Journal of Advanced Research 80 (2026) 503–533

Contents lists available at ScienceDirect

Journal of Advanced Research

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

Original Manuscript

Statistical agnostic regression: A machine learning method to validate regression models

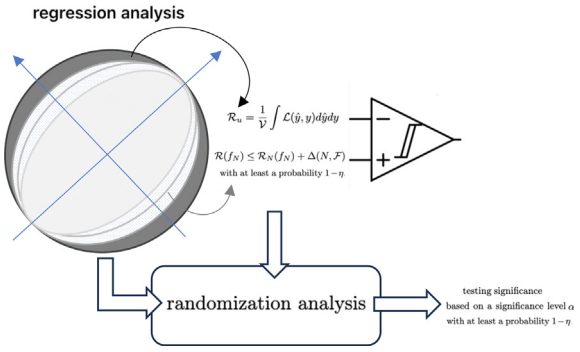
J.M. Gorriz [a,b,c,*], J. Ramirez [b], F. Segovia [b], C. Jimenez-Mesa [b], F.J. Martinez-Murcia [b], J. Suckling [a]

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

HIGHLIGHTS

- A novel approach, SAR evaluates statistical significance in ML-based linear regression models by analyzing concentration inequalities of the expected loss (actual risk).
- SAR introduces a threshold ensuring evidence of a linear relationship in the population, with a probability of at least 1-η1 - η1-η, under non-parametric assumptions.
- Simulations show SAR can emulate the classical multivariate FFF-test for slope parameters, offering comparable analyses of variance without relying on traditional assumptions.
- Residuals computed from SAR balance characteristics of ML-based and classical OLS residuals, bridging gaps between these methodologies.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 29 December 2024
Revised 3 March 2025
Accepted 18 April 2025
Available online 1 May 2025

Keywords:
Ordinary least squares
K-fold cross-validation
linear support vector machines
Statistical learning theory
Permutation tests
Upper bounding

ABSTRACT

Introduction: Regression analysis is a central topic in statistical modeling, aimed at estimating the relationships between a dependent variable, commonly referred to as the response variable, and one or more independent variables, i.e., explanatory variables. Linear regression is by far the most popular method for performing this task in various fields of research, such as data integration and predictive modeling when combining information from multiple sources.

Objectives: Classical methods for solving linear regression problems, such as Ordinary Least Squares (OLS), Ridge, or Lasso regressions, often form the foundation for more advanced machine learning (ML) techniques, which have been successfully applied, though without a formal definition of statistical significance. At most, permutation or analyses based on empirical measures (e.g., residuals or accuracy) have been conducted, leveraging the greater sensitivity of ML estimations for detection.

Methods: In this paper, we introduce Statistical Agnostic Regression (SAR) for evaluating the statistical significance of ML-based linear regression models. This is achieved by analyzing concentration inequalities of the actual risk (expected loss) and considering the worst-case scenario. To this end, we define a threshold that ensures there is sufficient evidence, with a probability of at least $1-\eta$, to conclude the existence of a linear relationship in the population between the explanatory (feature) and the response (label) variables.

* Corresponding author at: Dpt. of Psychiatry, University of Cambridge, UK.
E-mail address: jg825@cam.ac.uk (J.M. Gorriz).

https://doi.org/10.1016/j.jare.2025.04.026

Conclusions: Simulations demonstrate that the proposed agnostic (non-parametric) test can perform an analysis of variance comparable to the classical multivariate *F*-test for the slope parameter, without relying on the underlying assumptions of classical methods. A power analysis on a putative regression task revealed an overinflated false positive rate in standard ML methods, whereas the SAR test exhibited excellent control. Moreover, the residuals computed using this method represent a trade-off between those obtained from ML approaches and classical OLS.

Introduction

Ordinary Least Squares (OLS) is the most popular method to perform linear regression analysis due to its optimal statistical properties assuming the linear model:

$$y = \beta_1^T \mathbf{x} + \beta_0 + \epsilon \tag{1}$$

where y is the response variable or observation, $\mathbf{x}$ is the $P \times 1$ explanatory variable or predictor, β_1 and β_0 are unknown parameters, slope and intercept, respectively, that define the linear regressor (hyperplane) and ϵ is a random variable with zero mean and variance σ^2. The model above can be rewritten if we define $\beta = [\beta_1, ^T \beta_0]^T$ and $\hat{\mathbf{x}} = [\mathbf{x}^T, 1]^T$ as:

$$y = \beta^T \hat{\mathbf{x}} + \epsilon \tag{2}$$

Using a set of observations $\mathbf{y} = [y_1, \ldots, y_N]^T$ and the predictor matrix $\mathbf{X} = [\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N]^T$, the OLS estimator is obtained by minimizing the sum of squares $||\mathbf{y} - \mathbf{X}\beta||_2^2$. Following the Gauss–Markov theorem, the best linear unbiased estimator of any linear function is obtained by OLS if the assumptions of linearity, independence, homoscedasticity, and no perfect multicollinearity are fulfilled [1].

The link between OLS and machine learning (ML) regression methods lies in the fundamental principles of linear regression, which are shared by both traditional statistical approaches and more advanced ML techniques. ML regression methods, such as Ridge regression [2], Lasso regression [3], or Support Vector Machines (SVM) [4], are formulated incorporating a regularizer $||\beta||^2$ into the minimization of the risk functional. In this context, the objective is to find a linear function $f(\hat{\mathbf{x}}) = \beta^T \hat{\mathbf{x}}$ that minimizes the expected loss or actual risk $\mathcal{R}(f) = E[\mathcal{L}(f, \mathbf{x}, y)]$, considering two terms. One term is akin to the OLS approximation and is referred to as the empirical risk, while the other is associated with model complexity. This insight is fundamental to the statistical learning theory (SLT), which serves as a strong foundation for all current artificial intelligence (AI) approaches [5].

Related works

Various approaches have been proposed to validate regression models [6–10]. [6] recommended comparing model predictions and coefficients with theoretical expectations, collecting new data, and employing techniques such as data splitting or cross-validation [7]. In [8] the use of bootstrapping for validating metamodels was introduced, identified as a potent methodology whilst [9], concentrated on logistic regression models and put forth a comprehensive approach that incorporated measures of goodness-of-fit, calibration, and refinement. Moreover, in [10] a comparative analysis of the predictive index accuracy between data splitting and residual resampling bootstraps was conducted, concluding that bootstrapping yields a more precise estimate of goodness-of-fit indexes.

The main drawback of current AI approaches for classification and regression [12,13] is their lack of rigorous analysis of significance in comparison to classical approaches such as the analysis of residuals in OLS linear regression using hypothesis testing. These approaches often limit their analyses to the use of permutation testing on empirical measures derived during the training stages in limited sample sizes, such as p-value analysis using cross-validation (CV). Moreover, although resampling techniques introduce little variation among bootstrap distributions, they can be quite variable, retaining the original variability of the random sample from the population. Randomization-based inference from a small sample may therefore be unreliable [11]. For instance, numerous commentaries in the neuroimaging literature [14–18] highlight the high variability of performance across CV folds in various analytic designs, with clear implications for predictive inference. More critically, in pattern recognition, and particularly in regression problems, there is significant concern about formulating ML analyses exclusively based on learning curves derived from selected loss functions [22] that merely demonstrate that the learning algorithm is converging to a potentially unreliable solution.

Another explored possibility for testing significance is to apply classical analysis on the residuals produced by these AI approaches, which heavily relies on the assumptions of classical statistics, such as Gaussianity. Nonetheless, deep feature extraction analyses are typically conducted in low-dimensional spaces using linear classifiers [19–21]. In this paper, we propose the use of SLT to formulate a non-parametric statistical test for assessing the significance of ML regression models. Initially, we establish an upper bound on the actual risk (expected loss) of a (linear) support vector regressor under the worst-case scenario. Subsequently, we compare the bounded actual risk with that obtained under the null hypothesis H_0, similar to the 50% rate in a classification problem, meaning there is no linear relationship between the predictors and the observed variables. Whenever the *corrected* risk is less than this threshold, we reject H_0 and conclude that, with at least a probability $1 - \eta$, there is a linear relationship between the predictor and observation.

Background, definitions and notation:

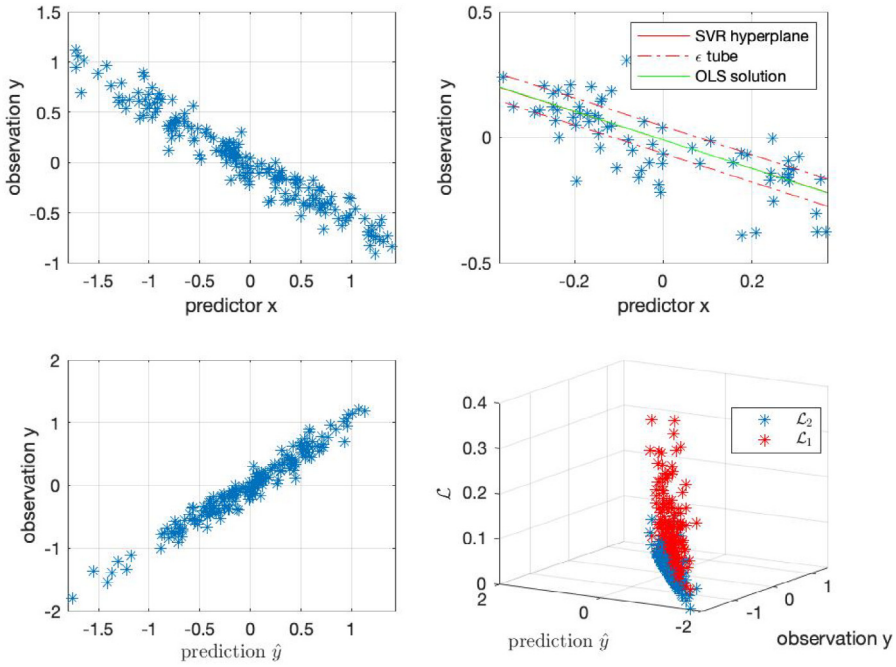
The primary objective of SLT is to establish a structured approach for tackling the challenge of statistical inference, as highlighted in works such as [30,32]. We consider the supervised learning problem where a pair of predictors and observations $\{\mathbf{x}, y\}$ follows an unknown distribution P that draws independent and identically distributed (iid) data. These pairs are used to estimate linear functions $f \in \mathcal{F}$ with a small risk.

$$\mathcal{R}(f) = \int \mathcal{L}(f, \mathbf{x}, y) dP(\mathbf{x}, y), \tag{3}$$

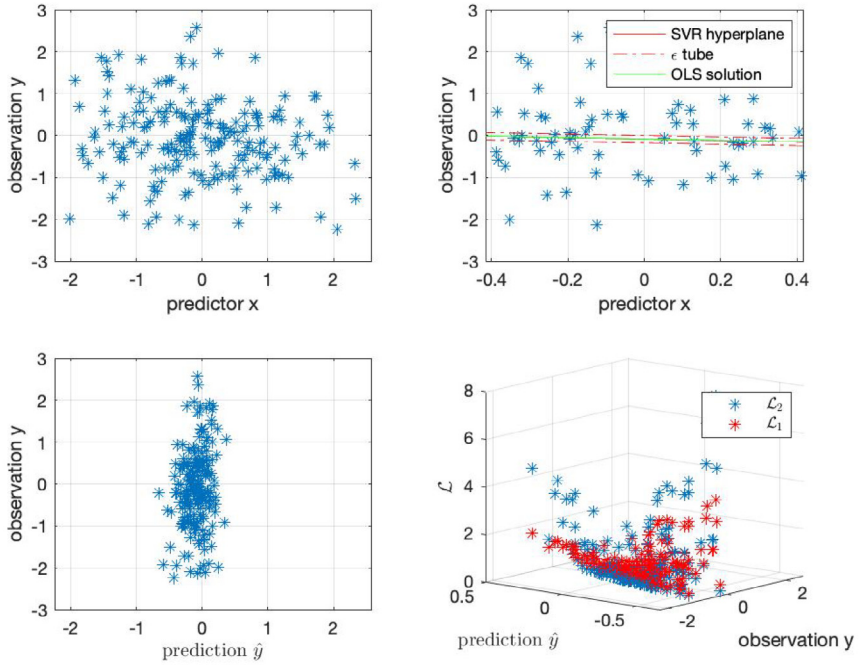
where $\mathcal{L}$ is a loss function, e.g. $\mathcal{L} = (f(\mathbf{x}) - y)^2$. The minimization of the actual risk is only possible through the evaluation of the empirical risk:

$$\mathcal{R}_N = E[\mathcal{L}(f, \mathbf{x}, \mathbf{y}) | (\mathbf{X}, \mathbf{y})] \tag{4}$$

given the sample $\mathbf{X} = \{\mathbf{x}_i\}, \mathbf{y} = \{y_i\}$ for $i = 1, \ldots, N$.



(a) Regression fitting for $\tau = 0.9$.

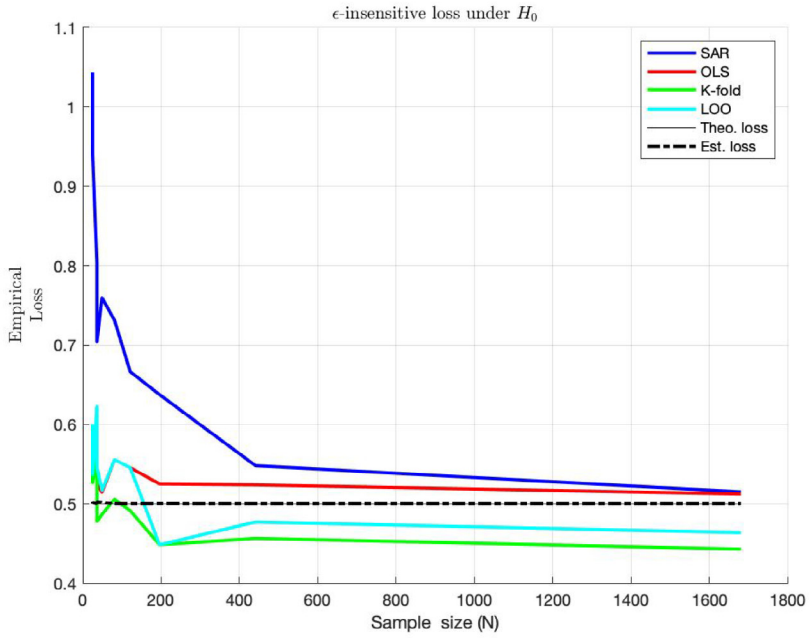


(b) Regression fitting for $\tau = 0.1$.

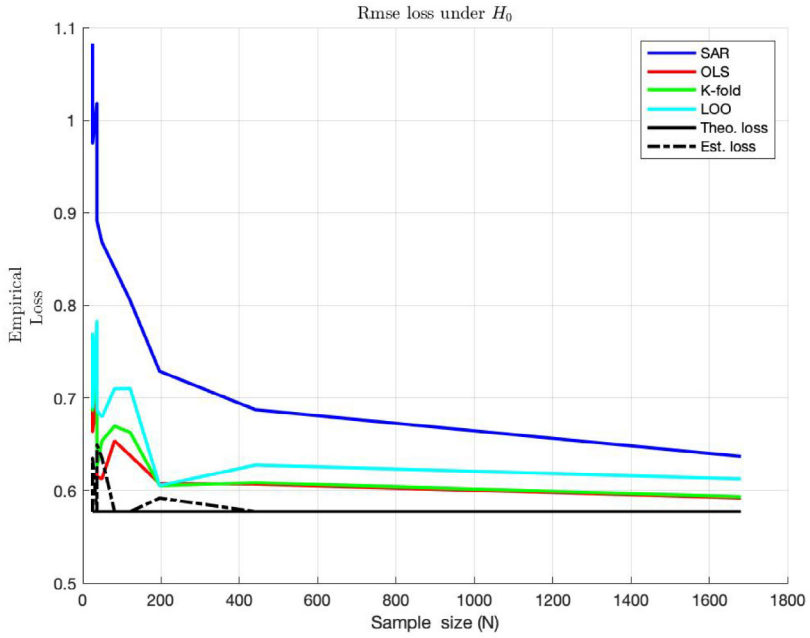
Fig. 1. In this 2D example of regression fitting, we explore effects ranging from large to very small. Observe the flatness of the linear functions in subFig. 1b. In the regression examples of subFigs. 1a and 1b, we present several representations: y vs. x; the result of regression; y vs. $\hat{y}$; and $\mathcal{L}$ vs. (y, $\hat{y}$).

A notable accomplishment of SLT is the development of straightforward and robust confidence intervals that effectively delimit the actual risk $\mathcal{R}(f)$ [33]. Our specific focus lies in the estimation of risk derived from an empirical quantity, ensuring a probability of at least $1 - \eta$. This estimation is articulated through the concentration inequalities (CI) presented as:

$$\mathcal{R}(f_N) \leqslant \mathcal{R}_N(f_N) + \Delta(N, \mathcal{F}) \tag{5}$$



(a) Averaged (empirical) loss under H_0 ($\mathcal{L}_1$).



(b) Averaged (empirical) loss under H_0 ($\mathcal{L}_2$)..

Fig. 2. 2D example (continuation). . .In Figs. 2a and 2b, we plot the empirical losses for all the methods using uncorrelated data within a mesh grid and compare them with the theoretical value.

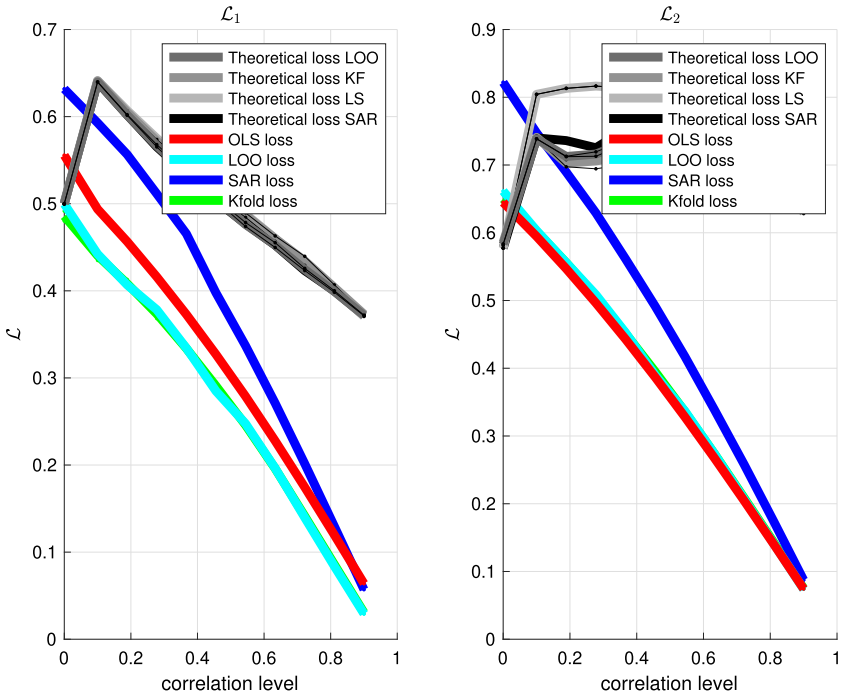
where f_N is carefully chosen to prevent overfitting by restricting the class of functions $\mathcal{F}$, such that:

$$f_N = \arg\min_{f\in\mathcal{F}} \mathcal{R}_N(f) \tag{6}$$

and the term $\Delta(N,\mathcal{F})$ acts as an upper bound for the actual risk, depending on the complexity of the class $\mathcal{F}$ and the sample $Z_N = \{\mathbf{X},\mathbf{y}\}$. In the worst-case scenario, this inequality becomes an equality. This deviation can be understood through various perspectives in classical probability theory offering insights into how closely the sum of independent random variables (empirical risk) are to their expectations (actual risk) [30].

