Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

A Goodness-of-Fit Assessment for General Learning Procedures in High Dimensions

Chenxuan He, Canyi Chen & Liping Zhu

To cite this article: Chenxuan He, Canyi Chen & Liping Zhu (29 Sep 2025): A Goodness-of-Fit Assessment for General Learning Procedures in High Dimensions, Journal of the American Statistical Association, DOI: 10.1080/01621459.2025.2529602

To link to this article: https://doi.org/10.1080/01621459.2025.2529602

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2025, VOL. 00, NO. 0, 1–12: Theory and Methods
https://doi.org/10.1080/01621459.2025.2529602

A Goodness-of-Fit Assessment for General Learning Procedures in High Dimensions

Chenxuan He[a,*], Canyi Chen[b,*], and Liping Zhu[a]

[a]Institute of Statistics and Big Data, Renmin University of China, Beijing, China; [b]Department of Biostatistics, University of Michigan, Ann Arbor, MI

ABSTRACT

Black-box learners have demonstrated remarkable success across various fields due to their high predictive accuracy. However, the complexity of their learning procedures poses significant challenges in evaluating whether a given learner has achieved optimal performance on datasets with unknown data-generating mechanisms. We propose a general goodness-of-fit test for assessing different learning procedures involving high-dimensional predictors, encompassing methods from classical linear regression to advanced neural networks. Our goodness-of-fit test leverages data-splitting, using the test set to evaluate the black-box learner trained on the training set. By examining the cumulative covariance of the residuals, our method can effectively handle high-dimensional predictors. Extensive simulations and three real data analyses validate the effectiveness of our method. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received September 2024
Accepted June 2025

KEYWORDS
Cumulative covariance;
General learning procedures;
High-dimensional predictors;
Model checking

1. Introduction

Black-box learners, such as random forests (Breiman 2001a), gradient boosting (Friedman 2001), and deep neural networks (LeCun, Bengio, and Hinton 2015), have garnered significant attention recently due to their high predictive accuracy (Breiman 2001b). Successful applications include handwriting digit recognition, protein structure prediction, and voice transcription (Fan et al. 2020). However, the complexity of their learning processes raises concerns about overfitting, potentially limiting their generalizability. To mitigate this issue, empirical techniques such as dropout (Srivastava et al. 2014) and early stopping (Girosi, Jones, and Poggio 1995) have been developed. Despite these advances, there remains a strong demand for theoretical tools to assess whether a specific black-box learner has achieved the necessary convergence rate to the data-generating process for a given dataset.

Classical prediction metrics, such as mean squared error (MSE), may not serve this purpose well. Consider the example of spurious correlation (Fan, Guo, and Hao 2012; Fan and Zhou 2016). Suppose we have a random sample $\{(\mathbf{x}_i, Y_i)\}_{i=1}^{100}$, where each element of $\mathbf{x}_i \in \mathbb{R}^p$ is independently generated from a standard normal distribution $\mathcal{N}(0,1)$, and each $Y_i \in \mathbb{R}$ is independently generated from a normal distribution $\mathcal{N}(0,\sigma^2)$. We vary the variance σ^2 of Y over $\{1,2\}$ and consider the dimension $p \in \{100, 500, 1000\}$. Clearly, the predictors $\mathbf{x}_i$ are independent of the response Y_i, and the mean of Y_i is the best prediction under ℓ_2-risk. The irreducible error of MSE is hence σ^2. However, the high dimensionality of irrelevant variables introduces significant spurious correlations with the response, placing advanced black-box learners at risk of overfitting. Consider a training set ratio of 80% and the following competitors: the average of the responses, least absolute shrinkage and selection operator (Tibshirani 1996, LASSO), smoothly clipped absolute deviation (Fan and Li 2001, SCAD), support vector regression (Smola and Schölkopf 2004, SVR), XGBoost (Chen and Guestrin 2016), random forest (Breiman 2001a, RF), and feedforward neural network (Schmidhuber 2015, FNN). The average MSEs on the training and test sets over 100 replications are reported in Table 1.

