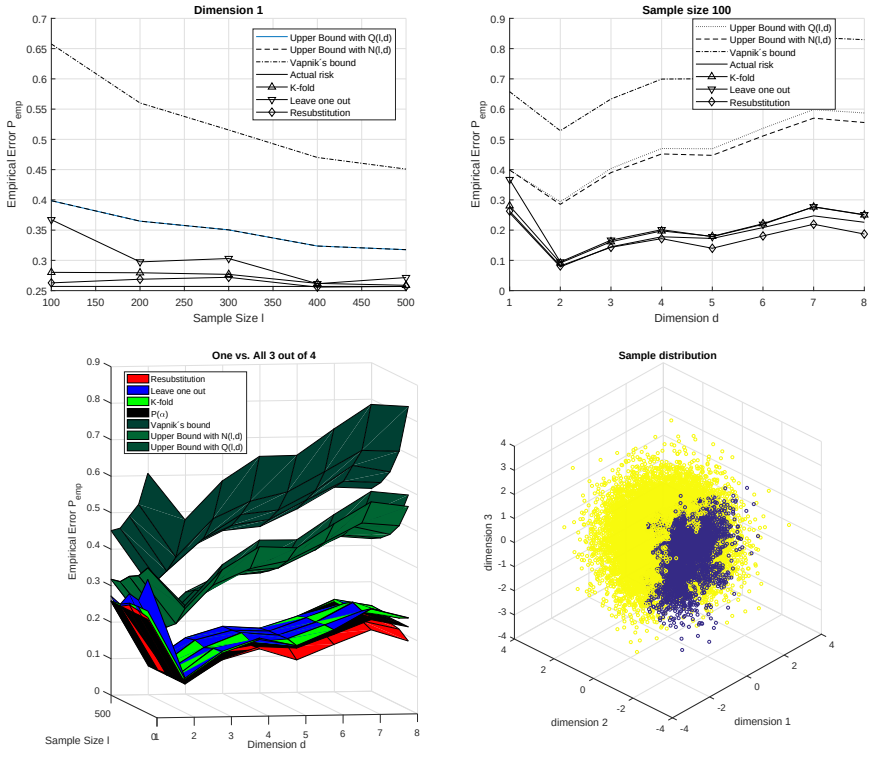


ing binary classification problem. We then compared the resubstitution, the leave-one-out and the K-fold CV estimates. Finally, we repeated the experiment fifty times and averaged the results. To approximate the actual risk (the real error) we fitted the same linear SVM classifier in each ova experiment using all the available samples, i.e $l = 10^3 \sim \infty$ balanced binary samples. Thus, we have a robust estimation of the theoretical error to compare with the empirical risks obtained from each cross-validation model. Each ova experiment defined a different pattern recognition problem, as shown in figure 4, where the actual risk is also displayed.

In figure 5 we show the upper bounds, the averaged error estimates of the resubstitution error and the commonly used CV-estimation errors in a Monte-

Carlo simulation (50 runs). We show, for linear classifiers and low dimension d, the out-performance of the proposed upper bound in comparison with Vapnik's upper bound, which is under conservative. Both limit the empirical risk obtained with all the CV error estimates. In addition, we plot the approximate actual risk computed using the balanced version of the whole data set (20.000 samples) in black. It is worth mentioning from this figure that: i) for the different tasks, the resubstitution error estimate is more optimistic than LOO or K-fold CVs as expected, but very similar to the mean of both; ii) the empirical errors are non-monotonic increasing functions (related to the upper bound behavior) as the dimension (task complexity) increases.

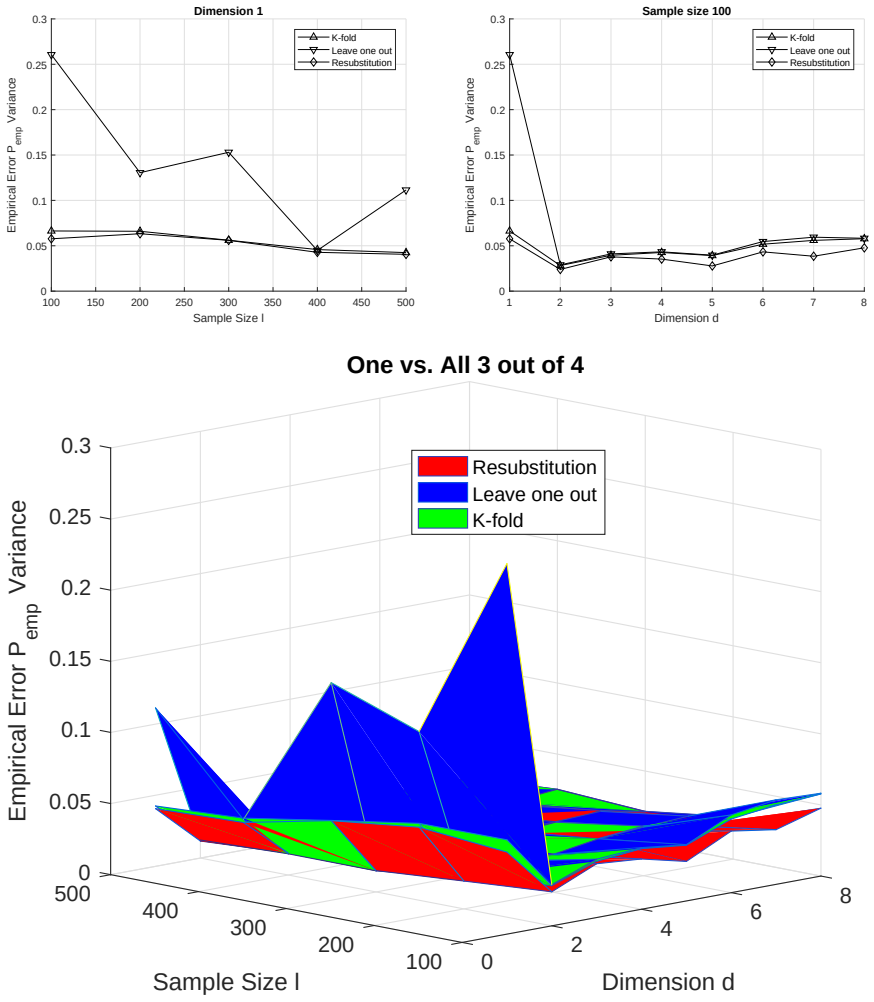
Figure 5: Error Estimates (averaged on 50 runs) in the proposed one vs. all classification problem. Comparison of the resubstitution error estimate with leave-one-out and K-fold cross validations in a low dimension d for several downsample sizes.