Support vector regression: theory and practice

The Support Vector Regression (SVR) algorithm can be generalized to the case of regression estimation [23], where the sparseness



(a) Averaged losses vs correlation for $\mathcal{L}_{1,2}$

Fig. 3. 2D example (continuation). . .In this figure, we illustrate how the theoretical losses under H_0 for all tested methods enclose the empirical losses, except in the case of uncorrelated data, where the correlation level is zero. This suggests that average losses could serve as a valid measure for testing linearity.

property is preserved by the definition of the ϵ-insensitive loss function:

$$|y - f(\mathbf{x})|_\epsilon = \max\{0, |y - f(\mathbf{x})| - \epsilon\}, \tag{7}$$

where the parameter ϵ is automatically computed [24]. The expected loss to be minimized is the regularized risk functional:

$$\frac{1}{2}||\beta||^2 + \frac{c}{N}\sum_{i=1}^{N}|y_i - f(\mathbf{x_i})|_\epsilon \tag{8}$$

Observe that c is the trade-off constant between complexity and training error. The minimization of Eq. 8 is equivalent to a constrained optimization problem detailed elsewhere [4]. It is important to note that the choice of loss function depends on the specific regression problem at hand [25]. In this paper, we will employ the two most commonly used losses: the squared loss (also known as $\mathcal{L}_2$) and the absolute value loss or $\mathcal{L}_1$ (the insensitive loss for ϵ=0). However, previous studies have indicated that when additive noise deviates from Gaussianity, better approximations to the regression problem are achieved by using estimators based on alternative loss functions, beyond the quadratic loss function employed in OLS [25,23]. To make the two losses comparable, we will employ the rounded quadratic loss in the case of the OLS algorithm whilst ML models are using the least modulo method, e.g., ϵ-insensitive loss. Finally, we point out that the quadratic problem associated with Eq. 8 is connected to support vector classification [24]. Furthermore, the hyperplane that is constructed by this minimization, and lies close to as many of the data points as possible, is a self-supervised training strategy to classify the samples outside the ϵ-tube.

Flatness and the empirical loss in testing for linearity

As an example, and connected to the novel test proposed in the next section, let's assume we are provided with a dataset of observations y and predictors x in 2 dimensions. Starting from an i.i.d. 2D Gaussian distribution, we can define scaling and rotation transforms $\mathbf{T} = \mathbf{SR}$, where $\mathbf{S} = \text{diag}(1, 1-\tau)$, where $\tau \in (0,1)$ is the correlation level, and $\mathbf{R} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ sin(\theta) & \cos(\theta) \end{pmatrix}$. These transforms are applied to the data to create a non-diagonal covariance distribution, i.e. a linear relationship (see Figs. 1a and 1b).Then, we assess the individual losses and the expected loss (average) of the problem in the same figures.

From these examples, we readily see that with decreasing correlation level, the flatness property of the OLS and SVR solutions dominate over the minimization of the empirical risk that, with uncorrelated data, provides an extreme value. For instance, within the framework of the OLS method, the expression for the squared loss function, denoted as $\mathcal{L}_2$, is given by $(f(\mathbf{x}) - y)^2$. Recall that $\boldsymbol{\beta}$ is determined by the OLS solution $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ when the rank of the design matrix $\mathbf{X}$ is equal to the number of observations N [1]. The solution becomes null ($\boldsymbol{\beta} = \mathbf{0}$) when $\mathbf{X}^T\mathbf{y} = \mathbf{0}$, indicating a perfect orthogonality issue in the dataset, which is the opposite of multicollinearity. Therefore, the expected loss simplifies to $\mathcal{R}_N|_{\beta=0} = \frac{1}{N}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) = \frac{1}{N}\mathbf{y}^T\mathbf{y}$.

In this hypothetical case, the effect of sampling and sample size is depicted in Figs. 2, where we additionally illustrate individual and averaged losses converging to the theoretical value. This theoretical value (gray curves in Fig. 3a) could serve as a threshold to establish the significance level of a regression problem: regressions with losses lower than this threshold, with at least a certain probability, allow us to conclude that there is enough evidence of a linear relationship in the population. Under H_0 (correlation level equal to zero), the expected loss can be computed on a meshgrid of uniformly distributed points $(\mathbf{x}, y)$ within a standardized hypercube of $P+1$ dimensions centered at the origin. After some calculus with $P = 1$, we find (see appendix) $\mathcal{R} = \frac{a^2}{6b} + \frac{1}{2}b$ for $\mathcal{L}_1$ and $\mathcal{R} = \frac{b^2+a^2}{3}$ for $\mathcal{L}_2$, where a and b are the maximum values for

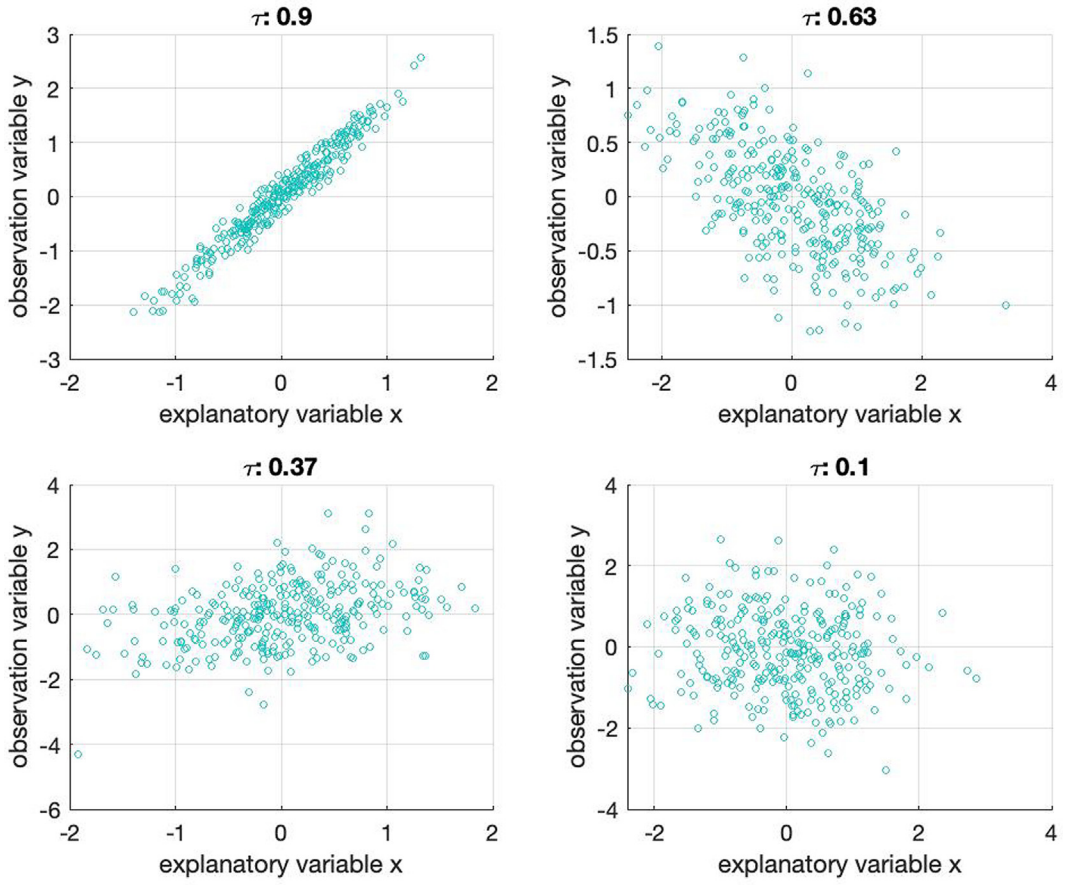


Fig. 4. Data transformed by rotation and scaling with a non-diagonal covariance matrix, assuming a Gaussian distribution.

$\hat{y}$ and y, respectively. These values can be easily approximated under H_0, as illustrated in Figs. 2a and 2b. It is worth mentioning that when using the $\mathcal{L}_1$ loss function, widely employed in predictive models for linear regression, such as K-fold and leave-one-out CV, the results are overly optimistic. This is because they provide convergent values lower than the theoretical expected loss under H_0.

Classical tests of hypothesis in a linear model

A learning algorithm fits a linear function $f(x) \equiv \hat{y} = \beta_1 x + \beta_0$ using a loss function that penalizes the difference between the observation and prediction, e.g., the OLS algorithm. A classical test for linearity on the OLS estimates conducts a test of hypothesis about the regression parameter β_1, where errors are assumed to be independent random quantities normally distributed with mean zero and a common variance. The sampling distributions of the OLS estimates $\hat{\beta}_i$ are indeed normal and a suitable test statistic for testing the null distribution on individual β_i's $H_0 : \beta_i = 0$ against the alternative hypothesis $H_1 : \beta \neq 0$, is the t-test:

$$t = \frac{\hat{\beta}_i}{std(\hat{\beta}_i)} \tag{9}$$

In addition to examining individual hypotheses by the t-test, the classical F-test is defined to test P multiple hypothesis given N observations. It is defined as the ratio

$$F = \frac{MSR}{MSE} \tag{10}$$

with P and $N-P-1$ degrees of freedom, where $MSE = \sum (y_i - \hat{y}_i)^2/(N-P-1)$ is the mean square error and $MSR = \sum (\hat{y}_i - \bar{y})^2/P$ is the mean square due to regression [26].

After linearity is confirmed at a significance level α the quality of the fit can be measured by the magnitude of the t/F-test, the correlation coefficient $Cor(y,x)$, coefficient of determination $R^2 = Cor(y,\hat{y})$, etc.; however they all require the aforementioned assumptions. In the following sections we show how the empirical risk, described as a functional area, can be used as a test for linearity and to assess the quality of the fitting, without the assumptions stated earlier.

A test for linearity using error estimation in ML algorithms

In this section, we present a non-parametric method for testing the null hypothesis utilizing a common measure employed by ML researchers, namely the expected loss $\mathcal{R}$. Under H_0, we expect the loss value obtained with the linear regressor $f_N \in \mathcal{F}$ to be comparable to that of uncorrelated variables. Conversely, this value is expected to be lower when the regression coefficients are significant. The power of the test $(1-\beta)$, control of false positives (FPs) and more statistical properties will be empirically assessed in the experimental section.

The SAR test

The proposed non-parametric test, named the Statistical Agnostic Regression (SAR) test, evaluates the significance level α used in the regression analysis by comparing the actual risk with that

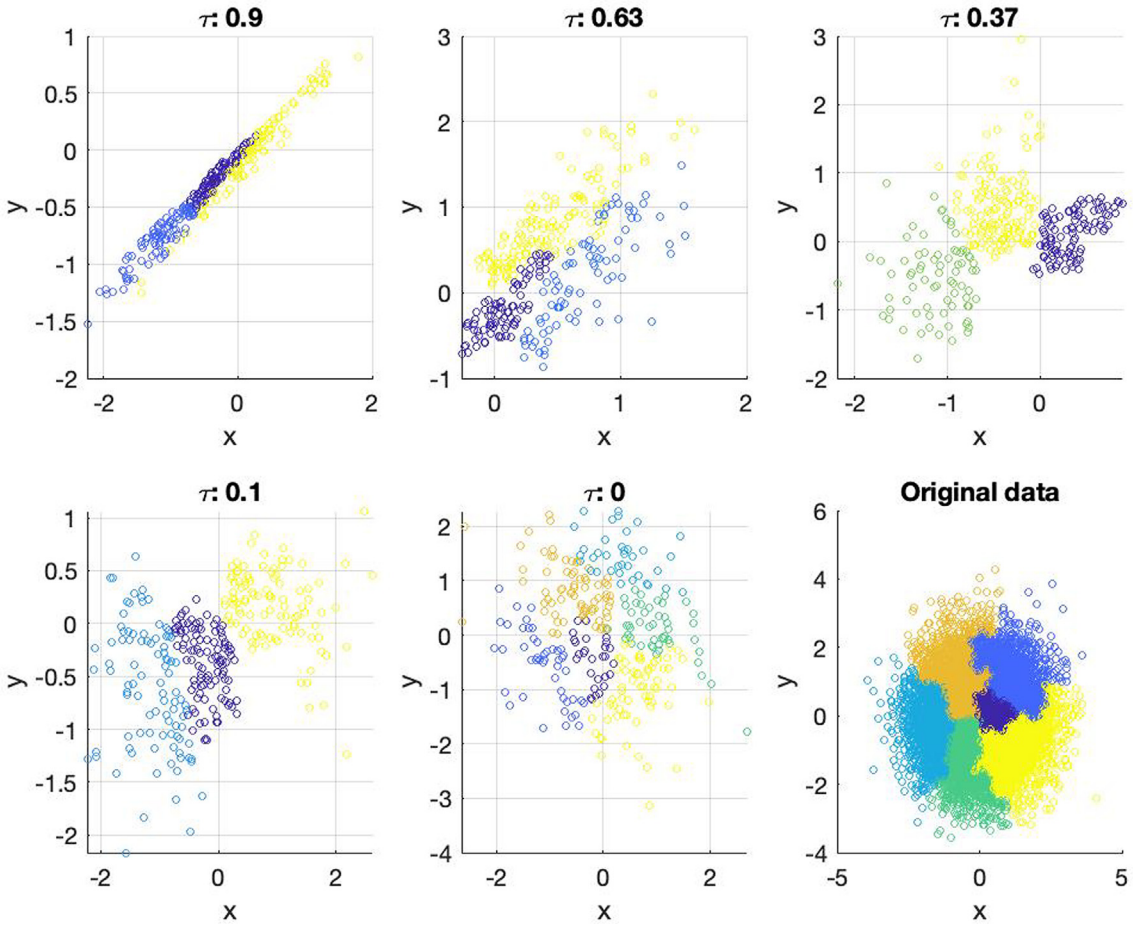


Fig. 5. Gaussian data transformed by rotation and scaling, along with cluster pruning. Colors simply indicate the applied transformations to the data and identify the clusters that were removed.

obtained under the assumption of no correlation between predictors and observations. At a significance level α, the non-parametric SAR test for linearity is formulated as follows:

$$\begin{aligned} H_0 &: \beta = 0 \\ H_1 &: \beta \neq 0 \end{aligned} \tag{11}$$

In this framework, the test statistic is $\mathcal{R}(f_N)$, and it is calculated with at least a probability of $1-\eta$ by considering the *worst-case scenario* as:

$$\mathcal{R}(f_N) = \mathcal{R}_N(f_N) + \Delta(N, \mathcal{F}) \tag{12}$$

where we incorporate the empirical risk $\mathcal{R}_N(f_N)$ and an upper bound $\Delta(N, \mathcal{F})$. We reject H_0 if the $\mathcal{R}$ statistic, computed with at least a probability of $1-\eta$, is less than the critical value γ, or if its p-value, derived after randomization, is less than the level of significance α. Otherwise, we fail to reject it.

Under H_0, the critical value γ is equal to the expected loss when $\beta = 0$; for example, $\gamma = \frac{1}{N}\mathbf{y}^T\mathbf{y}$ for $\mathcal{L}_2$, similar to the perfect orthogonality in OLS. For comparison purposes, we propose an extension of the permutation test, where the threshold γ varies with the sampling process (bootstrapping). We determine how likely the evidence for linearity is by resampling the available data and using the empirical risk distribution, the observed statistic $\mathcal{R}$, compared to the critical value from R realizations. The probability of the observed value of the expected loss under H_0 through permutation π is:

$$p_{value} = \frac{\#(\mathcal{R}^\pi \geqslant \gamma)}{R+1} \tag{13}$$

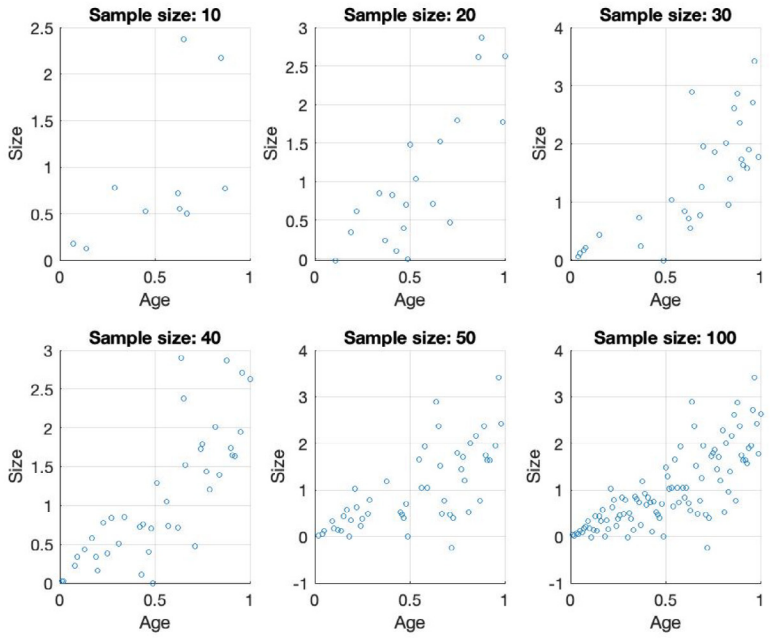
where $\#(.)$ represents the number of times the risk in the permutation is greater than or equal to the error obtained under the null hypothesis H_0. If this p-value is less than our level of significance, e.g. $\alpha = 0.05$, then there is evidence to reject H_0.

The former test is similar to the aforementioned F-test (also known as analysis of variance). The difference lies in the model assumptions, e.g., in classical approaches we assume $E[MSE] = \sigma^2$, and in the methods used to estimate the expected losses. If the aforementioned assumptions are fulfilled, the F^* ratio transforms into an F distribution, and the regression problem becomes a hypothesis test or the analysis of p-values. Even if the linear model assumptions are not fulfilled, we can test our ML measures in the same manner [28] and compare them with the test provided in Eq. 11 (see the experimental section). Moreover, previous approaches to achieve statistical significance using ML and statistical agnostic theory in classification tasks [29] are naturally extended in this regression analysis.

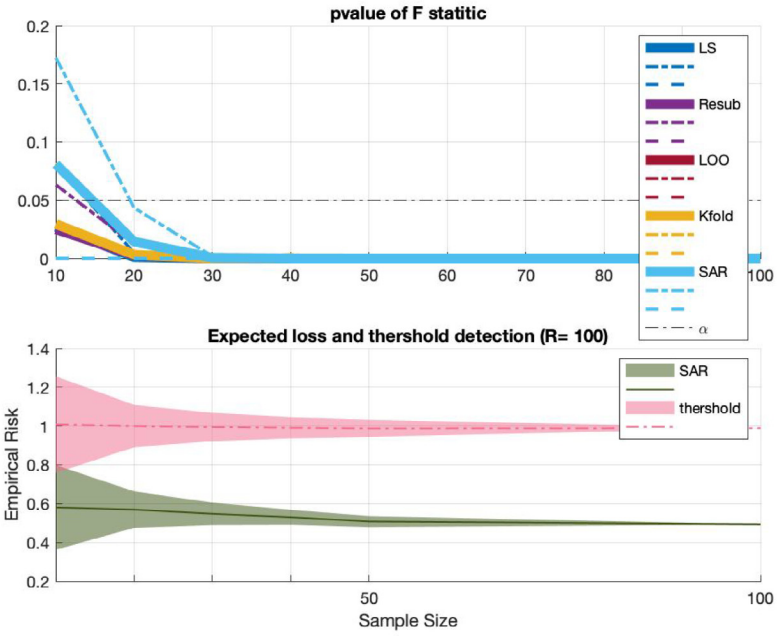
Materials and methods

A probably approximately correct bayesian bound

In this work, we leverage a significant advancement in the field rooted in Probably Approximately Correct (PAC)-Bayesian theory



(a) Datasets.