Despite achieving a low training MSE, the black-box learner neural network performs the worst on the test set, and the MSE on the test set deteriorates as p increases. This indicates that prediction-oriented black-box learners can over-extract information from irrelevant predictors, leading to overfitting. Moreover, the MSE is unsuitable for assessing the goodness-of-fit because it incorporates an irreducible error. When the variance of the response $\sigma^2 = 1$, the irreducible error of MSE is 1, implying that a learner is anticipated to have a good fit when its MSE is approximately 1. However, this anticipation changes when $\sigma^2 = 2$, at which point the irreducible error increases to 2. As a result, MSE alone does not provide a reliable assessment of whether an optimal fit has been achieved. This example underscores the need for new toolkits that assess the goodness-of-fit rather than relying solely on prediction capabilities.

The classical goodness-of-fit tests typically count on restricted model classes and are hard to handle general learning procedures. Hosmer and Lemesbow (1980), McCullagh (1985), le Cessie and van Houwelingen (1991), and Janková et al. (2020) worked under the generalized linear model. Continuous responses are tackled by Stute (1997) with nonparametric regression and by Fan and Huang (2001) and Shah and Bühlmann (2018) with parametric regression. In particular, Fan and Huang (2001) introduced an adaptive Neyman test for assessing the goodness-of-fit for parametric regression models by examining whether the expectation of residual terms conditional on

CONTACT Liping Zhu zhu.liping@ruc.edu.cn Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China.
*Chenxuan He and Canyi Chen contributed equally to this work.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

Table 1. Mean squared errors (MSEs) for various learners averaged over 100 replications.

MSE	σ^2	p	Average	LASSO	SCAD	SVR	RF	XGBoost	FNN
Training	1	100	0.974	0.879	0.908	0.225	0.171	*	0.005
		500	0.961	0.792	0.842	0.202	0.160	*	0.004
		1000	0.992	0.803	0.842	0.205	0.162	*	0.005
Test	1	100	0.956	1.047	1.038	1.042	1.057	1.291	2.002
		500	0.939	1.033	1.026	1.007	1.031	1.210	2.317
		1000	0.917	1.025	1.013	0.984	1.003	1.234	2.629
Training	2	100	1.966	1.705	1.796	0.454	0.348	*	0.007
		500	1.986	1.641	1.752	0.406	0.329	*	0.016
		1000	1.969	1.675	1.753	0.411	0.319	*	0.012
Test	2	100	1.932	2.184	2.114	2.150	2.158	2.612	3.547
		500	1.948	2.098	2.092	2.067	2.122	2.593	3.689
		1000	1.873	2.086	2.063	2.027	2.039	2.671	3.610

* indicates less than 0.001.

predictors is close to zero. Other methods, such as kernel-based tests and empirical processes, are also considered (González-Manteiga and Crujeiras 2013). Goodness-of-fit tests for linear and generalized linear models in high dimensions are discussed by Shah and Bühlmann (2018) and Janková et al. (2020), who proposed residual prediction tests and generalized residual prediction testing applicable for both binary and continuous responses in high-dimensional linear models. Their approach involves predicting the residual terms using nonparametric techniques and assessing whether additional information can be captured.

Extending traditional goodness-of-fit tests to black-box learners poses a challenge due to the complex learning procedure. Black-box learners prioritize prediction performance, resulting in limited literature on the assessment of goodness-of-fit. Recently, Zhang, Ding, and Yang (2023) and Javanmard and Mehrabi (2024) proposed unified frameworks for conducting goodness-of-fit tests for binary classifiers, known as BAGofT and GRASP. However, these two tests are tailored for binary data and cannot be applied to continuous data.

In this article, we introduce a goodness-of-fit test procedure for both traditional model-based learners and general black-box learners. We adopt a data-splitting procedure, using the training set to establish the learning process and the test set to evaluate the goodness-of-fit. Our work makes the following three contributions to the literature: (a) We develop a unified framework for conducting goodness-of-fit tests that is applicable to both classic model-based learners and general black-box learners with high-dimensional predictors. In contrast, existing works such as Shah and Bühlmann (2018) are limited to parametric models. (b) Our method accommodates both regression and binary classification tasks, whereas recent works on goodness-of-fit for general learning procedures, such as BAGofT (Zhang, Ding, and Yang 2023) and GRASP (Javanmard and Mehrabi 2024), are restricted to binary classifiers. (c) Through simulation studies, we compare our method with several existing approaches, demonstrating its robust performance.