From figure 6, we note an interesting finding regarding the variance of the estimations that, under the stability conditions stated in [24], is inversely proportional to the number of samples; that is, $var(P_{emp}(l)) = P_{emp}(l)(1-P_{emp}(l))/l$, approximately. The variance of estimators exhibits a non-monotonically increasing behaviour with increasing task complexity. However, the robustness of the resubstitution estimate to perturbed datasets is observed, unlike the LOO-CV estimate (see for example the "3 out of 4" task in the same figure). The stability condition of the *inducer* must be held in any of the random selected K-folds to obtain robust estimations [24]; i.e. the prevalence of each subgroup within a heterogeneous class must be equal in the folds. As an example, given a population size of l samples, and assuming two subgroups with the same prevalence $l/2$ in the heterogeneous class, the probability of selecting m samples from one subgroup in each fold of l/K samples follows a hypergeometric distribution $p(X = m) = \binom{l/2}{m}\binom{l/2}{l/K-m}/\binom{l}{l/K}$. Thus, the probability of having an unbalanced fold is $P(X \leq m) = \sum_{i=0}^{m} p(X = i)$, for $m < l/2K$. For $l = 100$, $K = 10$ and $m = 4$, $P(X \leq m) \simeq 37\%$ in each fold.

This is related to the inherent problem of failure of cross-validation methods using majority inducers as shown in [24]. The stable condition is partially unfulfilled in the latter task, where the cross-validation methods provide error estimations above the true error rate and the resubstitution estimate (see figure 7), thus increasing the false negative rate in the detection problem. This hypothesis may be tested in heterogeneous datasets including Autism or MCI individuals, which have been described as classes including several subgroups [25, 11]. Thus, the resubstitution error is a good candidate, in low dimensional scenarios and small samples sizes, to estimate the performance of the classifier since it avoids dividing data into folds with a small upper bound. This behaviour was also observed for a different experimental setups with the number of clusters $C = 8, 12$, reinforcing the heterogeneous class.

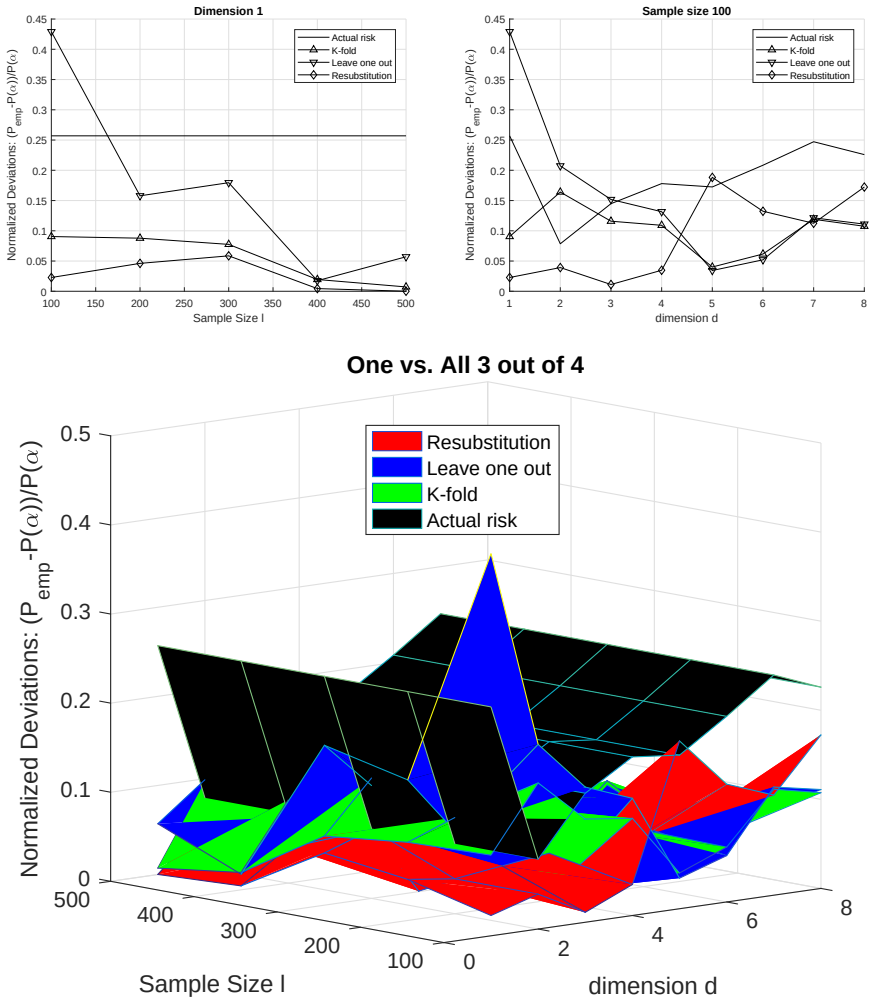
Figure 6: Error Estimate Variances (averaged on 50 runs) in the proposed one vs. all classification problem. Comparison of the resubstitution error estimate with leave-one-out and K-fold cross validations in a low dimension d for several downsample sizes.