(b) Tests for linearity.

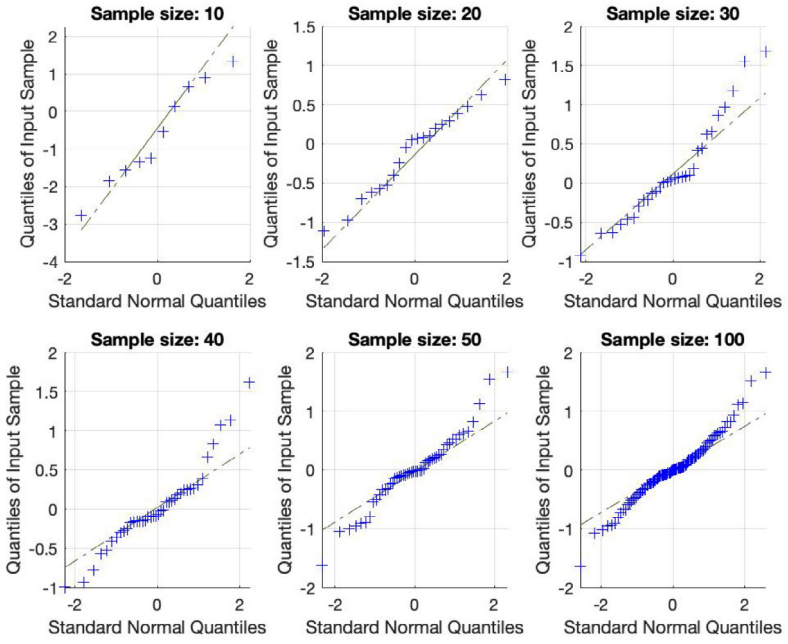
Fig. 6. Dataset with heteroscedascity and increasing sample size in Fig. 6a and SAR and F tests on linearity in Fig. 6b.

[34]. Specifically, we assess a dropout bound inspired by the recent success of dropout training in deep neural networks. The bound, as represented in Eq. 5, is articulated with respect to the underlying distribution Q, which samples the function f from the set of 'rules' denoted as $\mathcal{F}$.

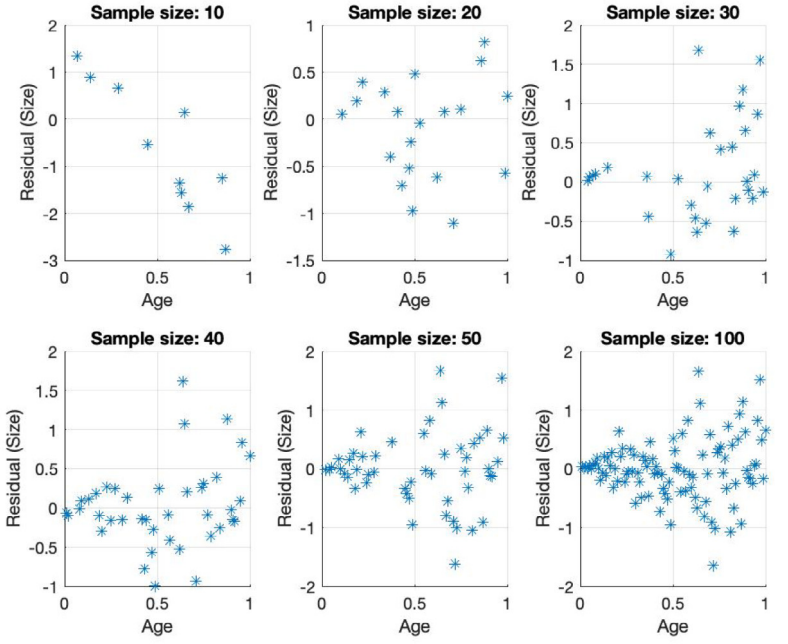
For any constant $\lambda > 1/2$ and a class of linear regressors, $\beta \in \mathbb{R}^{P+1}$, selected according to the distribution Q, we establish that with a probability of at least $1-\eta$ over the sample draw, the following CI are valid for all distributions Q with dropout rate δ:[1]

$$\mathcal{R}(\beta) \leqslant \mathcal{R}_N(\beta) + \min_{1\leqslant i\leqslant k} \frac{1}{2\lambda_i - 1}\left(\mathcal{R}_N(\beta) + \frac{2\lambda_i^2 \mathcal{L}_{max}}{N}\left(\mathcal{D}(Q, Q_u)) + \ln\frac{k}{\eta}\right)\right) \tag{14}$$

[1] Full demonstration in [29], with a preliminary demonstration in the appendix.



(a) Q-Q plot revealing non-Gaussianity.



(b) Explanatory variable vs. residuals.

Fig. 7. Dataset with heteroscedascity (continuation)...Assumptions needed to perform the F-test are not fulfilled as shown in the Q-Q plots in Fig. 7a and residuals vs. explanatory variable plot in Fig. 7a.

Here, $\mathcal{D}(Q,Q_u)=\frac{1-\delta}{2}||\boldsymbol{\beta}||^2$ represents the Kullback–Leibler divergence from Q to the uniform distribution Q_u that is formalised as an isotropic unit-variance prior $\mathcal{N}(0,1)^{P+1}, \lambda \in (1/2,10)$ can assume k different values and $\mathcal{L}_{max}$ is an outlier threshold [34]. Other approaches to formulate upper bounds can be considered based on more general assumptions [30,31].

[1]*ANOVA and the analysis of residuals in ML*

In classical regression [26] results are tested for significance by assessing the analysis of variance of the residuals $r=y-\hat{y}$. As aforementioned, the F^*-statistic is compared to an F-distribution with $P=1$ numerator and $N-2$ denominator degrees of freedom

and the p-value, or the probability that we get this statistic as large as we did under the null-hypothesis, is determined. This is also known as the formal test for the slope parameter β_1 and, if all the model assumptions are fulfilled, can be extended to evaluate the residuals obtained by ML methods.

Another possibility for drawing conclusions about a population (and not only a particular observed sample) is through the use of confidence intervals. Confidence intervals and hypothesis tests are two different ways of learning about the values of population parameters (β_0 and β_1). Both approximations can be conducted on ML residuals as well to demonstrate the reliability of the results, however this is rarely found in the ML literature and there is a trend to exclusively show different performance scores based on empirical measures; e.g. accuracy in learning curves. Similar to the permutation analysis proposed for group comparisons [27], we can test the power $(1-\beta)$ of the proposed SAR test by comparing a set of Monte Carlo simulations with the expected loss under H_0, following the ideas presented in previous sections. We estimate the power $(1-\beta)$ by counting the number of times the corrected empirical loss was less than the averaged loss when $\beta = 0$.

Heteroscedascity and the Breusch-Pagan test

Beyond linearity, we explore the concept of heteroscedascity [35] using the ML approaches vs. OLS. When the residuals are heterogeneously distributed along the explanatory variable (predictor), we encounter the issue of heteroscedasticity in the data. The Breusch-Pagan (BP) test checks the null hypothesis (homoscedasticity) by re-fitting the explanatory (independent) variable to the squared residuals as observation variables. A simple way to assess this condition is by using the test statistic $T = N \cdot R^2$, where R is the coefficient of determination of the auxiliary fitting, and follows a χ^2 distribution with P degrees of freedom under H_0 [36]. If the p-value associated with it is less than the level of significance, we reject H_0. The homoscedasticity condition is assumed in the linear regression model, and when it fails, further analyses based on the F-test are not only imprecise (biased) but also invalidated from a statistical point of view. ML residuals could be tested in the BP test as well, to assess the ability of data-driven models to detect heteroscedasticity in regression problems with limited sample sizes.

Table 1
Summary of datasets and types. ** More information in Section 4.4.1. *** More information in Section 4.4.2.

Parameter	Synthetic Data	ADNI dataset*	Cancer dataset**
Dimension (P + 1)	2	2–7	2–7
Sample Sizes (N)		10–300	
Number of Clusters N_c	1/6	-	-
Total Samples (N_T)	20000	818	2809
Number of Experiments	4	2	
# Correlation Levels	10	-	-

Table 2
Demographics details of the ADNI dataset with group means with their standard deviation.

Status	Number	Age	Gender (M/F)	MMSE
NC	229	75.97 ± 5.0	119/110	29.00 ± 1.0
MCI	401	74.85 ± 7.4	258/143	27.01 ± 1.8
AD	188	75.36 ± 7.5	99/89	23.28 ± 2.0

Synthetic, realistic & real datasets

First, we assessed the 2D problem with Gaussian-distributed variables ranging from perfectly uncorrelated to strongly correlated signals. To control the degree of correlation, we drew samples $\mathbf{Z}$ from a 2D normal distribution, i.e., $\mathcal{N}(0,1)$. Subsequently, we computed a random linear transformation matrix $\mathbf{T}$ and modified its singular value decomposition (SVD) diagonal matrix using a scaling transform similar to the earlier simulated dataset (1). Finally, we applied this transformation matrix to the original data, obtaining $\hat{\mathbf{Z}} = \mathbf{ZT}$ that generated observations and predictors (see Fig. 4). We fitted our models on this data with two losses $\mathcal{L}_1$ and $\mathcal{L}_2$, several samples sizes $N = 10, 20, \ldots, 300$ and tested for significance using the F-test and the SAR test. We drew up to $R = 100$ realizations (sampling) from the ideal distribution and then averaged the regression results to study the effect of sample size [37]. Whenever CV methods split the dataset into folds, we analyzed the variability of the performance measures accordingly. This represents a realistic situation with a fixed realization or dataset instead of having an ensemble of R realizations.

To enhance the realism of our data, we employed the procedure used in [31] to segment the data into N_c clusters. By selectively removing some of these clusters, we introduced non-Gaussian characteristics. Under this experimental setting, the F-test and, in particular, the OLS method are no longer optimal. This allows us to truly test the robustness of the SAR test developed in Section 2.1. We conducted tests for up to four experimental settings by randomly removing a set of $N_c - N_g$ clusters, with $N_g = 3$. In Fig. 5 we plotted the remaining clusters with different colors to illustrate the complete data generation.

Finally, we assessed how the BP test worked on the residuals obtained from the set of methods previously described. Specifically, we evaluated the ability of the resulting models to detect heteroscedasticity versus sample size. For this purpose, we employed the dummy dataset provided in the XLSTAT software [38] designed to compare a homoscedastic model to another with strong heteroscedasticity. In particular, we generated a dependent variable (“Size”) based on an independent variable (“Age”), where the residuals are defined as the product of the independent variable by the random normal error (Fig. 6a). In this case, the residuals were obviously correlated with age, and the problem (Fig. 6) was apparently linear, but heteroscedastic.(see Table 1).

A neuroimaging dataset to study Alzheimer Disease

Data used in preparation of this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI database contains 1.5 T and 3.0 T T1w MRI scans for AD, Mild Cognitive Impairment (MCI), and cognitively normal controls (NC) which are acquired at multiple time points. Here we only included 1.5T sMRI corresponding to the three different groups of subjects. The original database contained more than 1000 T1-weighted MRI images, comprising 229 NC, 401 MCI (252 stable MCI and 149 progressive MCI) and 188 AD, although for the proposed study, only the first medical examination of each subject is considered, resulting in $N = 818$ segmented grey matter (GM) images after standard preprocessing. Demographic data of subjects in the database is summarized in Table 2.

Feature extraction based on principal component analysis (PCA) was performed on the set of segmented GM images to reduce high-dimensional data while retaining key variance and information. A simple pairwise scatter plot and histogram of these novel features for the NOR class with respect to demographic factors, such as the

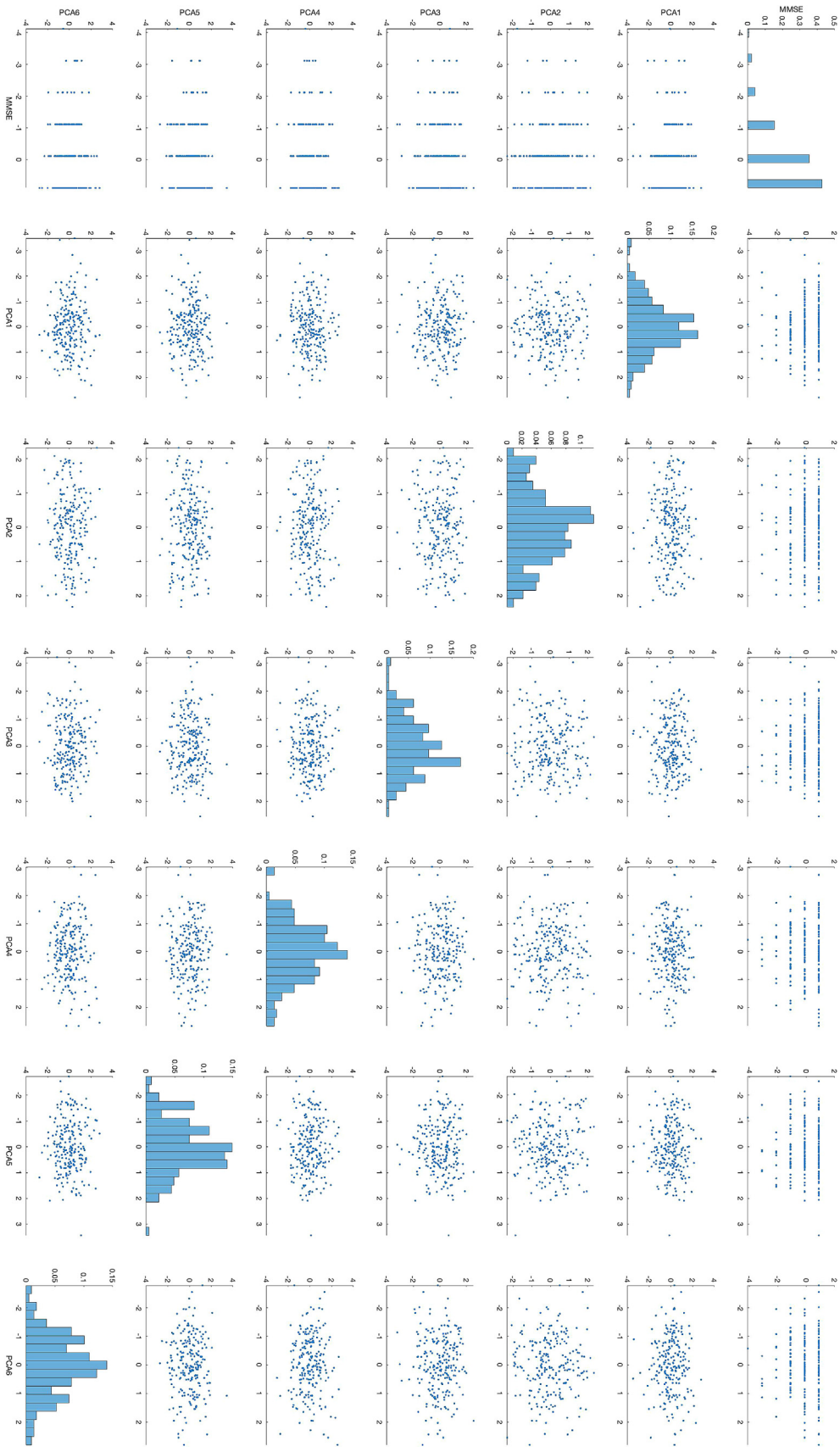
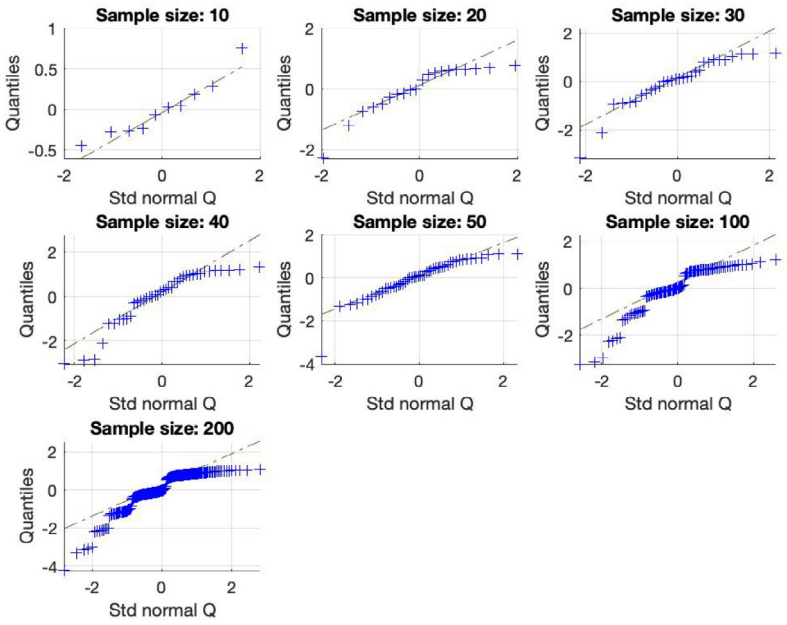
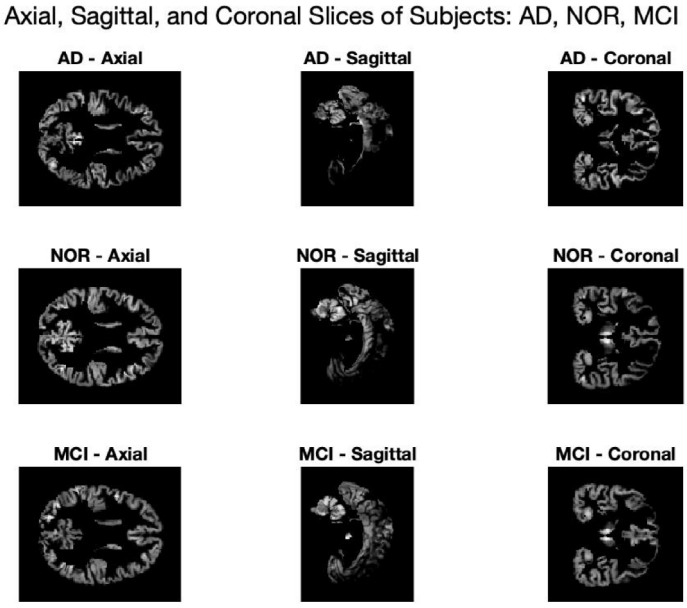


Fig. 8. ADNI Dataset with multiple predictors (6). The first column represents the MMSE as the observable variable. The remainder of columns are the predictors (PCA components) extracted from the set of segmented GM images.



(a) Q-Q plot revealing non-Gaussianity in the NC class.



(b) Three samples of the segmented GM dataset

Fig. 9. ADNI Dataset (continuation). . .In Fig. 9a we represent the Q-Q plot following the analysis given in the reference provided in the text. Assumptions needed to perform the F-test are not fulfilled as shown in the Q-Q plots.

Mini Mental State Exam (MMSE), along with Q-Q plots of the corresponding OLS residuals, revealed non-Gaussianity and non-standard data distributions in the control class (Fig. 8).