We organized this article as follows: In Section 2, we propose a testing procedure and provide its theoretical foundations. Section 3 covers various topics related to implementing the proposed test statistic. We present simulation results and real-data examples in Sections 4 and 5, respectively. Section 6 concludes the article with brief discussions. Technical proofs and additional simulation results are relegated to the online supplementary material.

We use the following notations throughout this article. We use $\xrightarrow{p}$ to represent convergence in probability and $\xrightarrow{d}$ to denote convergence in distribution. For two sequences a_n and b_n, the notation $a_n = O(b_n)$ or $a_n \lesssim b_n$ indicates that there exists a constant C_0 such that $|a_n| \leq C_0|b_n|$ as n goes to infinity. The notation $a_n \asymp b_n$ means $a_n = O(b_n)$ and $b_n = O(a_n)$. The notation $a_n = O_p(b_n)$ implies that $a_n = O(b_n)$ holds with high probability. Furthermore, $a_n = o(b_n)$ denotes that $a_n/b_n \to 0$ as $n \to \infty$, and $a_n = o_p(b_n)$ indicates that $a_n/b_n \to 0$ with high probability. We write $a_n = \omega(b_n)$ if $b_n = o(a_n)$, and $a_n = \Omega(b_n)$ if $b_n = O(a_n)$. Similarly, $a_n = \omega_p(b_n)$ or $a_n = \Omega_p(b_n)$ signifies that $a_n = \omega(b_n)$ or $a_n = \Omega(b_n)$ holds with high probability, respectively. Let $\mathbf{I}_p$ be the p-dimensional identity matrix.

2. Methodology

2.1. Problem Formulation

In this section, we formalize the problem and the underlying hypothesis. Consider the following general regression model:

$$Y = E(Y \mid \mathbf{x}) + \epsilon \stackrel{\text{def}}{=} m(\mathbf{x}) + \epsilon, \tag{2.1}$$

where $Y \in \mathbb{R}$ denotes the response variable, which is decomposed into a systematic component $m(\mathbf{x})$ that characterizes the potentially complex relationship with the predictors $\mathbf{x} \in \mathbb{S} \subseteq \mathbb{R}^p$, and a stochastic noise term ϵ satisfying $E(\epsilon \mid \mathbf{x}) = 0$.

Let $\mathcal{D}_n = \{(\mathbf{x}_i, Y_i)\}_{i=1}^n$ be a random sample of size n drawn from the joint distribution of $(\mathbf{x}, Y)$. Given a general learning procedure, let $m_{\mathcal{D}_n}(\mathbf{x})$ denote the estimate of the conditional expectation $m(\mathbf{x})$ based on the sample $\mathcal{D}_n$. Using $m_{\mathcal{D}_n}(\mathbf{x})$, we can rewrite (2.1) as

$$Y = m_{\mathcal{D}_n}(\mathbf{x}) + \varepsilon, \tag{2.2}$$

where $\varepsilon = Y - m_{\mathcal{D}_n}(\mathbf{x}) = \epsilon + m(\mathbf{x}) - m_{\mathcal{D}_n}(\mathbf{x})$ represents the residual term.

Remark 1 (Error decomposition). The mean squared error $E\{Y - m_{\mathcal{D}_n}(\mathbf{x})\}^2$ decomposes into two components:

$$E\{Y - m_{\mathcal{D}_n}(\mathbf{x})\}^2 = E\{Y - m(\mathbf{x})\}^2 + E\{m(\mathbf{x}) - m_{\mathcal{D}_n}(\mathbf{x})\}^2,$$

which correspond the variance of $\epsilon = Y - m(\mathbf{x})$ and the deviance of $m_{\mathcal{D}_n}(\mathbf{x})$ relative to $m(\mathbf{x})$. Notably, the variance of ϵ is irreducible. Our primary objective is to assess the extent to which a given learning procedure controls the deviance of $m_{\mathcal{D}_n}(\mathbf{x})$, which can be quantified by the condition $\sup_{\mathbf{x}\in\mathbb{S}} |m_{\mathcal{D}_n}(\mathbf{x}) - m(\mathbf{x})| = O_p(r_n)$ for some rate $r_n > 0$.

Remark 2 (Binary classification). Our framework naturally extends to binary classification. When the response Y takes values in $\{0, 1\}$, the conditional expectation reduces to $m(\mathbf{x}) = E(Y \mid \mathbf{x}) = \Pr(Y = 1 \mid \mathbf{x})$. The objective is to estimate the underlying conditional probability $\Pr(Y = 1 \mid \mathbf{x})$, which aligns with the model specified in (2.1).