6.2. Application to a neuroimaging challenge dataset

Datasets provided by the International challenge for automated prediction of *mild cognitive impairment* from MRI data (https://inclass.kaggle.com/c/mci-prediction) were considered for the evaluation of the proposed method in a neuroimaging context. MRIs were selected from the Alzheimer's disease Neuroimag-

Figure 7: Deviations of the empirical risk under an unstable condition due to perturbed datasets (including subgroups).



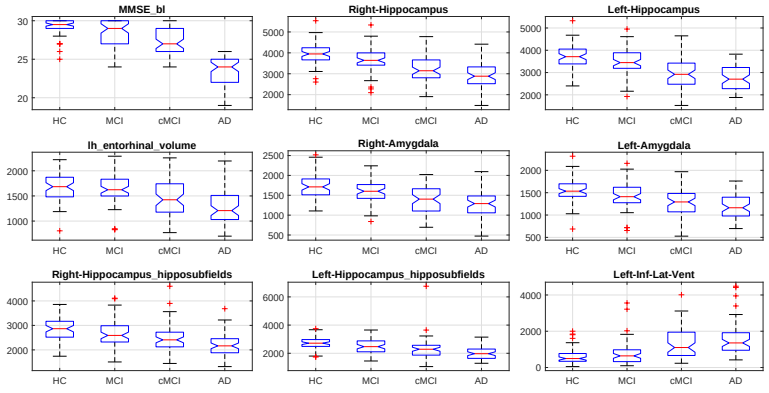
ing Initiative (ADNI, http://www.adni-info.org) and preprocessed by Freesurfer (v5.3) [33]. The dataset consisted of 429 demographical, clinical as well as cortical and subcortical MRI features for each subject. The challenge provided a real classification problem defined on two separated datasets, one for training purposes and another for testing a hold-out set in a four class-classification problem

with unknown labels. Subjects were grouped into four classes according to their diagnosis status: healthy control (HC) subjects, AD patients, MCI individuals whose diagnosis did not change in the follow-up and converter MCI (cMCI) individuals that progressed from MCI to AD in the follow-up period. The training dataset contained 240 ADNI individuals (60 HC, 60 MCI, 60 cMCI and 60 AD). The testing dataset consisted of 500 individuals: 160 real participants and 340 dummy subjects, artificially generated from the real data. The demographic information of the participants can be found in [33, 32]. No information about the class labels of the test set was available during the competition. The test set was half split into public and private test sets and only the accuracy score on the public dataset was available for competitors until the challenge ended. Once the challenge finished, class labels for the images in the test set were provided to the competitors. The accuracy score on the real participants of the testing set was used as the figure of merit in the competition.

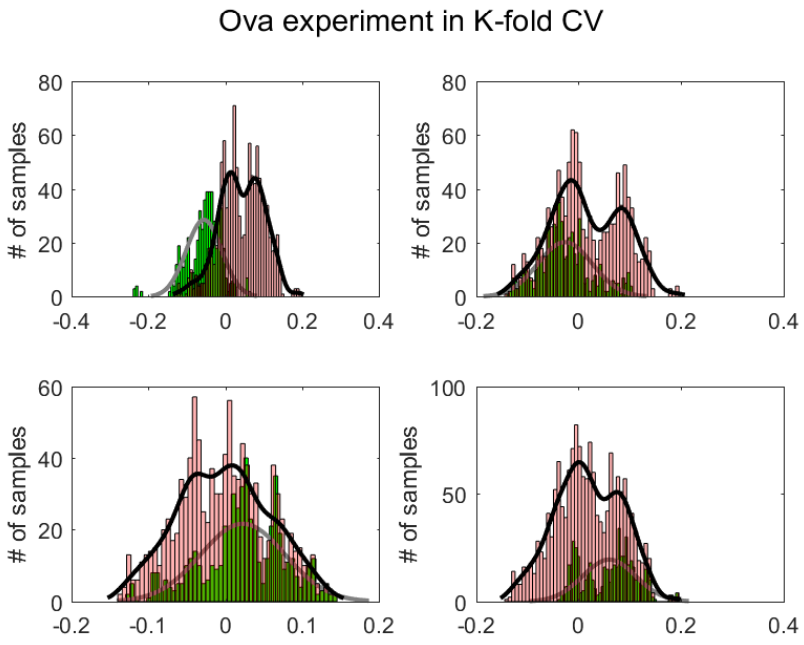
First, we demonstrate that the previous simulations are in accordance with the statistical properties of the PLS features extracted from the MRI-study groups. In figure 8 the group box plots of the main 9 features selected by the proposed ANOVA test is shown, where a "notch" analysis is included to highlight the overlap between classes. In the same figure (bottom) we show the resulting one dimensional PLS-features after FE for each ova experiment. It is clear from the latter figures that each data division into folds including the multi-modal class will provide an unstable inducer as claimed in [24].