Cancer dataset in multiple dimensions

We tested our multivariate methods on real data set with data downloaded from US National Cancer Institute and the US Census American Community Survey to explore the linear relationship between between cancer mortality rate and several predictors (socio-economic status) in US counties. Seehttps://data.world/nrippner/cancer-linear-regression-model-tutorial/ for a full description of the data).

Following the OLS analysis (Fig. 10) of this dataset and eliminating multicollinearity by applying variance inflation factors, we employed up to six independent predictors including: Both male and female reported below poverty line per capita (All_Poverty_PC), median income of all ethnicities (Med_Income), males and females with and without health insurance per capita (All_-With/out_PC), lung cancer incidence rate (Incidence_Rate), and the population estimate in 2015 (Pop_estimate). A dataset with a

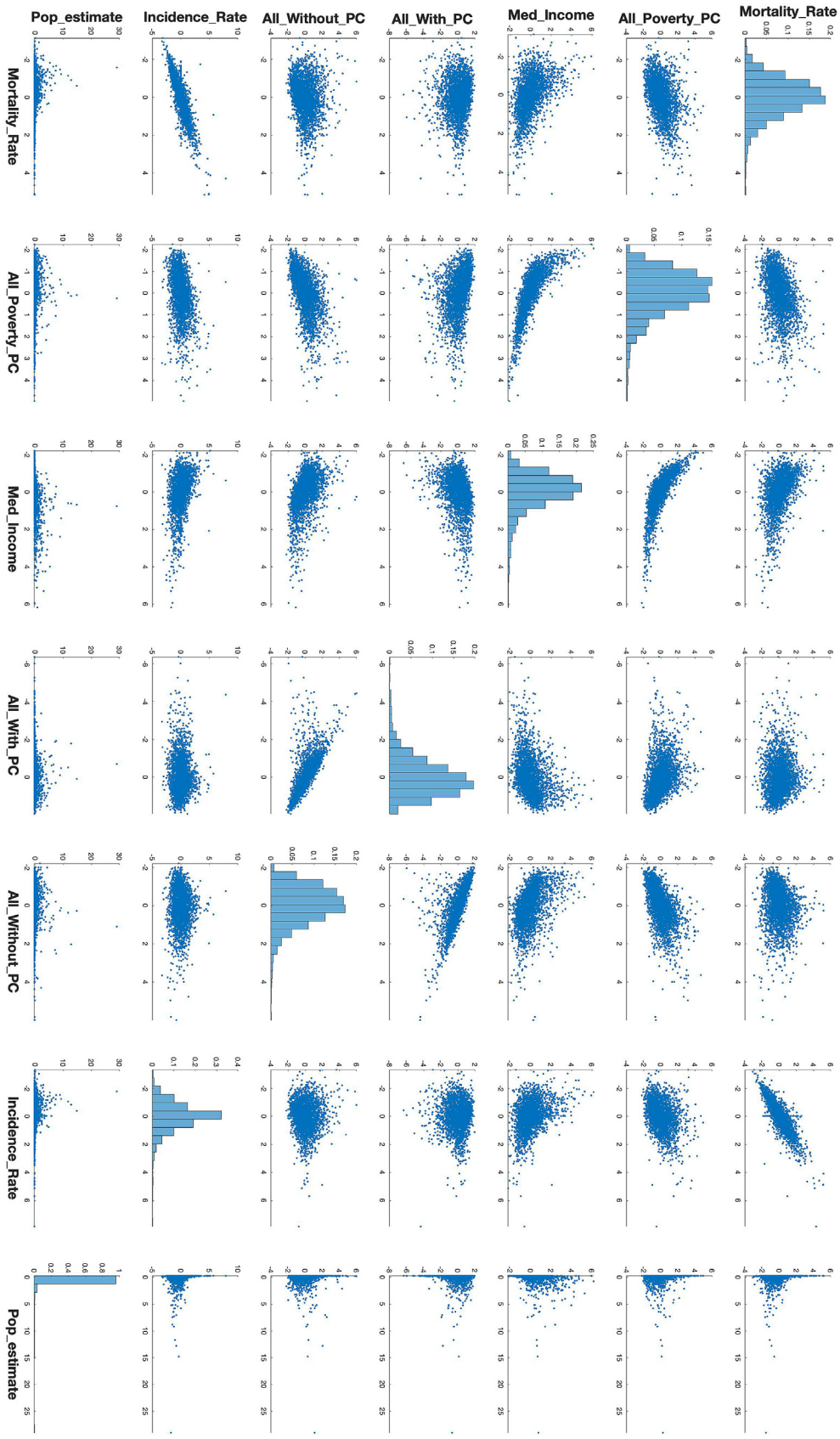
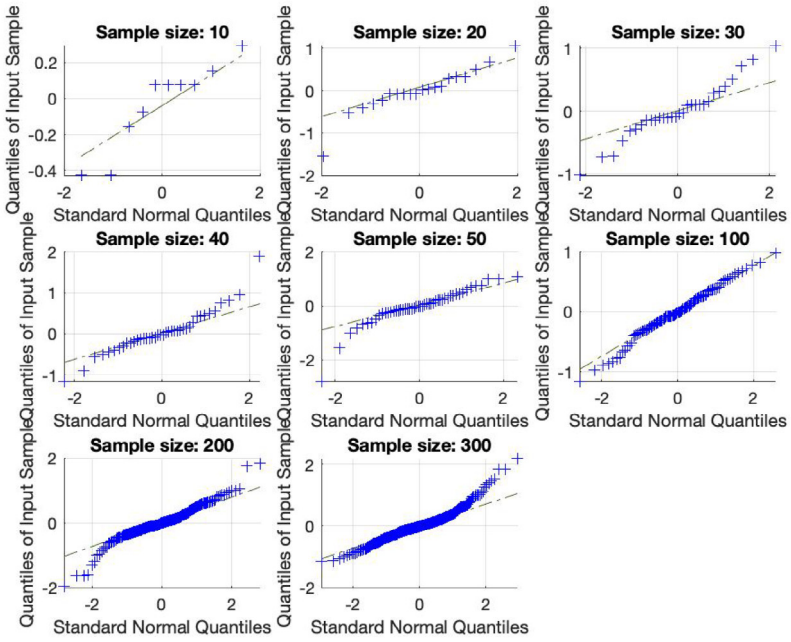
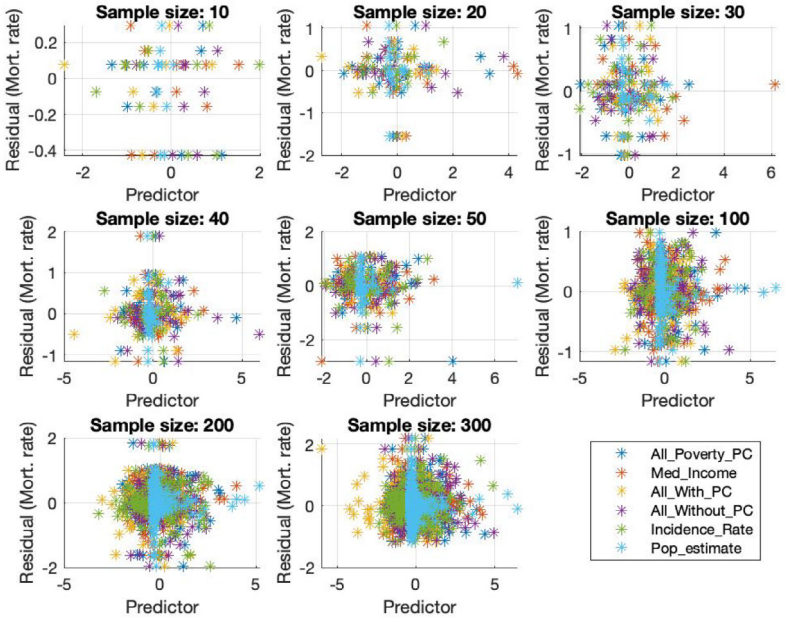


Fig. 10. Cancer Dataset with multiple predictors (6). The first column represents the mortality rate as the observable variable.



(a) Q-Q plot revealing non-Gaussianity.



(b) Explanatory variable vs. residuals.

Fig. 11. Cancer Dataset (continuation)...In Fig. 11a we represent the Q-Q plot following the analysis explained in Section 4.4.2. Assumptions needed to perform the F-test are not fulfilled as shown Figs. 11a and 11b.

sample size of $N = 2809$ was created and randomly down sampled for randomization analysis.

Results

We evaluated linear SVR models using the following validation techniques: K-fold, LOO, Resubstitution, and SAR, along with the OLS method, on synthetic, realistic, and real datasets. No hyperparameter optimization was applied to the linear ML models, therefore overfitting was not expected and nested cross-validation was unnecessary in this experimental setup. Table 3 summarizes the experimental setup used for all analyses conducted. Algorithms 1 and 2 describe the statistical inference procedure for comparing both approaches (classical vs. ML).

Table 3
Methods and parameters tested in the regression analysis.

Parameter	Value/Description
Methods	OLS, LOO, K-fold, SAR
Loss	$\mathcal{L}_1, \mathcal{L}_2$ — Specifies the loss function used for evaluating model performance.
SAR Mode	No data splitting (available for any CV method)
η	0.5 — Confidence level parameter for PAC-Bayes bounds
δ	0.5 — Dropout rate used in PAC-Bayes bounds computation
λ	$\in (1/2, 10)$ — Parameter for upper-bound optimization
K	10 — Number of folds for CV
L_{max}	Median of maximum absolute errors beyond the ϵ margin
R	100 — # of iterations used to assess model performance variability
α	0.05 — Significance level.

Algorithm 1. Statistical Inference with Linear Models: Classical vs. Machine Learning with Upper Bounding

1: **Set parameters:**
2: Set those defined in Table 3.
3: **Load Data and Normalization:**
4: Load **X** and *Y* data matrices (e.g., MRI data and MMSE scores for MCI, NOR, AD)
5: Extract relevant features (e.g., PCA)
6: Normalize data (e.g. z-score)
7: **for** each P + 1 dimension **do** ▷Loop over dimensions:

(*continued on next page*)

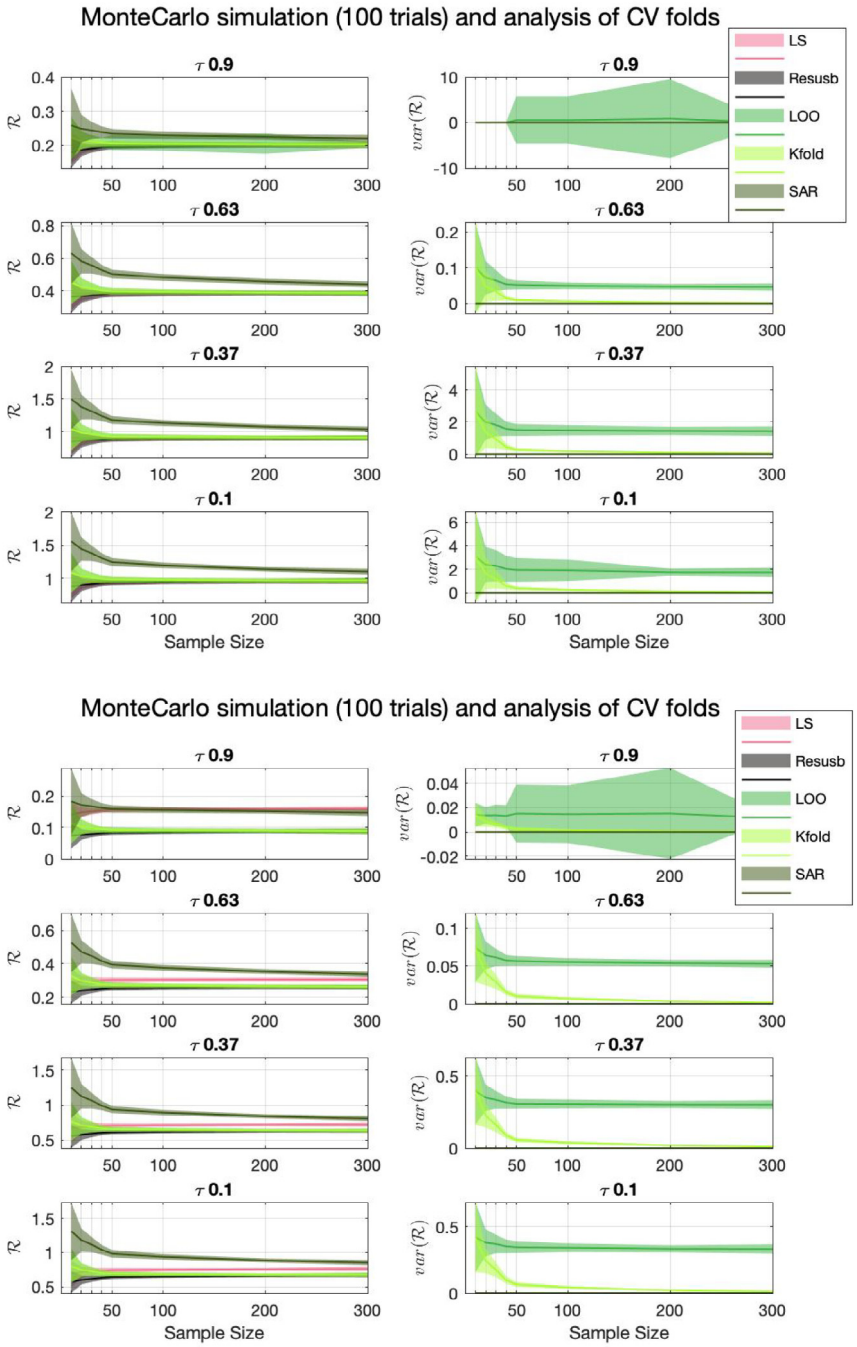


Fig. 12. Expected losses ($\mathcal{L}_2$ above and $\mathcal{L}_1$ below) and analysis of variance across folds. Note that the variance from folds reflects the variability encountered in real scenarios with a single sample realization.

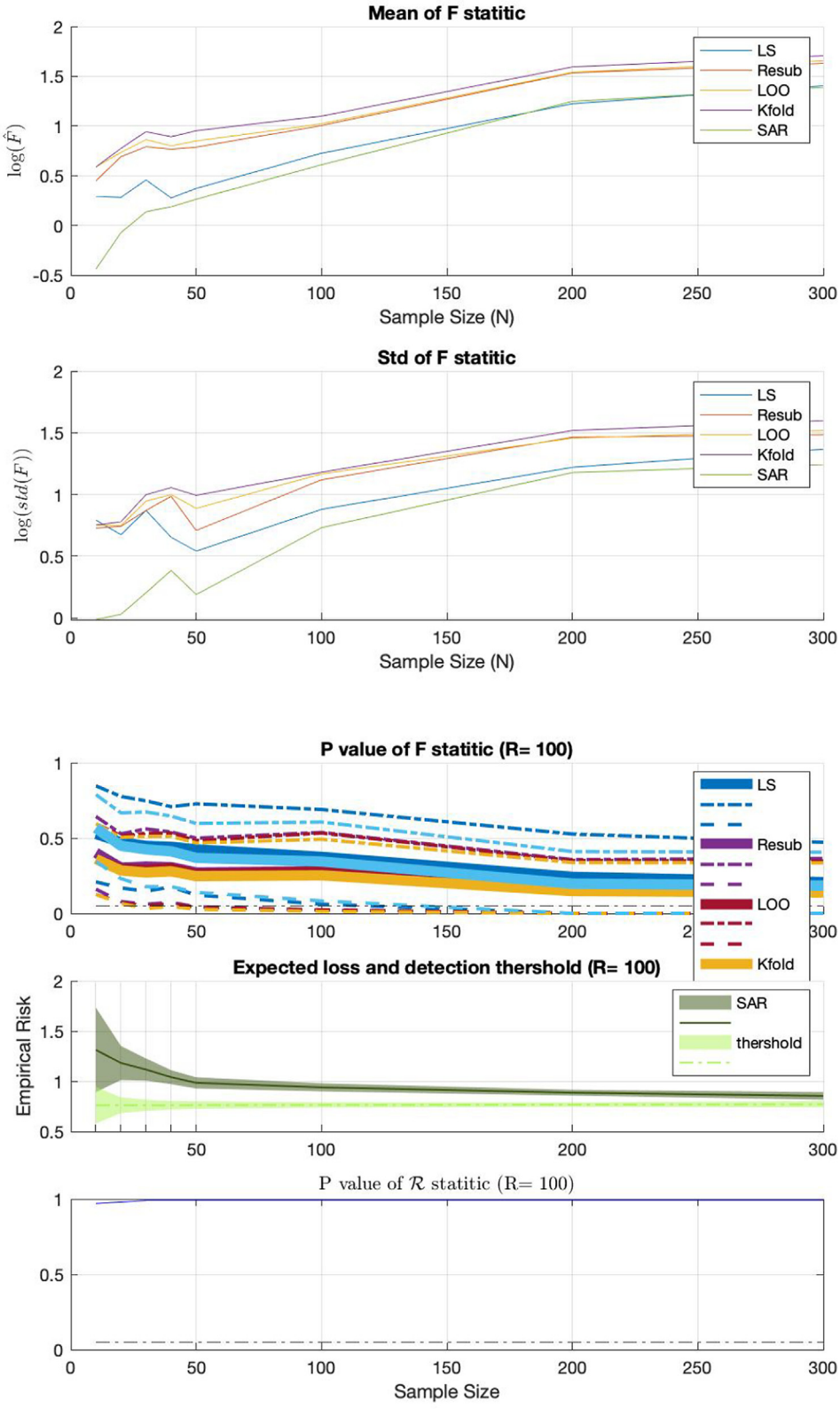


Fig. 13. The F-test for the slope β_1 in the case of a correlation level equal to 0.1. We utilized the residuals derived from each method and averaged the F-statistic and its variability over $R = 100$ repetitions. It is noteworthy that SAR provides a trade-off between ML methods and OLS, exhibiting less variability in the repetitions. The SAR test is obtained with a probability of at least $\eta = 0.5$.