In table 1 we show the Acc results when applying Resubstituion vs K-fold to the training set. We notice a larger variance in the K-fold approach and an a-priori optimistic estimation of the Acc provided by the resubstitution-based method. However, this estimation is good enough for low dimensional PLS feature vectors as shown in the figure 9, where the K-fold subestimates the actual risk, and the resubstitution error estimate allows the hidden patterns to be effectively distinguished in each ova experiment. As can be readily seen from the latter figure, the resubstitution error estimate outperforms the K-fold approach up to $d = 10$, where the upper bound (variance) of the estimator increases

Figure 8: Ova experiments for multi-label classification in the CAD system, a) The nine more discriminant features using ANOVA-based feature selection, b) PLS-feature distribution showing multi-modal classes within the remaining classes for each binary dichotomizer.



(a)



(b)

providing an optimistic classification accuracy, unlike the K-fold estimate that provides a pessimistic estimation. The real accuracy (plotted in red) is estimated by the application of the fitted CAD system, using K-fold cross validation, to the real hold-out set (unseen new samples, excluding fake ones). The box-plot on each PLS-dimension for the K-fold estimate reveals that the median of the Acc distribution over the folds is, in 16 out of 20 experiments, lower than the actual risk. Acc values above 55% were ranked in second position of the challenge [8], thus only 10 submissions to the kaggle platform (including one with an Acc of 58.13%) would be enough to achieve such a score using linear classifiers in a ECOC-ova experiment (see table 2).

Table 1: Final performance (Acc(var)) of the 4-label classification system applied to the training set. The class comparisons were extracted from confusion matrices and the Acc results were averaged on folds and # of PLS components

Method	**Overall**	**MCI**	**HC vs. AD**	Elapse Time
	%	%	%	(s)
Resubs.	**57.77(0.27)**	22.38(1.01)	93.17(0.07)	40.45
K-fold	51.10(0.61)	16.13(1.30)	86.08(1.20)	424.95

7. Conclusions

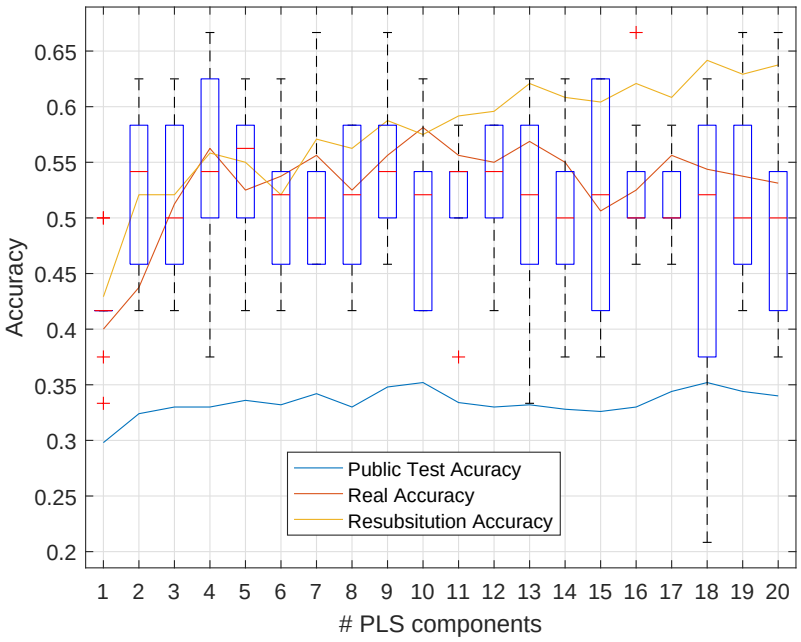
In this paper we focus on the development of novel upper bounds for the resubstitution error estimate. In this sense we provide upper bounds tighter than the commonly used Vapnik' bounds, that are similar in computational demands to our proposed method. To obtain such findings we use some results from classical combinatorial geometry, such as the *Function-Counting theorem*, to estimate the cardinality of the set of linear decision rules or dichotomizers under two assumptions; that is, the number of roots of these functions on the training set is zero or non zero. Moreover, we applied these upper bounds to synthetic and real databases and compared with standard CV-methods. We

Table 2: Ranking of the classifier's accuracies (identified through the name of the teams) as calculated on the platform at the closing of the competition, not including the fake subjects and extended to the whole test set [8].