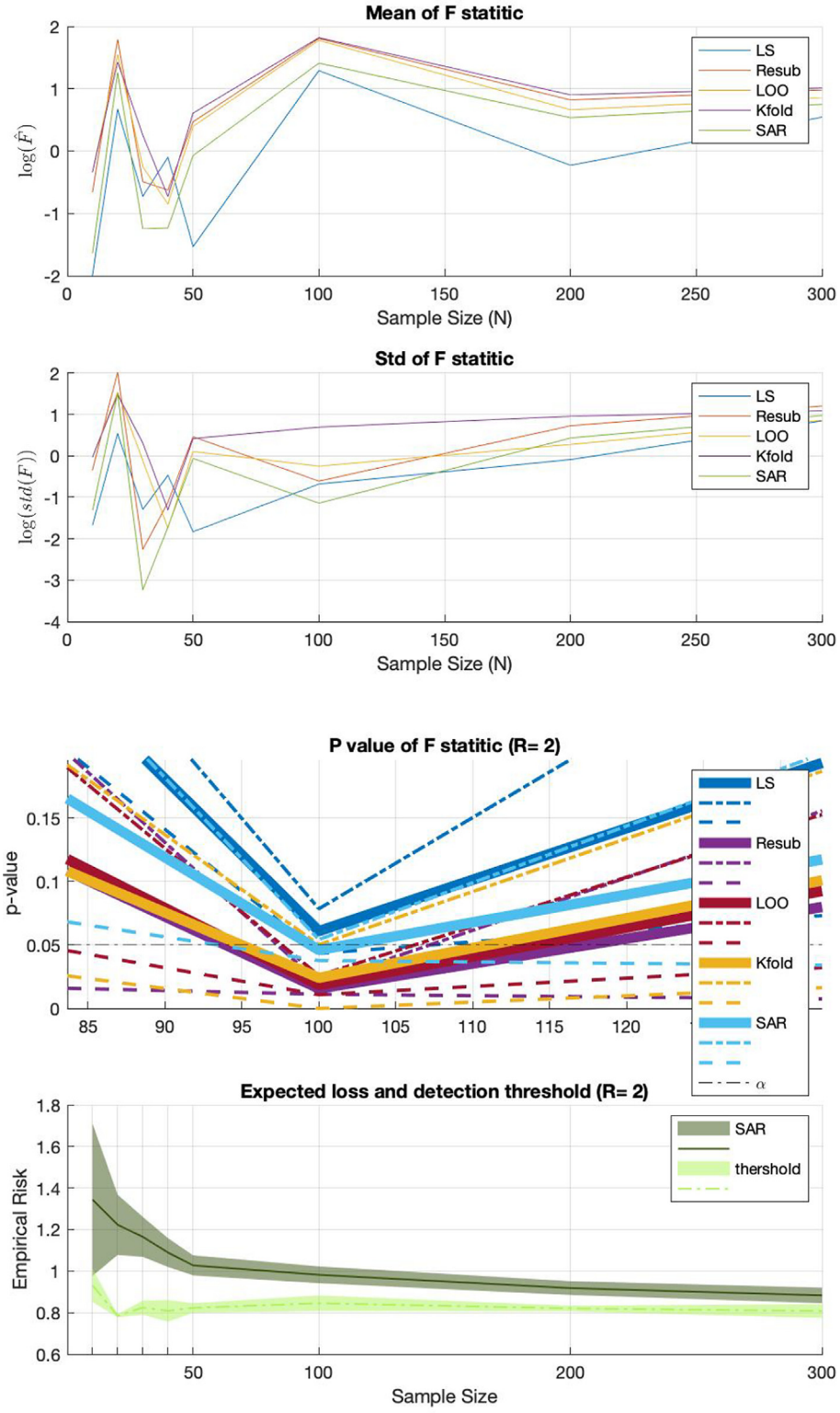


Fig. 14. The F-test for the slope β_1 in the case of a correlation level equal to 0.1. We utilized the residuals derived from each method and averaged the F-statistic and its variability over $R = 2$ repetitions.

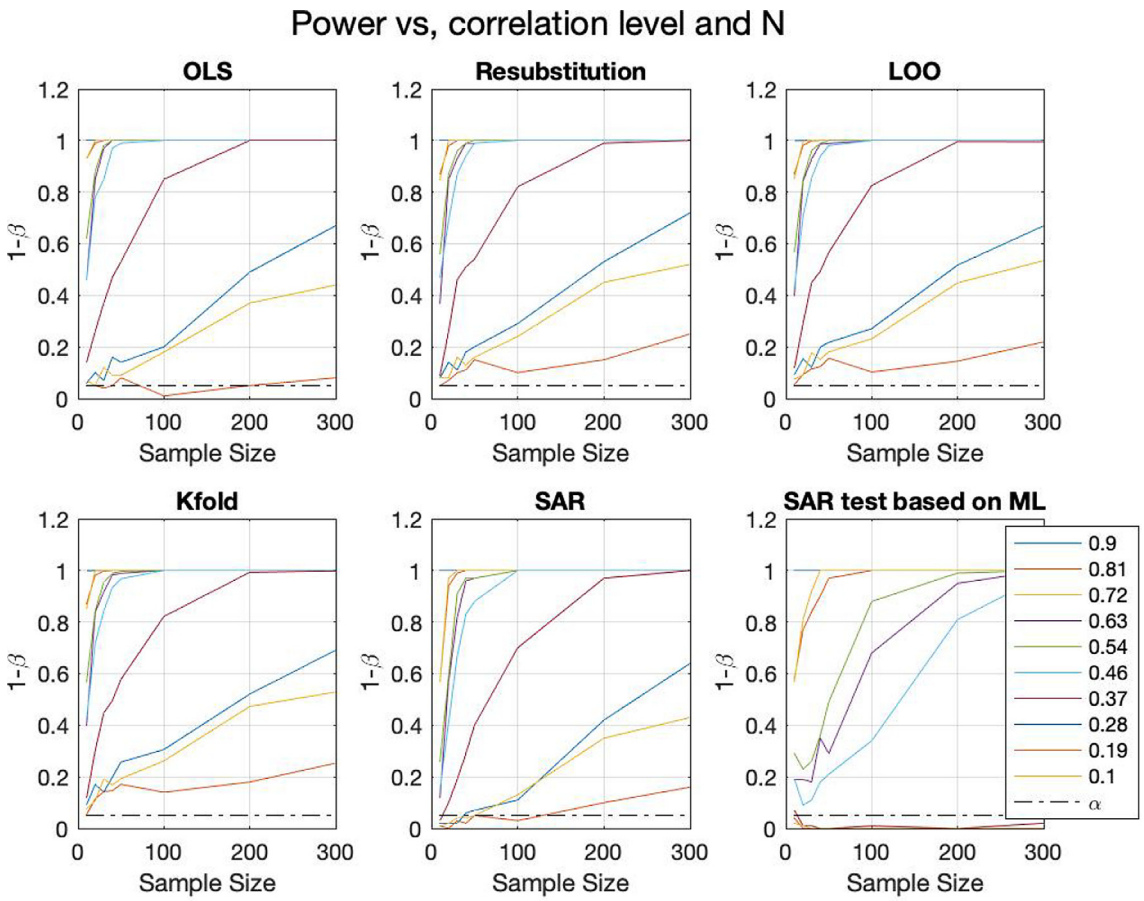


Fig. 15. Power analysis derived from the F-test in Gaussian data. The selected loss for ML methods is $\mathcal{L}_1$. Observe how the results obtained by the SAR features are by far the most similar to those obtained by the OLS. The SAR test based on expected loss is more conservative than the other methods (correlation levels below 0.37 are not considered significant based on the worst-case analysis with a probability of at least $\eta = 0.5$).

Define the dataset:
  Extract features $\mathbf{X}_T$ and target values Y_T
for each N sample size **do** ▷Loop over sample sizes
  for each R iteration **do** ▷Loop over repetitions
    Perform resampling of data $\mathbf{X_{Ts}}$ and Y_{Ts}
    Shuffle Y_{Ts} or process data as required
      ▷Perform regression and validation
    Linear Regression using OLS at α
      Compute regression coefficients β, and residuals
      Compute F-stats (parametric p-value from Eq. 10)
    Cross Validation with ML (K-fold, LOO)
      Perform CV to compute risk $\mathcal{R}_N$ and F-stats
    SAR regression at η
      Compute risk $\{\mathcal{R}, \mathcal{R}_u\}$ with **CUB** (Eq. 14)
      Compute F/SAR-stats (Eq. 12)
  end for
end for
Randomization analysis ▷Statistical Analysis
  Compute permutation p-value (F and SAR tests) in Eq. 13.
end for

Algorithm 2. Computation of PAC-Bayesian bound

Subfunction: CUB
  Input: $\mathcal{R}_N, \eta, \delta, L_{max}, \beta$
  Output: Δ
  Set parameters:
  Define $\lambda \in \left[\frac{1}{2}, 0.1, 10\right]$
  Compute $a = \frac{1}{1-\frac{1}{2\lambda}}$
  $\Theta = [\beta_1; \beta_0]$
  For each λ **in range:**
    $\Delta = \min_\lambda\left((a-1)(1-\mathcal{R}_N) + a \cdot \left(\frac{\lambda L_{\max}}{N}\right)\right.$
    $\left.\cdot\left(\frac{(1-\delta)}{2}\|\Theta\|^2 + \log\left(\frac{k}{\alpha}\right)\right)\right)$

Gaussian data: OLS is the gold standard

In this section, linearly transformed Gaussian data is considered, where the OLS method is the gold standard. We averaged $R = 100$ realizations and used the same loss ($\mathcal{L}_2$) with standard ML algorithms, which performed quite well (SAR was conservative

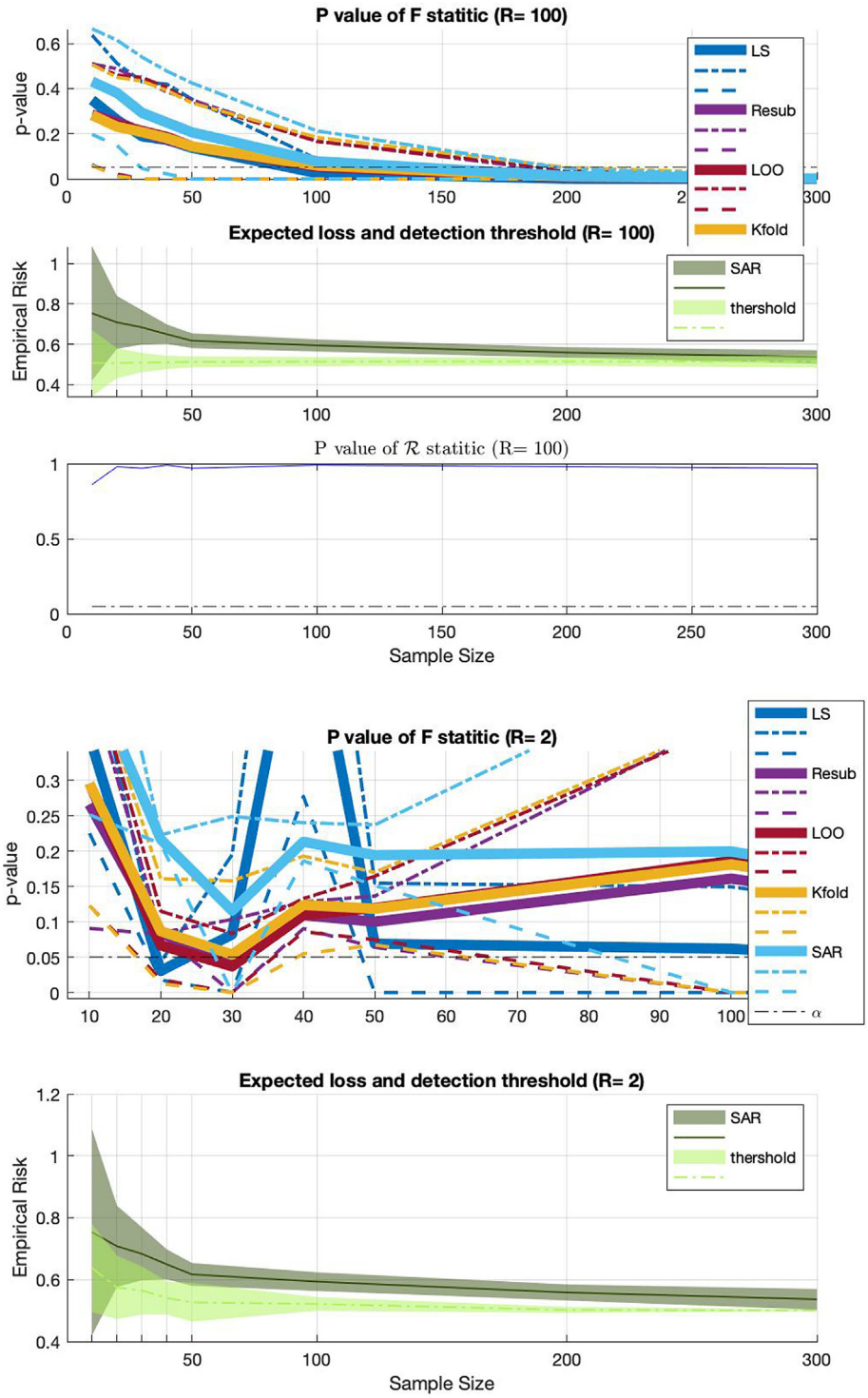


Fig. 16. The F-test for the slope β_1 in the case of correlation levels equal to 0.37 with non-Gaussian data. We employed the residuals derived from each method and averaged the F-statistic and its variability in $R = 100$ and 2 repetitions for comparison purposes.

but converged to OLS as the sample size increased). However, there was high variability in the expected loss from folds when the number of samples was limited ($N < 50$). It is worth noting that when using the $\mathcal{L}_1$ loss, the typical loss in ML applications, the correction applied to the resubstitution error made SAR converge to OLS, unlike the other ML validation methods (see Fig. 12). The methods are now tested with the formal F-test for the slope to assess their ability to detect a slight linear relationship with a correlation level equal to 0.1. The results reveal optimistic behavior of the ML methods, except for the SAR method which struck a trade-off between OLS and ML methods. It is essential to recall that OLS is the gold standard in this case, and the conservative behavior of the SAR allows us to obtain outputs similar to ML. In Fig. 13, we display the p-value in this challenging case where the null hypothesis is not rejected. At the bottom of the figure, we provide the SAR test based on the detection threshold that mimics classical tests for

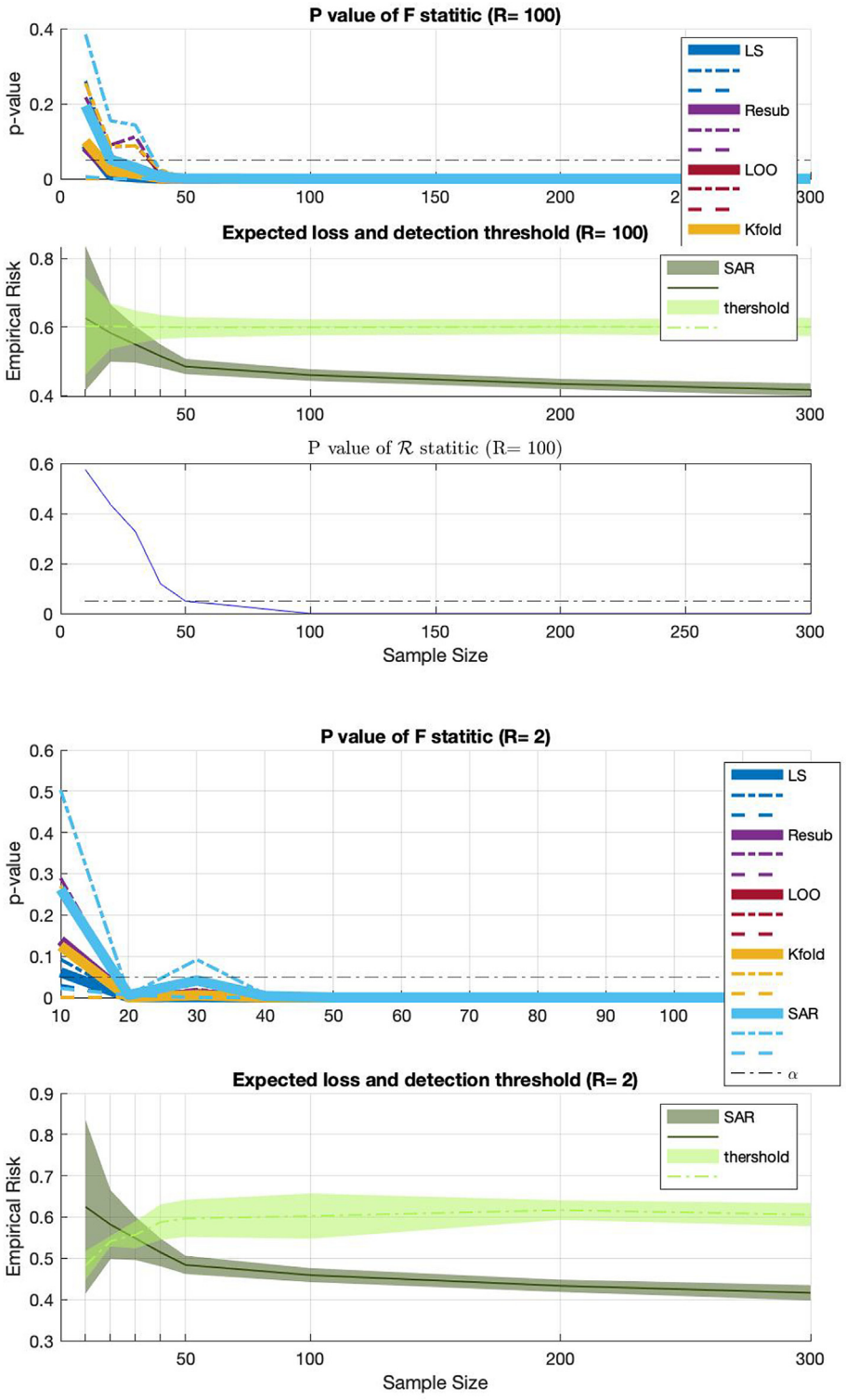


Fig. 17. The F-test for the slope β_1 in the case of correlation levels equal to 0.63 with non-Gaussian data. We employed the residuals derived from each method and averaged the F-statistic and its variability in $R = 100$ and 2 repetitions for comparison purposes.

the slope based on a significance level. The optimistic results obtained by CV methods with a limited sample size, which could indicate their better ability for linearity detection, are indeed a consequence of poor control of FPs. To demonstrate this issue, we designed a putative task consisting of a regression problem with no correlation at all, i.e., a correlation level equal to 0. In this

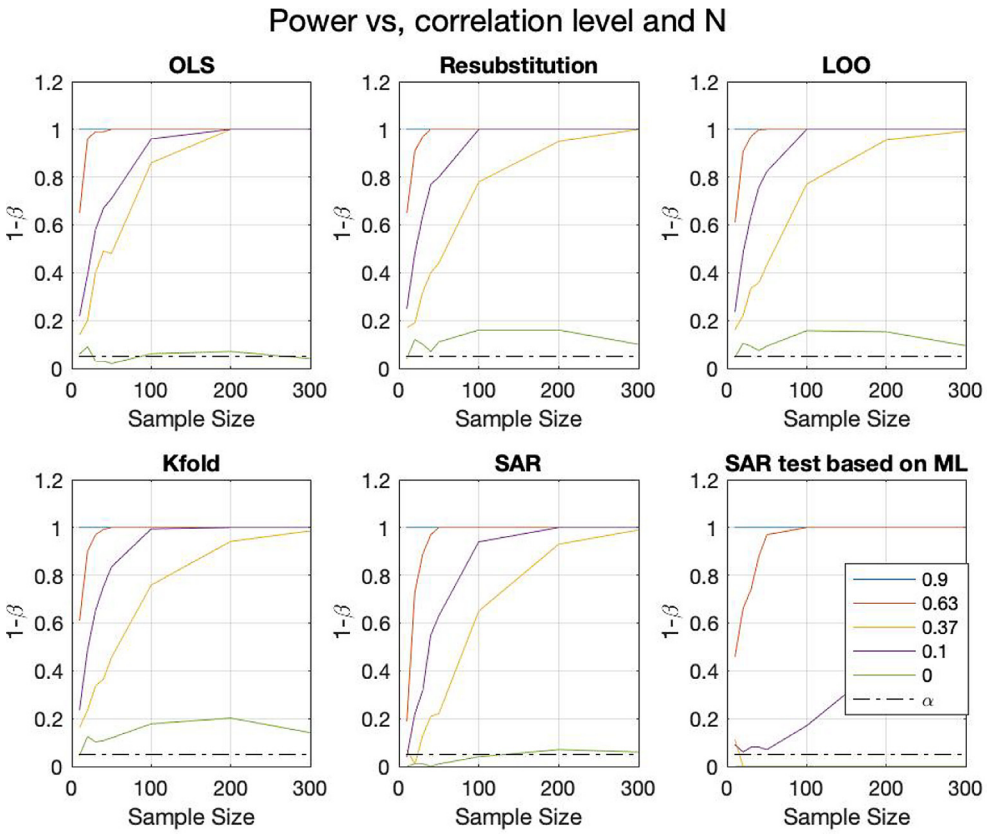


Fig. 18. Power analysis derived from the F-test in non-Gaussian data (except for the green solid line that represents uncorrelated data). The selected loss for ML methods is $\mathcal{L}_2$. Observe how the results obtained by the SAR features are by far the most similar to those obtained by the OLS for low correlation levels. However, for high levels, the performance of SAR is a trade-off between ML methods and OLS. The SAR test based on expected loss is more conservative than the other methods (correlation levels below 0.37 are not considered significant based on the worst-case analysis with a probability of at least $\eta = 0.5$).

case, we analyze the real case with almost no repetitions ($R = 2$), where the variability of errors in folds affects the evaluation of the F-statistic. In this scenario, we observe a non-smooth behavior of the F-statistic curves, but only ML CV lines cross below the significance level (e.g., $N = 100$), unlike the SAR test or the F-test for the OLS estimation, which are always above the detection threshold, controlling the rate of FPs (see Fig. 14). Moreover, to illustrate this issue, we performed a power analysis in Fig. 15 with a wide variety of correlation levels, showing that CV methods are inflating the rate of FPs (detection rates improved compared to the OLS method, for low correlation levels).