Team	**Accuracy on test set**	**# of submissions**
S. Dimitriadis/D. Liparas	0.61875	18
SiPBA-UGR	0.5625	56
Sørensen	0.55	41
Bari Medical Physics Group	0.55	15
GRAAL	0.54375	9
Jean-Baptiste Schiratti	0.54375	4
Neuroimage Div./CIFASIS/ARG	0.54375	22
Salvatore C./Castiglioni I.	0.5375	79
Loris Nanni	0.53125	37
BrainE	0.525	38
utaphys	0.525	2
gogogo	0.525	6
ChaseCowart	0.51875	7
agrickard	0.50625	2
fengxy	0.5	9
JocelynHoye	0.5	4
DevinAnnaWilley	0.46875	3
BoyX	0.4625	3
Webiolab	0.2125	2
Proposed CAD System	**0.5813**	10

demonstrate that under unstable conditions (perturbations caused by subclass imbalanced folds at the training stage) the variance of CV-method increases and the resubstitution error estimate is good choice in terms of bias, variance and computational demand.

Figure 9: Accuracy results for the K-fold (box plots) and resubstitution (orange line) approaches at the training stage and the actual error or accuracy on the test set (red line).



Appendix 1

By induction, consider l points $\mathbf{x}_1, \ldots, \mathbf{x}_l$ i.g.p. and the $N(l,d)$ homogeneously linearly separable dichotomies $\{\mathbf{X}^+, \mathbf{X}^-\}$ in d-space. If each of the $N(l,d)$ dichotomies of $\mathbf{X}$ is separable by the set of weight vectors $\{\alpha\}$ then either $\{\mathbf{X}^+ \cup \mathbf{x}_{l+1}, \mathbf{X}^-\}$ or $\{\mathbf{X}^+, \mathbf{X}^- \cup \mathbf{x}_{l+1}\}$ are also separable by the sets of weight vectors $\{\alpha^+\} \subset \alpha$ or $\{\alpha^-\} \subset \alpha$, respectively. In addition, we can build a weight vector $\hat{\alpha} = -(\alpha^- \cdot \mathbf{x}_{l+1})\alpha^+ + (\alpha^+ \cdot \mathbf{x}_{l+1})\alpha^-$ orthogonal to $\mathbf{x}_{l+1}$, i.e. $\hat{\alpha} \cdot \mathbf{x}_{l+1} = 0$, by picking up a couple of separating vectors within the previous sets $\{\alpha^+, \alpha^-\} \subset \alpha$, that separates the dichotomy $\{\mathbf{X}^+, \mathbf{X}^-\}$. Therefore, the projection of the l points to the $(d-1)$ dimensional orthogonal space to $\mathbf{x}_{l+1}$ is also separable in $N(l, d-1)$ dichotomies by the induction assumption. The

total number of dichotomies, after some manipulations, is then:

$$\begin{aligned} &N(l+1,d) = N(l,d) + N(l,d-1) = \\ &2\sum_{k=0}^{d-1}\binom{l-1}{k} + 2\sum_{k=0}^{d-2}\binom{l-1}{k} = \\ &2\sum_{k=0}^{d-1}\left[\binom{l-1}{k} + \binom{l-1}{k-1}\right] = 2\sum_{k=0}^{d-1}\binom{l}{k} \end{aligned} \tag{17}$$

where the last equality holds using Pascal's rule †.

A generalization of this result can be found in [10] (Theorem 4, page 328) by imposing $p = 0$ on the number of functions in α when p points are on the decision surface:

$$Q(l,d) = 2\sum_{p=0}^{d-1}\sum_{m=0}^{d-p-1}\binom{l}{p}\binom{l-p-1}{m} \tag{18}$$

At the same time this result can be demonstrated by induction on l and d [10] or using the results shown in the appendix in [18].

Appendix 2

Furthermore, the relationship between both upper bounds can be obtained as the following. The squared ratio between the two bounds is proportional to:

$$\Gamma_{\gamma_{emp}/\gamma_{VC}} \approx \frac{\log\left(N(l,d)\right)}{(d+1)(\log\left(2l/(d+1)\right)+1)} \tag{19}$$

Assuming that $l = \beta \cdot d$, with $\beta \geq 2$ and using the Badahur's expansion [1] the numerator of equation 19 can be expressed as:

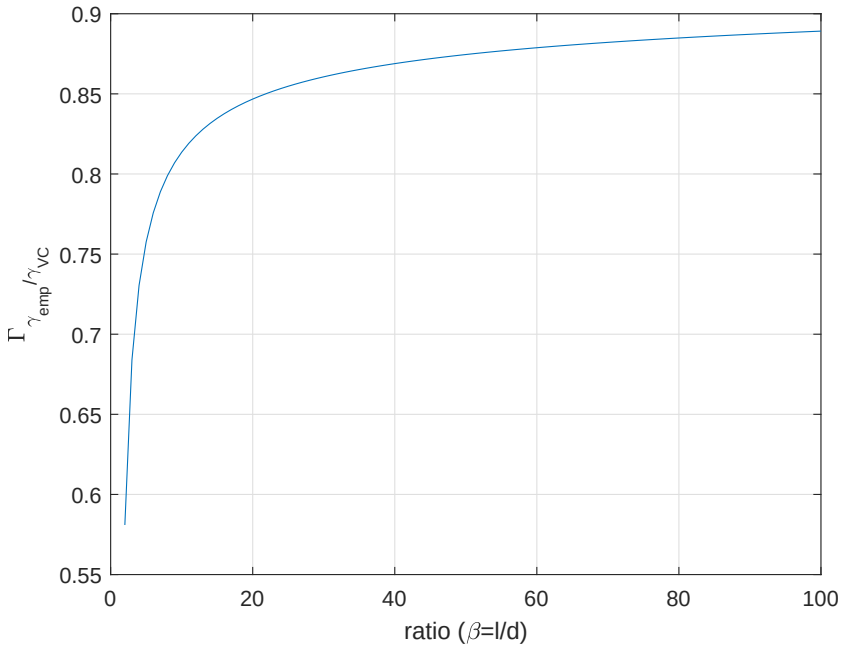
$$\begin{aligned} &\log\left(N(l,d)\right) = log\left(\tfrac{1}{2}\tbinom{l-1}{d-1}F(l,1;l-d+1;1/2)\right) \\ &\approx (l-1)\log(l-1) - (d-1)\log(d-1) - (l-d)\log(l-d) \\ &+ \log\left(F(l,1;l-d+1;1/2)\right)) \end{aligned} \tag{20}$$

where F is the hypergeometric function and the second approximation follows from the Stirling's formula. Taking the limit $d \to \infty$ in equation 19, and after some manipulations, we have:

$$\begin{aligned} &\lim_{d\to\infty}\Gamma_{\gamma_{emp}/\gamma_{VC}} \approx \lim_{d\to\infty}\tfrac{\log(N(l,d))}{(d+1)(\log(2\beta d/(d+1))+1)} = \\ &\lim_{d\to\infty}\tfrac{d(\beta\log(\beta)+(1-\beta)\log(\beta-1)+\log(F(\beta d,1;d(\beta-1)+1;1/2))}{d(\log(2\beta+1))} = \tfrac{(\beta\log(\beta)+(1-\beta)\log(\beta-1)}{\log(2\beta+1)} \end{aligned} \tag{21}$$

where the second term of the numerator term, $\lim_{d\to\infty} \log(F(\beta d, 1; d(\beta-1)+1; 1/2)) = log(\frac{\beta-1}{\beta-2})$, for $\beta \geq 2$, is cancelled by the denominator. Note that this relation is always less than 1 as shown in figure 10, thus our upper bound is theoretically demonstrated to be tighter than Vapnik's bound.

Figure 10: Upper Bound ratio Γ between both approximations for increasing d, where $l = \beta d$ and $\beta \geq 2$.



Appendix 3

In the derivation of a sufficient condition for the uniform convergence of frequencies to their probabilities [41], that is, $\lim_{l\to\infty} \frac{G(l)}{l} = 0$, the growth function is bounded by the function $\Phi(l, d)$ as:

$$\sup_{\mathbf{x}_1,\ldots,\mathbf{x}_l} N(\mathbf{x}_1, \ldots, \mathbf{x}_l) < \Phi(l, d) \tag{22}$$

This function $\Phi(l,d)$ verifies a recursive relation similar to the one found in equation 17 (see page 164 in [41]), and is uniquely determined for $l > 0$ and $d > 0$ as $\Phi(l,d) = \sum_{k=0}^{d-1} \binom{l}{k}$, assuming $\Phi(l,1) = 1$ and $\Phi(l \leq d+1, d) = 2^l$. Surprisingly, this function also appears in other theoretical works in combinatorial theory when introducing the concept of shattering finite sets by classes of measurable subsets [34, 36], i.e. the Sauer-Shelah lemma.

Firstly, it is easy to see that $\Phi(l,d) \leq N(l,d) \leq 2^l$ [4]. Then, following similar steps as in the appendix to chapter six in [41], the inequality $N(l,d) < \Gamma(l,d) \equiv 1.5\frac{l^{d-1}}{(d-1)!}$ is easily proved by checking it on the boundary $l = d+1$ and by the recursion in equation 17. Indeed, on the boundary $N(l,d) < 2^d$, and $\Gamma(l,d) \geq 1.2\frac{1}{\sqrt{2\cdot\pi\cdot l}}e^l$, for $l \geq 5$, using the Stirling's formula. Given, $1.2\frac{1}{\sqrt{2\cdot\pi\cdot l}}e^l > 2^{l-1}$, then the inequality is fulfilled on the boundary. For all $l > d+1$, the inequality is easily demonstrated by an induction on l and verifying that $\Gamma(l+1,d) \geq \Gamma(l,d) + \Gamma(l,d-1)$ †.

Acknowledgement

This study was conducted in association with the National Institute for Health Research Collaborations for Leadership in Applied Health Research and Care (NIHR CLAHRC) East of England (EoE). The authors are extremely grateful for all participants of this study. The Medical Research Council Autism Imaging Multicentre Study Consortium (MRC AIMS Consortium) is a UK collaboration between the Institute of Psychiatry, Psychology & Neuroscience (IoPPN) at King's College London, the Autism Research Centre, University of Cambridge and the Autism Research Group, University of Oxford. The Consortium members are in alphabetical order: Anthony J Bailey (Oxford), Simon Baron-Cohen (Cambridge), Patrick F Bolton (IoPPN), Edward T Bullmore (Cambridge), Sarah Carrington (Oxford), Marco Catani (IoPPN), Bhismadev Chakrabarti (Cambridge), Michael C Craig (IoPPN), Eileen M Daly (IoPPN),

[4] In fact, $N(l,d) = \Phi(l,d) + \binom{l-1}{d-1}$

Sean CL Deoni (IoPPN), Christine Ecker (IoPPN), Francesca HappÃ© (IoPPN), Julian Henty (Cambridge), Peter Jezzard (Oxford), Patrick Johnston (IoPPN), Derek K Jones (IoPPN), Meng-Chuan Lai (Cambridge), Michael V Lombardo (Cambridge), Anya Madden (IoPPN), Diane Mullins (IoPPN), Clodagh M Murphy (IoPPN), Declan GM Murphy (IoPPN), Greg Pasco (Cambridge), Amber NV Ruigrok (Cambridge), Susan A Sadek (Cambridge), Debbie Spain (IoPPN), Rose Stewart (Oxford), John Suckling (Cambridge), Sally J Wheelwright (Cambridge) and Steven C Williams (IoPPN).