Non-linearly transformed Gaussian data: large Cross-Validation residuals and False Positives

In this case, the OLS method may not be the most efficient estimator, and inferences based on the normality assumption, such as the F-test, can lead to unreliable results. Here we find the true utility of the SAR test: to validate the significance of a prior regression approach. We analyzed two cases, as shown in Fig. 5, with correlation levels $\tau = 0.63$ and $\tau = 0.37$. Both exhibited similar p-values for the standard ML approaches and OLS. However, this is not always good news, as the Gaussianity assumption is not fulfilled. Particularly, for $R = 100$ ideal realizations, from approximately $N = 120$ and $\tau = 0.37$, all methods rejected the null hypothesis, as shown in Fig. 16. One might be tempted to conclude that there is sufficient evidence for a linear relationship in the data. Unlike the SAR test, where the values of the expected loss converge to the detection threshold but always remain above it (p-value greater than the significance level). In realistic scenarios with $R = 2$, that is, the typical hypothesis test using a classical approach, all methods provided a p-value above 0.05 across a wide range of sample sizes, contravening the previous results. Thus, for example, results from a single laboratory with only two samples cannot be extrapolated to the rest of the realizations. In other words, depending on the repetition, we may have a significant result, or the null hypothesis cannot be rejected. As shown in the same figure, the SAR test, being conservative, is robust in this scenario compared to the F-test.

Another example is shown in Fig. 17 with $\tau = 0.63$ (indicating significant linearity), where the SAR test validates the significance of the results from approximately $N = 50$. Note that the SAR test is robust against the number of repetitions in the bootstrap approach, since for $R = 100$ and $R = 2$ repetitions, we reject the null hypothesis at the same sample size, and above $N = 50$, leading to good replication. Finally, we performed a power analysis, including a putative task with a correlation level equal to zero to assess how the methods control the rate of FPs [39]. In this case, the rate of FPs should be approximately equal to the level of significance α. Fig. 18 shows that the ML methods are inflating FPs (green solid line) above the level of significance, i.e., $(1 - \beta) > \alpha$. Recall that the results of OLS under conditions of non-Gaussianity should be subject to scrutiny (except for the case $\tau = 0$).

Heteroscedascity data: superior detection using ML techniques

In this section, we tested the ability of the methods to detect heteroscedasticity using the BP test described in Section 4. The

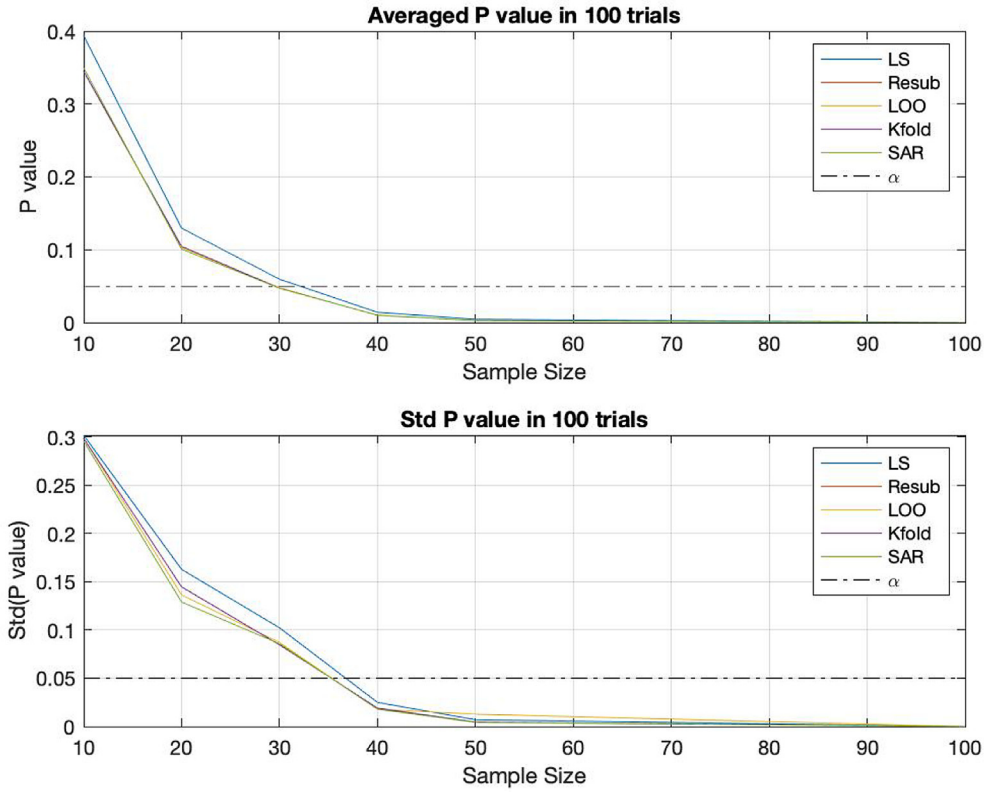


Fig. 19. The BP test on the residuals.

problem analyzed here is clearly linear, but the distribution of residuals is not homogeneous (see Fig. 6). A simple Q-Q plot analysis of the data reveals its strong heteroscedasticity.

We evaluated the BP test on the residuals obtained with all methods and compared the averaged and standard deviation (std) p-value with increasing sample size in $R = 100$ repetitions. The BP test using the SAR residuals provides a faster detection of heteroscedasticity on average and a lower standard deviation of the p-value, as shown in Fig. 19.

Multivariate linear regression: SAR features effectively detect linearity.

The cancer dataset

We repeated the test for linearity and the power analysis with increasing dimensions, from one to six predictors, using the Cancer dataset. We observed two primary effects. First, the variance of CV methods is extremely large for small sample sizes ($N = 10 - 50$), especially with the $\mathcal{L}_2$ loss, although it is significant for the rest of the simulations. This effect depends on the predictor added in the regression analysis. Notably, the last predictor that was added, 'Pop_ estimate' (predictor 6), increases the variance of the estimators resulting in non-reliable estimates within training folds, mainly with small sample sizes (Fig. 20). The reason for this anomalous operation was previously shown in Fig. 11b in light blue. Here, we readily see in the SAR test (bottom of the figure) that evidence for a linear relationship in the data, under non-ideal conditions, is achieved with only 10 samples (the risk is always less than the threshold).

The second effect worth commenting on is the detection ability of the SAR method when the number of predictors increases, as shown in Fig. 21a. Moreover, the SAR method converges to the OLS method with the two tested losses, unlike the CV-based ML methods. This suspicious feature found for the SAR test (and for any other ML method) can be explained by overfitting since it goes against its common behavior. To discard this possibility, we increased the dimensions or the number of predictors used in the simulation representing the putative task where there is no correlation. In this task, we additionally employed a general upper bound based on the assumption of samples in general position (i.g.p.), as proposed in [31]. The analysis for the two bounds is given in Figs. 22a and 22b. As shown in this analysis, when we increase the number of dimensions (only in this section), the bound proposed in [34] is overly optimistic compared to that proposed in [31] when $N < 100$. In this case, PAC approaches should be reformulated, and analytical bounds such as those proposed in [30,31] are valid solutions to formulate the SAR test as they effectively control the FP rate.

The ADNI dataset

The correlation between MMSE scores and MRI images is important because it provides valuable insights into the relationship between cognitive function and structural brain changes. We can clearly observe this effect in the set of PCA features extracted from the original dataset, prior to the regression analysis shown in Fig. 23a. In the same figure, we present the regression fit using both OLS and SAR approaches (including the ϵ-tube), where the observed MMSE scores and predictors are normalized using the z-score method. Note how the AD features are primarily located below the hyperplanes, while the NC features are above them.

We first analyzed this relationship in the NC group to assess the "quality" of the baseline group. The correlation between MMSE scores and MRI control images might not be as strong or consistent as in groups with MCI or AD. In healthy individuals, both MMSE

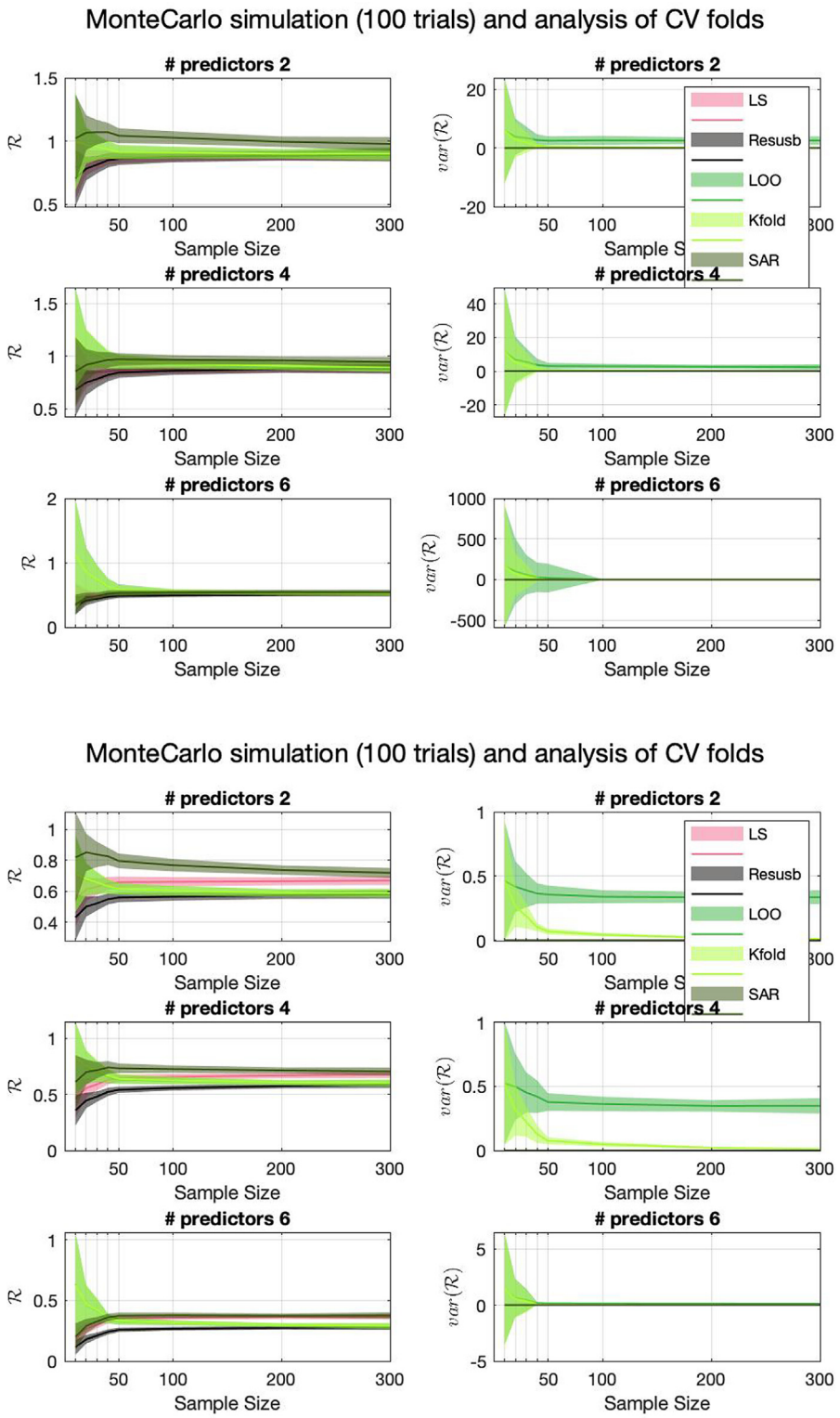
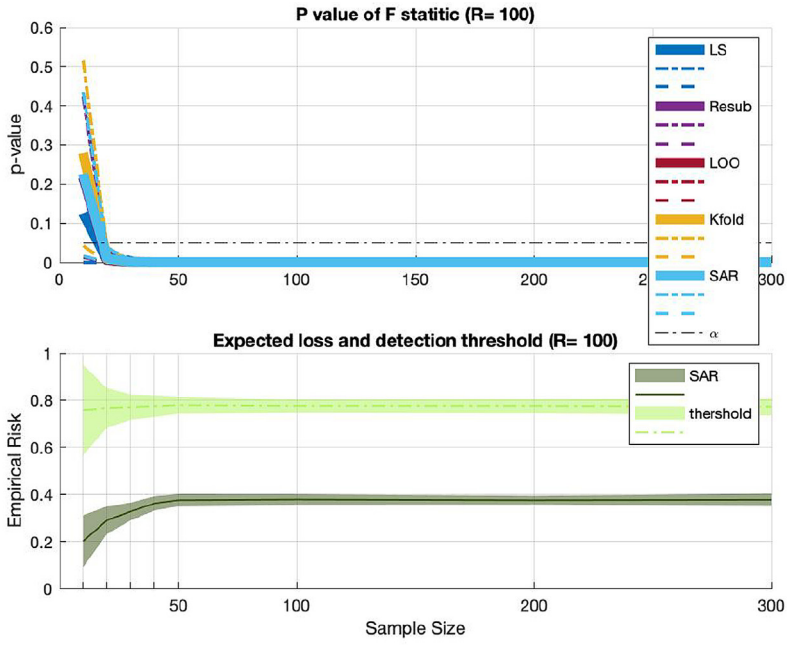


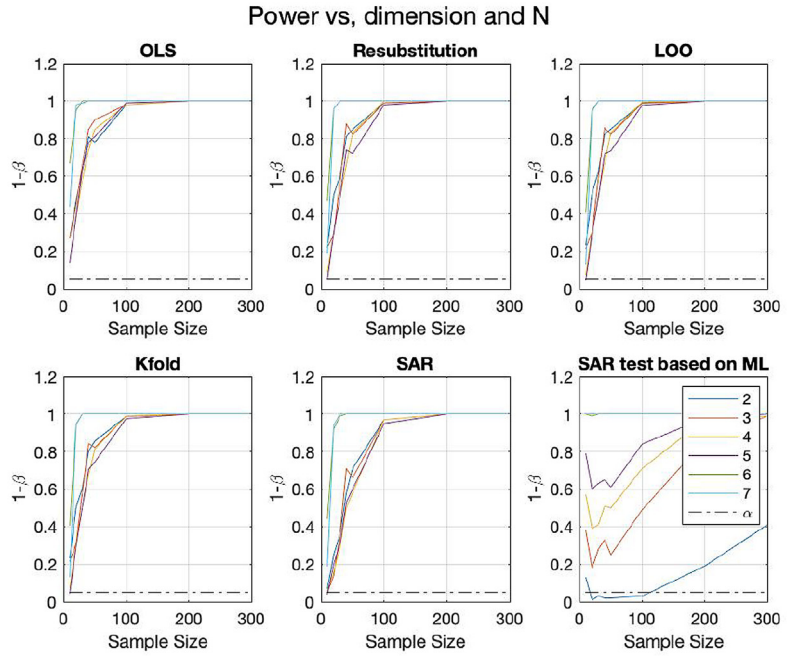
Fig. 20. Expected losses ($\mathcal{L}_2$ above and $\mathcal{L}_1$ below) and analysis of variance across folds. Once again, the variance using the $\mathcal{L}_2$ loss is greater than that of the $\mathcal{L}_1$ loss for ML methods employing cross-validation. However, with real data, the presence of large outliers in folds can lead to inadmissible loss estimations.

scores and brain structures are likely to be within normal ranges, with less variability in cognitive function and brain structure. We detected suspicious behaviour in the power of the F-test of the methods based on ML. They significantly correlated MMSE with the PCA features above the significance level (Fig. 24b). This effect is essential for comparing against groups with neurodegenerative conditions like MCI or AD. If the control group shows unexpected correlations, it raises concerns about the validity of the baseline for future comparisons.

However, a p-value analysis over a set of R repetitions showed that no significant correlation can be found between MMSE and the brain structure as expected (Fig. 24a). Despite the ML models indicating a significant correlation, a p-value analysis conducted over multiple repetitions (R) contradicts this finding, showing that no significant correlation actually exists between MMSE and brain



(a) Testing the linear relationship in Cancer dataset with 6 predictors using $\mathcal{L}_1$.



(b) Power analysis versus the # predictors and sample size. Note that results are obtained with the most competitive loss for ML methods ($\mathcal{L}_1$).

Fig. 21. Performance with increasing number of predictors in Cancer Dataset.

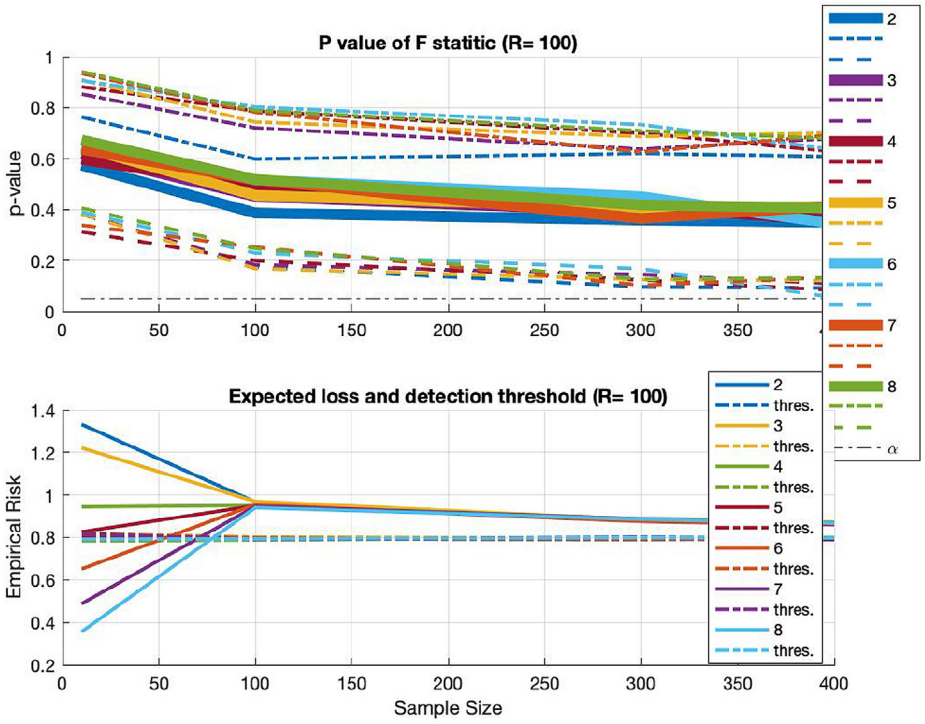
structure in the control group. This discrepancy highlights potential overfitting or spurious correlations within the ML methods when applied to group analyses in realistic experimental setups.