This work was partly supported by the MINECO under the TEC2015-64718-R project, the Salvador de Madariaga Mobility Grants 2017 and the Consejería de Economía, Innovación, Ciencia y Empleo (Junta de Andalucía, Spain) under the Excellence Project P11-TIC-7103. Finally, Prof. J.M. Gorriz would like to thank Prof. J. Suckling for the invitation to spend the best summer ever in rainy Cambridge, together with my boss (my wife Elisabet) and my everything Eva and Ivan for supporting my stay.

References

[1] Bahadur, R.R., 1960. Some Approximations to the Binomial Distribution Function. The Annals of Mathematical Statistics 31, 43–54. doi:10.1214/aoms/1177705986.

[2] Baldi, P., 2012. Autoencoders, Unsupervised Learning, and Deep Architectures, in: Proceedings of ICML Workshop on Unsupervised and Transfer Learning, pp. 37–49. URL: http://proceedings.mlr.press/v27/baldi12a.html.

[3] Baron-Cohen, S., 2002. The extreme male brain theory of autism. Trends in Cognitive Sciences 6, 248–254.

[4] Beleites, C., Salzer, R., 2008. Assessing and improving the stability of chemometric models in small sample size situations. Analytical and Bioanalytical Chemistry 390, 1261–1271.

[5] Braga-Neto, U., Hashimoto, R., Dougherty, E.R., Nguyen, D.V., Carroll, R.J., 2004. Is cross-validation better than resubstitution for ranking genes? Bioinformatics (Oxford, England) 20, 253–258.

[6] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont.

[7] Caragea, D., Cook, D., Honavar, V.G., 2001. Gaining Insights into Support Vector Machine Pattern Classifiers Using Projection-based Tour Methods, in: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA. pp. 251–256. doi:10.1145/502512.502547.

[8] Castiglioni, I., Salvatore, C., Ramírez, J., Górriz, J.M., 2018. Machine-learning neuroimaging challenge for automated diagnosis of mild cognitive impairment: Lessons learnt. Journal of Neuroscience Methods 302, 10–13. doi:10.1016/j.jneumeth.2017.12.019.

[9] Chang, M.W., Lin, C.J., 2005. Leave-One-Out Bounds for Support Vector Regression Model Selection. Neural Computation 17, 1188–1222. doi:10.1162/0899766053491869.

[10] Cover, T.M., 1965. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Transactions on Electronic Computers 14, 326–334.

[11] Delano-Wood, L., Bondi, M.W., Sacco, J., Abeles, N., Jak, A.J., Libon, D.J., Bozoki, A., 2009. Heterogeneity in mild cognitive impairment: Differences in neuropsychological profile and associated white matter lesion pathology. Journal of the International Neuropsychological Society : JINS 15, 906–914. doi:10.1017/S1355617709990257.

[12] Efron, B., 1983. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. Journal of the American Statistical Association 78, 316–331. doi:10.1080/01621459.1983.10477973.

[13] Escalera, S., Pujol, O., Radeva, P., 2010. On the Decoding Process in Ternary Error-Correcting Output Codes. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 120–134. doi:10.1109/TPAMI.2008.266.

[14] Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition (2Nd Ed.). Academic Press Professional, Inc., San Diego, CA, USA.

[15] Gascuel, O., Caraux, G., 1992. Distribution-free performance bounds with the resubstitution error rate. Pattern Recognition Letters 13, 757–764. doi:10.1016/0167-8655(92)90125-J.

[16] Gorriz, J., Ramírez, J., Suckling, J., Martínez-Murcia, F., Illan, I., Segovia, F., Ortiz, A., Salas-González, D., Castillo-Barnes, D., Puntonet, C., 2017. A Semi-Supervised Learning Approach for Model Selection Based on Class-Hypothesis Testing. Expert Systems with Applications 90. doi:10.1016/j.eswa.2017.08.006.

[17] Gorriz, J.M., Suckling, J., Lai, M., Lombardo, M., Baron-Cohen, S., Ramirez, J., Segovia, F., Martinez, F., 2019. A machine learning approach to reveal the neuro-phenotypes of autisms. International Journal of Neural Systems 1, 1–22. doi:https://doi.org/10.1142/S0129065718500582.

[18] Górriz, J.M., Ramírez, J., Suckling, J., Illán, I.A., Ortiz, A., Martinez-Murcia, F.J., Segovia, F., Salas-González, D., Wang, S., 2017. Case-Based Statistical Learning: A Non-Parametric Implementation With a Conditional-Error Rate SVM. IEEE Access 5, 11468–11478. doi:10.1109/ACCESS.2017.2714579.

[19] Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning 46, 389–422. doi:10.1023/A:1012487302797.

[20] Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer.

[21] Hoeffding, W., 1963. Probability Inequalities for Sums of Bounded Random Variables. Journal of the American Statistical Association 58, 13–30. doi:10.2307/2282952.