Finally, if we consider the three classes (AD, NC, and MCI) and an increasing sample size $N = 10$ to 800, we detect linearity with a few samples using the F-test, and above $N = 200$ with the SAR test as shown in Fig. 25a. Notice how the SAR test penalizes increasing dimensions, in contrast to classical residual-based tests for linearity (Fig. 25b).

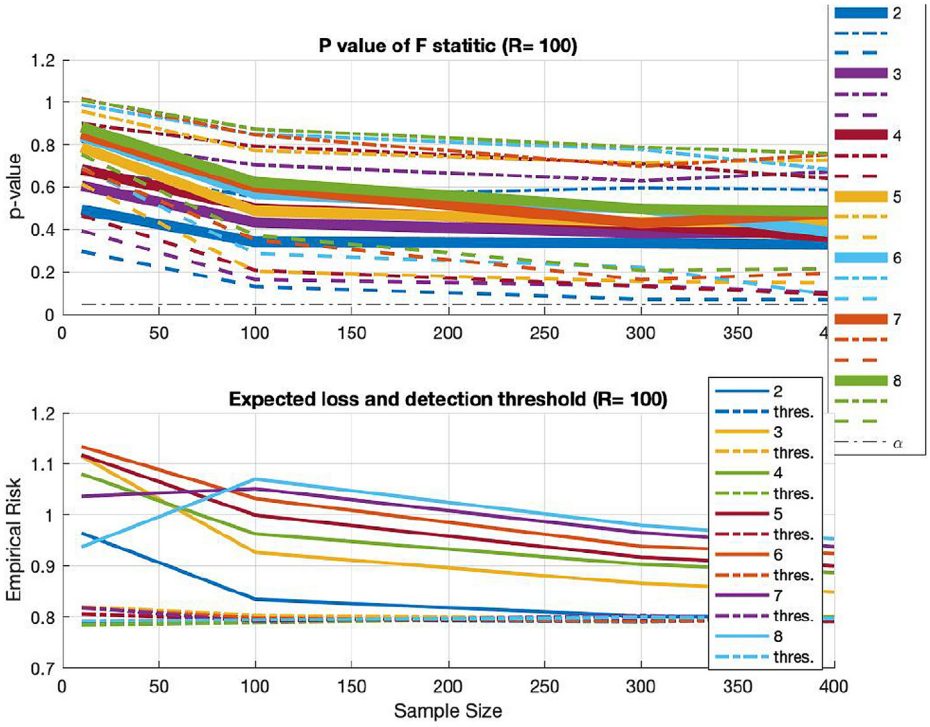
Discussion

OLS, machine learning, and statistical learning theory

Initially, we recapitulated the conventional application of OLS in linear regression, emphasizing its widespread adoption due to its optimal statistical properties. The transition to ML introduces regularization highlighting the shift in focus towards minimizing both empirical risk and model complexity. Critically, we highlighted a



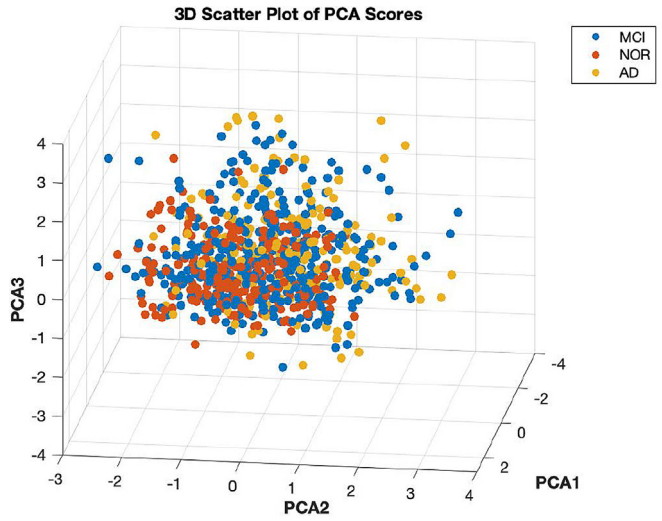
(a) Test for linearity in the putative task of uncorrelated samples using PAC bound.



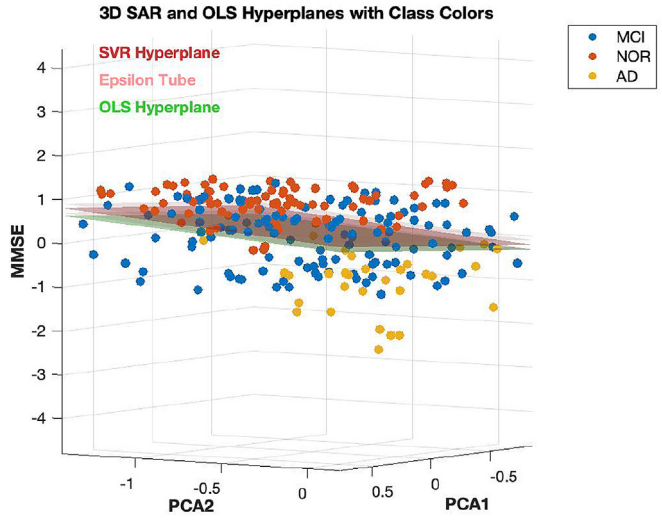
(b) Test for linearity in the putative task of uncorrelated samples using the i.g.p. bound.

Fig. 22. Performance with increasing number of predictors in Cancer Dataset (continuation)....

common deficiency in ML approaches, particularly in regression, where rigorous significance analysis is often neglected in favor of permutation testing based on cross-validation measures, especially with small sample sizes. The proposed SAR method seeks to bridge this gap by formulating a statistical test grounded in statistical learning theory. This involves establishing an upper bound on the



(a) 3D PCA clustering of the original ADNI dataset



(b) 3D regression hyperplanes for MMSE vs predictors PCA 1-2.

Fig. 23. PCA features in ADNI dataset (NC, AD and MCI subjects).

expected loss comparing it to the null hypothesis and rejecting it if the corrected risk is lower, signifying evidence of a linear relationship. The background of Support Vector regression was then explored presenting the ϵ-insensitive loss function and the regularized risk functional. The discussion explores the flatness property and empirical loss highlighting the differences in loss functions employed in ML models. Finally, the novel SAR method was introduced as a formal test for assessing the error estimation of ML algorithms drawing parallels with classical statistical tests, like the F-test. This section emphasized the significance of SAR in overcoming limitations in ML approaches and extended the application to regression analysis.

Experimental validation and significance analysis

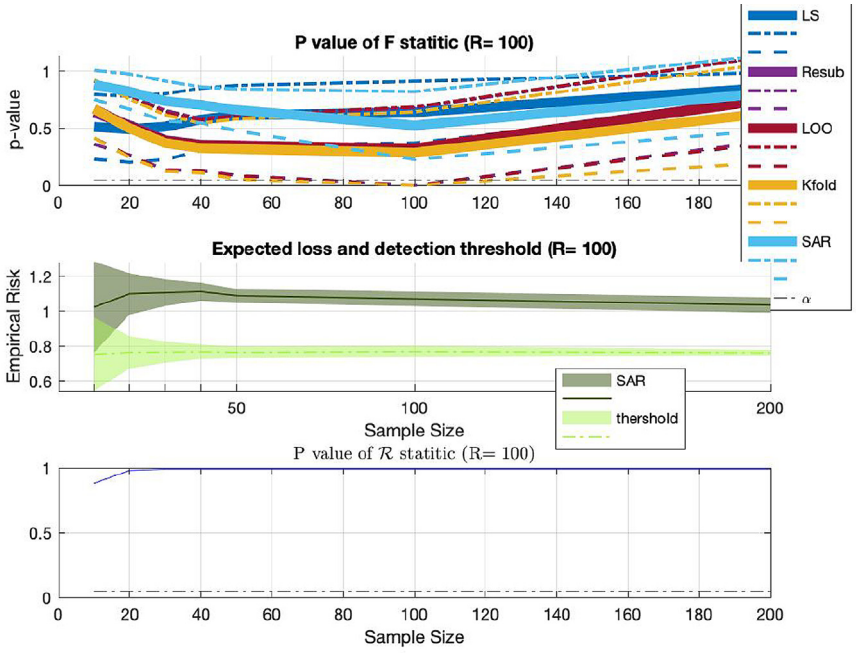
The results section presented a comprehensive examination of experimental outcomes focusing on the performance of the proposed SAR method in comparison to standard ML techniques across diverse scenarios. For Gaussian distributed data, where OLS serves as the gold standard, the SAR method exhibited conservatism converging to OLS with increasing sample sizes [41]. Testing with the F-test for the slope at a correlation level of 0.1 revealed the optimistic behavior of ML methods, except for SAR, which struck a trade-off with OLS [42]. Analysis of non-Gaussian distributed data emphasized the utility of SAR in validating significance (reproducibility), especially in scenarios with limited experimental repetitions ($R = 2$) [40]. The detection of heteroscedasticity using the BP test demonstrated SAR's faster detection and lower standard deviation of the p-values. This thorough analysis, illustrated through various figures, underscores SAR's versatility and robustness in assessing error variability, validating significance, and detecting heteroscedasticity across different data scenarios (sample sizes). To sum up, from the set of experiments undertaken here, we demonstrated that the proposed SAR is effective for assessing the variability of error values obtained by CV in ML methods. Moreover, the evidence derived from these experiments suggests that SAR is a robust approach to functional regression.

Challenges in CV and theoretical implications

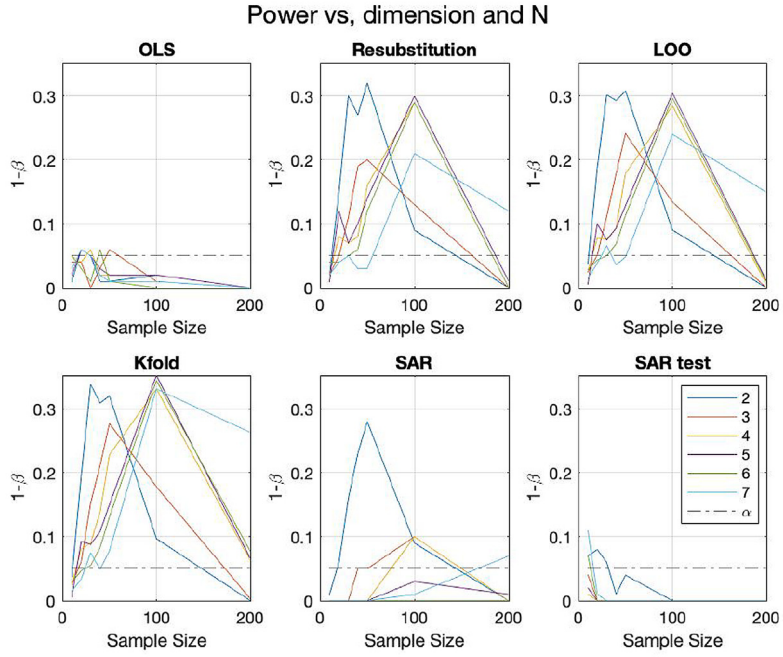
As mentioned in the introduction, there is increasing concern about the use of K-fold CV as a baseline method for model selection. AI is proposing more complex algorithmic architectures that are validated with the easy to implement CV method, potentially leading to significant variability if sample sizes are limited or the data heterogeneity is large. This raises the question of whether the giant has feet of clay. When the data allows us to derive a stable inducer by splitting (making a guess from some parts of it), it can be a very robust method to derive models beyond the current realization. However, this assumption is not always met. By performing this procedure, we can expect that what we learn in the training folds is not necessarily useful for prediction with the remaining samples. This is analogous to the judgment of Solomon; we ask the two women to determine who the baby belongs to, and the information provided by each party should be useful. In case the two women reply the same, Solomon would have to sadly split the baby in two. In the biblical story, the reaction of each of the two women to the suggestion of splitting the baby was not the same, and Solomon gave the baby to the real mother. This is akin to what we have done with data using SLT, although we correct this resubstitution (split in two) by considering the information contained in the data (reaction from data). In our case, we would give the two parts to the more affected woman.

Computational efficiency, limitations and future directions

Another key advantage of the proposed method is its relatively low computational cost. While PAC-Bayesian methods generally require more computation than other upper-bound approaches, such as those based on Rademacher averages (though less than those relying on growth functions), our use of linear classifiers in the proposed bound helps mitigate this burden. Specifically, the computational complexity is approximately $\mathcal{O}(N + j + s)$, where j is the number of iterations in the 1-D λ optimization and s is the number of samples used to estimate D_{KL}. In contrast, dataset-splitting methods like K-fold cross-validation involve multiple operations of order $\mathcal{O}(N)$ for linear models, while nested approaches further increase this complexity to $\mathcal{O}(K^2N)$. Although this difference may be negligible for small datasets, the computational cost can become substantial when statistical inferences involve permutation analyses with a large number of randomizations and complex null distribution modeling. The SAR method is based on a specific PAC-Bayesian upper bound when applied to a class of linear classifiers [29]. While the methodology is general,



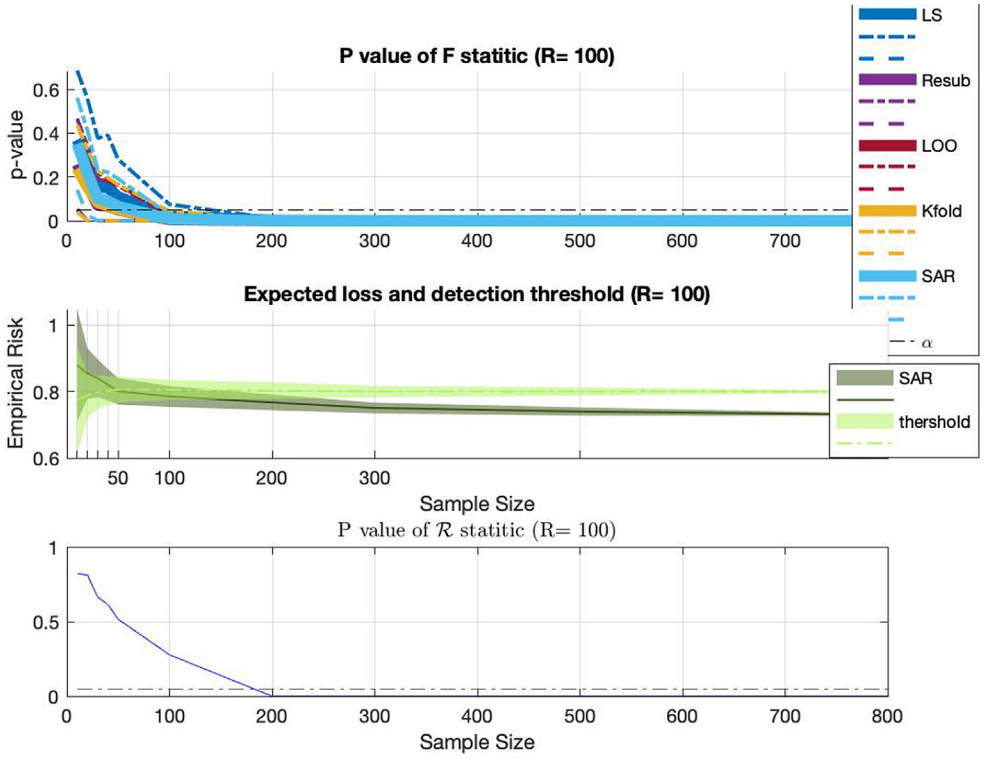
(a) Testing the linear relationship in the NC group ADNI dataset with 6 predictors using $\mathcal{L}_1$.



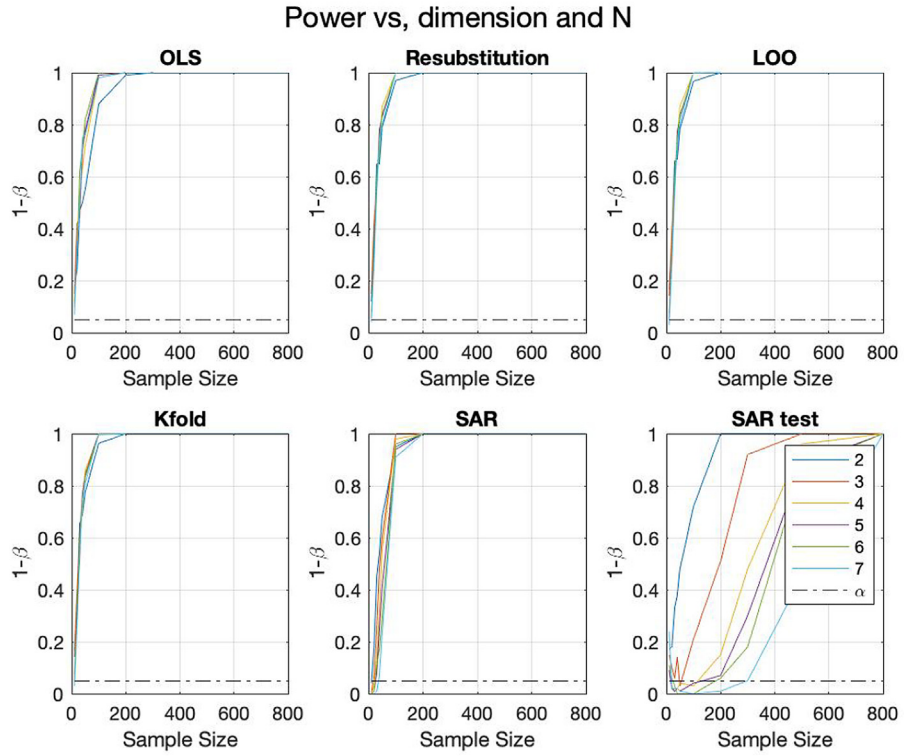
(b) Power analysis versus the number of predictors and sample size. Note that results are obtained with the most competitive loss for ML methods ($\mathcal{L}_1$).

Fig. 24. NC and NC-AD-MCI group analyses using the ADNI dataset.

extending it to more complex function classes would require defining new upper bounds which might be less strict than those proposed in this work. This limitation primarily motivates our focus on low-dimensional feature spaces (i.e., fewer predictors) and simple classifiers as more conservative risk bounds are typically achievable under these conditions [21]. In general, methods such as Rademacher complexity, covering numbers, and PAC-Bayesian bounds tend to provide tighter risk estimates, particularly for high-dimensional, complex, or probabilistic models. Conversely, approaches based on the VC dimension often yield less conservative bounds, especially for simple or finite hypothesis spaces, but they remain easier to compute and apply in basic settings.



(a) P-value analysis using R repetitions in one dimension for the ADNI dataset



(b) Power analysis with increasing dimensions using the ADNI dataset

Fig. 25. NOR and NOR-AD-MCI group analyses using the ADNI dataset.

Although we utilize PAC-Bayesian bounds in this work, the experimental setup allows us to compute them efficiently without incurring significant computational overhead. The theoretical findings are also applicable to real data in various fields of research. Indeed, the models devised in this paper, together with the simulation of realistic datasets, create suitable exemplars for characterizing performance in neuroimaging applications [43–48]. A simple comparison in classification tasks between the simulated datasets and those of real situations reveals a similarity in the results obtained [31]. Nevertheless, the scatter plots and data distributions projected on the dimensions were clear examples demonstrating that the conditions to provide stable inducers [7] were not met. Therefore, exploration of alternative validation methods is a priority.