[22] Joseph, R.D., 1960. The Number of Orthants in N-space Intersected by an S-dimensional Subspace. Defense Technical Information Center. Google-Books-ID: WRTUSgAACAAJ.

[23] Jr, J.H.W., 1963. Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association 58, 236–244. doi:10.1080/01621459.1963.10500845.

[24] Kohavi, R., 1995. A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. pp. 1137–1143.

[25] Lai, M.C., Lombardo, M.V., Chakrabarti, B., Baron-Cohen, S., 2013a. Subgrouping the Autism "Spectrum": Reflections on DSM-5. PLOS Biology 11, e1001544. doi:10.1371/journal.pbio.1001544.

[26] Lai, M.C., Lombardo, M.V., Suckling, J., Ruigrok, A.N.V., Chakrabarti, B., Ecker, C., Deoni, S.C.L., Craig, M.C., Murphy, D.G.M., Bullmore, E.T., Baron-Cohen, S., 2013b. Biological sex affects the neurobiology of autism. Brain 136, 2799–2815. doi:10.1093/brain/awt216.

[27] López, M., Ramírez, J., Górriz, J.M., Salas-González, D., Álvarez, I., Segovia, F., Puntonet, C.G., 2009. Automatic tool for the Alzheimer's disease diagnosis using PCA and bayesian classification rules. IET Electronics Letters 45, 389–391.

[28] McIntosh, A.R., Bookstein, F.L., Haxby, J.V., Grady, C.L., 1996. Spatial pattern analysis of functional brain images using partial least squares. NeuroImage 3, 143–157. doi:10.1006/nimg.1996.0016.

[29] Parrado-Hernández, E., Gómez-Verdejo, V., Martínez-Ramón, M., Shawe-Taylor, J., Alonso, P., Pujol, J., Menchón, J.M., Cardoner, N., Soriano-Mas, C., 2014. Discovering brain regions relevant to obsessive–compulsive disorder identification through bagging and transduction. Medical Image Analysis 18, 435–448. doi:10.1016/j.media.2014.01.006.

[30] Platt, J.C., 1998. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Technical Report. Advances in Kernel Methods-Support Vector Learning.

[31] Ramírez, J., Górriz, J.M., Chaves, R., López, M., Salas-Gonzalez, D., Álvarez, I., Segovia, F., 2009. SPECT image classification using random forests. Electronics Letters 45, 1–2.

[32] Ramírez, J., Górriz, J.M., Ortiz, A., Martínez-Murcia, F.J., Segovia, F., Salas-Gonzalez, D., Castillo-Barnes, D., Illán, I.A., Puntonet, C.G., Alzheimer's Disease Neuroimaging Initiative, 2018. Ensemble of random forests One vs. Rest classifiers for MCI and AD prediction using ANOVA cortical and subcortical feature selection and partial least squares. Journal of Neuroscience Methods 302, 47–57. doi:10.1016/j.jneumeth.2017.12.005.

[33] Sarica, A., Cerasa, A., Quattrone, A., Calhoun, V., 2018. Editorial on special issue: Machine learning on MCI. Journal of Neuroscience Methods 302, 1–2. doi:10.1016/j.jneumeth.2018.03.011.

[34] Sauer, N., 1972. On the density of families of sets. Journal of Combinatorial Theory, Series A 13, 145–147. doi:10.1016/0097-3165(72)90019-2.

[35] Segovia, F., Górriz, J., Ramírez, J., Alvarez, I., Jiménez-Hoyuela, J., Ortega, S., 2012. Improved parkinsonism diagnosis using a partial least squares based approach. Medical physics 39, 4395–4403.

[36] Shelah, S., 1972. A combinatorial problem; stability and order for models

and theories in infinitary languages. Pacific Journal of Mathematics 41, 247–261.

[37] Tian, Y., Deng, N., 2005. Leave-one-out Bounds for Support Vector Regression, in: International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06), pp. 1061–1066. doi:10.1109/CIMCA.2005.1631610.

[38] Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58, 267–288.

[39] Tohka, J., Dinov, I.D., Shattuck, D.W., Toga, A.W., 2010. Brain MRI tissue classification based on local Markov random fields. Magnetic Resonance Imaging 28, 557–573. doi:10.1016/j.mri.2009.12.012.

[40] Vapnik, V., Chapelle, O., 2000. Bounds on error expectation for support vector machines. Neural Computation 12, 2013–2036.

[41] Vapnik, V.N., 1982. Estimation of Dependences Based on Empirical Data. Springer-Verlag, New York.

[42] Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, Berlin.

[43] Vapnik, V.N., 1998. Statistical Learning Theory. John Wiley and Sons, Inc., New York.

[44] W Weiner, M., M. Górriz, J., Ramírez, J., Castiglioni, I., 2016. Editorial (Thematic Issue: Statistical Signal Processing in the Analysis, Characterization and Detection of Alzheimer's Disease). Current Alzheimer Research 13, 466–468.

[45] Winder, R.O., 1961. Single Stage Threshold Logic, in: Proceedings of the 2Nd Annual Symposium on Switching Circuit Theory and Logical Design

(SWCT 1961), IEEE Computer Society, Washington, DC, USA. pp. 321–332. doi:10.1109/FOCS.1961.29.

[46] Zou, H., Hastie, T., 2005. Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society, Series B 67, 301–320.