Emphasizing negative results in scientific research

In summary, echoing practices from our previous work, we underscore the importance of emphasizing negative results to enhance scientific understanding[2]. While research papers often prioritize positive outcomes, we infrequently assess our algorithms in hypothetical task scenarios where no discernible effect is expected. This is evident in our experiments with uncorrelated data. Such analyses play a crucial role serving to approximate the null-distribution of the test-statistic in permutation studies, such as evaluating performance or accuracy in a classification task using ML techniques. For instance, in the permutation analysis of classification tasks, the performance derived from paired data and labels is juxtapsed with that obtained by randomly permuting group labels numerous times with the anticipated distribution centered around 50%. If the performance distribution exhibits non-symmetry around random chance and bias, conclusions drawn from the test data may be compromised. This suggests a disparity in the data distributions across groups under the null hypothesis, violating the i.i.d. assumption. Consequently, the estimation of p-values might lead to inaccurate conclusions at the family-wise level [42].

Conclusions

In this paper, we present a method for validating regression models in the field of machine learning and its applications. The method is related to the F-test of classical hypothesis testing for establishing significance in linear models. We demonstrate that standard ML methods for model validation tend to overinflate FPs, thus requiring these approaches to ensure good replication and extrapolation of results in limited sample sizes. When SAR features are incorporated into classical statistical frameworks, they provide a trade-off between OLS and ML paradigms with excellent control of FPs (around the level of significance). Although these pipelines could face criticism when dealing with non-Gaussian data, we also conducted a formal test based on CIs. This test discarded regression problems with low correlation levels under the worst-case scenario and provided statistical significance for the rest of the cases. The use of this formal test is intended to be combined with classical hypothesis testing, allowing us to confirm the analysis with techniques that are not formally valid, or to inform the researcher that there is insufficient evidence from the perspective of SLT to establish such a linear relationship. This constitutes the formal definition of hypothesis testing: conclusions about the data can only be drawn when we reject the null hypothesis; otherwise, caution must be exercised with the findings.

[2] For further insights, refer to the column in Nature addressing this issue: https://www.nature.com/articles/d41586-019–02960-3

Acknowledgment and CREDIT Author Statement

Grant PID2022-137451OB-I00 funded by MICIU/AEI /10.13039/501100011033 and by ERDF, EU. This research is part of the PID2022-137451OB-I00 and PID2022-137629OA-I00 projects, funded by the CIN/AEI/10.13039/501100011033 and by FSE+. This work was (partially) supported by Programa Iberoamericano de Ciencia y TecnologÚa para el Desarrollo (CYTED) (through Red [225RT0169]). We would also like to thank the reviewers for their contributions to improving this manuscript. **J.M. Gorriz**: Conceptualization, Methodology, Project Administration, Writing- Original draft preparation. **J. Ramirez**: Data curation, Writing- Reviewing and Editing. **F. Segovia**: Visualization, Investigation, Writing- Reviewing and Editing. **C. Jimenez-Mesa**: Conceptualization, Writing- Reviewing and Editing. **F.J. Martinez-Murcia**: Conceptualization, Data Curation, Writing- Reviewing and Editing.: **J. Suckling**: Project Administration, Writing- Reviewing and Editing. 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 ADNIband/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).

Compliance with ethics requirements

The authors confirm that this submission complies with the journal's policies on ethical research and publication.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

A Refined Introduction to PAC-Bayes Bounds

Overview of PAC-Bayes Framework

The PAC-Bayes framework is a robust and elegant tool for establishing generalization bounds in machine learning (ML). The term *PAC-Bayes* stands for *Probably Approximately Correct Bayesian learning* and represents a synthesis of PAC learning principles and Bayesian analysis. This approach enables deriving probabilistic bounds on the generalization error of learning algorithms, which has found extensive applications in classification, regression, and, more recently, deep learning. By focusing on distributions over hypotheses rather than individual ones, PAC-Bayes bounds allow for a nuanced understanding of model performance. Foundational works such as those by [49–53] have established this framework as a key tool in modern ML theory.

At its core, the PAC-Bayes framework leverages a prior distribution $Q(h)$, chosen independently of the data, and a posterior distribution $P(h)$, updated based on observed data. The bounds aim to quantify the generalization error of a hypothesis h probabilistically, where the prior Q and posterior P play central roles. A typical PAC-Bayes bound is expressed as:

$$\mathbb{P}(|\hat{\mathcal{R}}(h) - \mathcal{R}(h)| > \epsilon) \leqslant 2\exp\left(-\frac{N\epsilon^2}{2D_{KL}(P\|Q)}\right),$$

where $\hat{\mathcal{R}}(h)$ denotes the empirical risk (or training error), $\mathcal{R}(h)$ is the true risk (or generalization error), $D_{KL}(P\|Q)$ is the Kullback–Leibler divergence between P and Q, N is the number of samples, and ϵ represents the allowable deviation. A smaller value of $D_{KL}(P\|Q)$, which

measures how much the posterior diverges from the prior, results in tighter generalization bounds.

Derivation of Basic PAC-Bayes Bounds

The derivation of PAC-Bayes bounds begins by quantifying the deviation of the empirical risk $\hat{\mathcal{R}}(h)$ from the true risk $\mathcal{R}(h)$. Let $X = \exp(N(\mathcal{R}(h) - \hat{\mathcal{R}}(h)))$ represent a deviation term. Using Markov's inequality, the probability of significant deviation is bounded by:

$$\mathbb{P}(X \geqslant e^{N\epsilon}) \leqslant \frac{\mathbb{E}[X]}{e^{N\epsilon}}.$$

Substituting X, we get:

$$\mathbb{P}(\mathcal{R}(h) - \hat{\mathcal{R}}(h) > \epsilon) \leqslant e^{-N\epsilon} \cdot \mathbb{E}\left[\exp(N(\mathcal{R}(h) - \hat{\mathcal{R}}(h)))\right].$$

The expectation is taken over the posterior distribution $P(h)$, allowing us to write:

$$\mathbb{E}_{h\sim P}\left[\exp(N(\mathcal{R}(h) - \hat{\mathcal{R}}(h)))\right].$$

Bounding this term relies on the Donsker-Varadhan inequality, which ensures:

$$\mathbb{E}_{h\sim P}\left[e^{N(\mathcal{R}(h)-\hat{\mathcal{R}}(h))}\right] \leqslant e^{D_{KL}(P\|Q)} \cdot \mathbb{E}_{h\sim Q}\left[e^{N(\mathcal{R}(h)-\hat{\mathcal{R}}(h))}\right].$$

Assuming that $\hat{\mathcal{R}}(h)$ is an unbiased estimator of $\mathcal{R}(h)$ under Q, the expectation $\mathbb{E}_{h\sim Q}\left[e^{N(\mathcal{R}(h)-\hat{\mathcal{R}}(h))}\right]$ simplifies to 1. Thus:

$$\mathbb{P}(\mathcal{R}(h) - \hat{\mathcal{R}}(h) > \epsilon) \leqslant e^{-N\epsilon} \cdot e^{D_{KL}(P\|Q)}.$$

Using the union bound to account for symmetric deviations, we obtain:

$$\mathbb{P}(|\mathcal{R}(h) - \hat{\mathcal{R}}(h)| > \epsilon) \leqslant 2\exp\left(-N\epsilon + D_{KL}(P\|Q)\right).$$

Finally, by setting the right-hand side of the inequality to η, we upper bound the actual risk with at least a probability η as a function of $D_{KL}(P\|Q)$. This result demonstrates how PAC-Bayes bounds relate the generalization error $\mathcal{R}(h)$ to the empirical risk $\hat{\mathcal{R}}(h)$ and to the complexity of the hypothesis distribution, as quantified by $D_{KL}(P\|Q)$.

Key Contributions to PAC-Bayes Theory

The development of PAC-Bayes bounds owes much to several seminal works. [49] first introduced the PAC-Bayes framework for classification tasks, showing how generalization error is governed by the KL divergence between posterior and prior distributions. Later, [50] refined these bounds by incorporating stronger concentration inequalities and alternative measures of divergence, such as the *average relative entropy*, leading to tighter results in high-dimensional settings. Subsequent studies, including those by [51,52], extended PAC-Bayes theory to more general hypothesis classes and complex distributions, enabling its application to high-dimensional data and large-scale ML problems. [53] further broadened the scope of PAC-Bayes by developing bounds for deep learning and kernel-based methods, deepening our understanding of generalization in modern ML.

Applications of PAC-Bayes bounds

The PAC-Bayes framework has proven particularly useful for analyzing algorithms that operate over distributions of hypotheses. Notable applications include evaluating generalization in SVMs with non-linear kernels, providing tight bounds for regularized deep learning models, and assessing performance in online learning scenarios with streaming data. By integrating Bayesian inference with prior knowledge, PAC-Bayes offers a powerful probabilistic guarantee on model performance, making it an indispensable tool in theoretical ML.

A note on the theoretical losses presented in caption of Fig. 1

In Fig. 1, we stated that the theoretical expected loss on a uniform distribution of pairs $y, \hat{y}$ was $\mathcal{R} = \frac{a^2}{6b}$ for $\mathcal{L}_1$ and $\mathcal{R} = \frac{b^2+a^2}{3}$ for $\mathcal{L}_2$, where a and b represent the maximum values for $\hat{y}$ and y, respectively, when the number of predictors is $P = 1$. In general, the expected loss is given by:

$$\mathcal{R} = E[\mathcal{L}(f, x, y)] = \int \mathcal{L}(f, x, y) dP(x, y) \tag{B.1}$$

In this simple case with only two dimensions, assume that x, y are a set of uncorrelated samples (uniformly) centered at the origin. Then, the difference $(\beta \cdot x - y)$ is distributed around zero and equally probable within a volume $\mathcal{V}$. Define a and b as the bounds for the uniformly distributed pairs $\{\hat{y} = \beta \cdot x, y\}$ contained in this volume. Then $\mathcal{V} = 2a \times 2b$, and the expected loss can be computed as:

$$\mathcal{R}_u = \frac{1}{\mathcal{V}} \int \mathcal{L}(\hat{y}, y) d\hat{y} dy \tag{B.2}$$

where $\mathcal{V} = 4ab$. If $\mathcal{L}(\hat{y}, y) = |y - \hat{y}|$, this expression simplifies to:

$$\begin{aligned} \mathcal{R}_u &= \tfrac{1}{\mathcal{V}} \int_{-a}^{a} \int_{-b}^{\hat{y}} (\hat{y} - y) d\hat{y} dy + \tfrac{1}{\mathcal{V}} \int_{-a}^{a} \int_{\hat{y}}^{b} (y - \hat{y}) d\hat{y} dy \\ &= \tfrac{1}{4ab}\left(\tfrac{1}{3}a^3 + b^2 a\right) + \tfrac{1}{4ab}\left(\tfrac{1}{3}a^3 + b^2 a\right) = \tfrac{1}{6}\tfrac{a^2}{b} + \tfrac{1}{2}b \end{aligned} \tag{B.3}$$

If we instead choose $\mathcal{L} = (y - \hat{y})^2$, the loss Eq. B.2 becomes:

$$\begin{aligned} \mathcal{R}_u &= \tfrac{1}{\mathcal{V}} \int_{-a}^{a} \int_{-b}^{b} \mathcal{L}(\hat{y}, y) d\hat{y} dy \\ &= \tfrac{1}{4ab}\left(\tfrac{4}{3}b^3 a + \tfrac{4}{3}b a^3\right) = \tfrac{1}{3}(b^2 + a^2) \end{aligned} \tag{B.4}$$

Note that we expect the solution to be flat ($\beta \sim 0$) with uncorrelated data. Consequently, as a approaches zero, the solutions should converge to the mean value of the 1-D loss evaluated on the observed variable (half of the interval and one-third of the square, respectively). Any algorithmic deviation from this ideal solution yields the theoretical expected loss described in Eqs. B.3 and B.4. By examining Fig. 1, we can evaluate how sampling and non-ideal flatness affect the convergence to these theoretical values.

References

[1] Marvin Zelen. Linear Estimation and Related Topics. in Survey of Numerical Analysis edited by John Todd, McGraw-Hill Book Co., Inc., New York; 1963. p. 558–577.

[2] Hilt Donald E, et al. Ridge, a computer program for calculating ridge regression estimates; 1977. doi:10.5962/bhl.title.68934.

[3] Tibshirani Robert. Regression Shrinkage and Selection via the lasso. J Roy Stat Soc Ser B (Methodol) 1996;58(1):267–88. Wiley.

[4] Burges CJC. A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 1998;2(2):121–67.

[5] P Grohs, et al. Mathematical Aspects of Deep Learning. Cambridge University Press. ISBN 9781009025096. doi: 10.1017/9781009025096.

[6] Snee Ronald D. Validation of regression models: methods and examples. Technometrics 1977;19(4):415–28. Nov..

[7] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International Joint Conference on Artificial Intelligence (IJCAI). p. 1–7.

[8] Kleijnen JPC et al. Validation of trace-driven simulation models: regression analysis revisited. Proceedings Winter Simulation Conference. 1996:352–9. 0-7803-3383-7. Dec.

[9] Miller Michael E et al. Validation techniques for logistic regression models. Stat Med 1991;10(8):1213–26. August.

[10] A.I Oredein et al. On Validating Regression Models with Bootstraps and Data Splitting Techniques. Global Journal of Science Frontier Research Volume 11 Issue 6 Version 1.0 September 2011.

[11] Moore, D.S., et al. (2003): Bootstrap Methods and Permutation Tests. In The Practice of Business Statistics Companion, chap 18. W.H. Freeman; First Edition ISBN 978-0716757269.

[12] LeCun Y et al. Deep learning. Nature 2015;521:436–44.
[13] Gorriz JM et al. Artificial intelligence within the interplay between natural and artificial computation: Advances in data science, trends and applications. Neurocomputing 2020;410:237–70. 14 October.
[14] Varoquaux G. Cross-validation failure: Small sample sizes lead to large error bars. NeuroImage 2018;180:68–77.
[15] A. Eklund, et al. Cluster failure: Inflated false positives for fMRI. Proceedings of the National Academy of Sciences Jul 2016, 113 (28) 7900–7905.
[16] Jollans L et al. Quantifying performance of machine learning methods for neuroimaging data. Neuroimage 2019;1(199):351–65. Oct.
[17] J.M. Górriz, et al. A Machine Learning Approach to Reveal the NeuroPhenotypes of Autisms. International journal of neural systems, 1850058. 2019.
[18] Braga-Neto UM, Dougherty ER. Is cross-validation valid for small-sample microarray classification? Bioinformatics 2004;20(3):374–80. doi: https://doi.org/10.1093/bioinformatics/btg419. Feb 12. PMID:14960464.
[19] L.van der Maaten et al. Visualizing Data using t-SNE. Journal of Machine Learning Research 2008 vol 9, num 86, 2579–2605.
[20] Jimenez-Mesa C et al. A non-parametric statistical inference framework for Deep Learning in current neuroimaging. Information Fusion 2023;91:598–611. March.
[21] Gorriz JM et al. Statistical Agnostic Mapping: A framework in neuroimaging based on concentration inequalities. Information Fusion 2021;66:198–212. February.
[22] Tom Viering et al. The Shape of Learning Curves: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence June 2023, pp. 7799–7819, vol. 45 DOI Bookmark: 10.1109/TPAMI.2022.3220744.
[23] Vapnik V. The nature of statistical learning theory. New York Inc: Springer-Verlag; 1995.
[24] Bernhard Scholkopf et al. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press ISBN:978-0-262-19475-4. December 2001.
[25] Huber PJ. Robust estimation of a location parameter. Ann Math Statist 1964;35:73–101.
[26] Chatterjee S, Hadi AS. Regression Analysis by Example. 4th Edition. Hoboken: John Wiley & Sons; 2006. doi: https://doi.org/10.1002/0470055464.
[27] E T Bullmore et al. Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain IEEE Trans Med Imaging (1999) Jan; 18(1):32–42.
[28] 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.
[29] Juan M Gorriz et al. Is K-fold cross validation the best model selection method for Machine Learning? arXiv:2401.16407.
[30] V. Vapnik. Estimation dependencies based on Empirical Data. Springer-Verlach. 1982 ISBN 0-387-90733-5.
[31] Górriz JM et al. On the computation of distribution-free performance bounds: application to small sample sizes in neuroimaging. Pattern Recogn 2019;93:1–13.
[32] Haussler D. Decision theoretic generalizations of the PAC model for neural net and other learning applications. Inf Comput 1992;100(1):78–150.
[33] S. Boucheron et al. Concentration Inequalities: A Nonasymptotic Theory of Independence ISBN: 9780199535255 Oxford University Press.
[34] D. McAllester, A PAC-Bayesian tutorial with a dropout bound. arXiv 10.48550/ARXIV.1307.2118 2013.
[35] T.S. Breusch et al. A Simple Test for heteroscedasticity and random coefficient variation. Econometrica, Sep., 1979, Vol. 47, No. 5 (Sep., 1979), pp. 1287–1294.
[36] R. Koenker. A note on studentizing a test for heteroscedascity. Journal of Econometrics 17. 107-l 12. North-Holland Publishing Company (1981).
[37] Friston KJ. Sample size and the fallacies of classical inference. NeuroImage 2013;81:503–4.
[38] Addinsoft, 2019. XLSTAT statistical and data analysis solution, Long Island, NY, USA. <https://www.xlstat.com>
[39] Rosenblatt JD et al. Better-than-chance classification for signal detection. Biostatistics 2016.
[40] National Academies of Sciences, Engineering, and Medicine. (2019). Reproducibility and Replicability in Science. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303.
[41] Noble S et al. Cluster failure or power failure? Evaluating sensitivity in cluster-level inference. NeuroImage 2020;209:116468.
[42] Phipson B et al. Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn. Statistical Applications in Genetics and Molecular Biology 2010;Vol. 9(Iss. 1). Article 39.
[43] Gorgen K et al. The same analysis approach: Practical protection against the pitfalls of novel neuroimaging analysis methods. NeuroImage 2018;180:19–30.
[44] Zhang Y et al. Multivariate lesion-symptom mapping using support vector regression. Hum Brain Mapp 2014;35(12):5861–76. Dec.
[45] Gorriz JM et al. A connection between pattern classification by machine learning and statistical inference with the General Linear Model. IEEE Journal of Biomedical and Health Informatics 2021.
[46] Gorriz JM et al. A hypothesis-driven method based on machine learning for neuroimaging data analysis. Neurocomputing 2022;510(21):159–71. October.
[47] Wang Z et al. Support vector machine learning-based fMRI data group analysis. NeuroImage 2007;36(4):1139–51.
[48] Wang Z. A hybrid SVM–GLM approach for fMRI data analysis. Neuroimage 2009;46(3):608–15.
[49] McAllester D. PAC-Bayesian model averaging. In: Proceedings of the 12th Annual Conference on Computational Learning Theory (COLT).
[50] Catoni O. PAC-Bayesian inequalities for general loss functions. Theoret Comput Sci 2007.
[51] Alquier P. PAC-Bayesian inequalities and their applications. J Mach Learn Res 2016.
[52] van Erven T et al. The PAC-Bayesian approach to machine learning. Foundations and Trends in Machine Learning 2014.
[53] Canonni F et al. PAC-Bayes bounds for deep learning models. Journal of Machine Learning 2021.