The general learning procedure may be either a parametric learner or a black-box learner. For a parametric learner, the fitted conditional expectation can be represented as $m_{\mathcal{D}_n}(\mathbf{x}) = f(\mathbf{x}, \hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ lies in a finite-dimensional parameter space. In contrast, a general black-box learner defines a mapping from the predictor space to the response space, that is, $m_{\mathcal{D}_n}: \mathbb{S} \to \mathbb{R}$. If the model is correctly specified, $m_{\mathcal{D}_n}(\mathbf{x})$ converges to the true conditional expectation $m(\mathbf{x})$ at some rate r_n. We quantify this convergence using the infinity norm $\sup_{\mathbf{x}\in\mathbb{S}} |m_{\mathcal{D}_n}(\mathbf{x}) - m(\mathbf{x})| = O_p(r_n)$. Consequently, the residual $\varepsilon = Y - m_{\mathcal{D}_n}(\mathbf{x})$ satisfies $\varepsilon = \epsilon + O_p(r_n)$. Because $E(\epsilon \mid \mathbf{x}) = 0$ by definition in (2.1), it follows that if $\sup_{\mathbf{x}\in\mathbb{S}} |m_{\mathcal{D}_n}(\mathbf{x}) - m(\mathbf{x})| = O_p(r_n)$, then $E(\varepsilon \mid \mathbf{x}) = O_p(r_n)$.

Thus, our test is designed to assess whether and how fast the conditional expectation $E(\varepsilon \mid \mathbf{x})$ converges to zero. In symbols, given the observed data $\mathcal{D}_n$ and the estimated conditional expectation $m_{\mathcal{D}_n}(\mathbf{x})$, we aim to test

$$\begin{aligned} H_0:&\ Y = m_{\mathcal{D}_n}(\mathbf{x}) + \varepsilon,\ \sup_{\mathbf{x}\in\mathbb{S}} |m_{\mathcal{D}_n}(\mathbf{x}) - m(\mathbf{x})| = O_p(r_n), \\ &\text{with } r_n \to 0 \text{ as } n \to \infty, \text{ versus} \\ H_1:&\ \exists \mathbb{M}_n \subseteq \mathbb{S} \text{ with } \Pr(\mathbf{x} \in \mathbb{M}_n) \text{ bounded away from 0} \\ &\text{such that} \\ &\inf_{\mathbf{x}\in\mathbb{M}_n} |m_{\mathcal{D}_n}(\mathbf{x}) - m(\mathbf{x})| = \Omega_p(r_n^{(a)}),\ \text{where } r_n = o(r_n^{(a)}), \end{aligned} \tag{2.3}$$

where r_n represents the convergence rate of a general learning procedure, which may be unknown for general learning procedures. Under the alternative hypothesis, $r_n^{(a)}$ may either decay at a slower rate than r_n or fail to converge to zero, indicating that the learning procedure converges more slowly or fails to converge to $m(\mathbf{x})$. Consequently, depending on the specific choice of r_n, a learning procedure may fall anywhere between H_0 and H_1, where a smaller r_n corresponds to faster convergence for a given learning procedure.

We illustrate the rates r_n and $r_n^{(a)}$ in our hypothesis formulation (2.3). For parametric learners with non-convex penalties such as SCAD, under suitable regularity conditions, the convergence of the infinity norm, $\sup_{\mathbf{x}\in\mathbb{S}} |m_{\mathcal{D}_n}(\mathbf{x}) - m(\mathbf{x})| \le (s/n)^{1/2}$, holds with high probability, where s denotes the number of nonzero coefficients in the high-dimensional linear model (Shi et al. 2019; Peng, Wang, and Wu 2016). Consequently, by selecting $r_n = (s/n)^{1/2}$, the null hypothesis H_0 is satisfied. Conversely, when the high-dimensional linear model is misspecified, under appropriate conditions, the model falls under the alternative hypothesis H_1, and $r_n^{(a)}$ becomes a constant. For parametric learners with Lasso-type penalties (Tibshirani 1996; Wang et al. 2020), the rate r_n is usually $(s \log p/n)^{1/2}$. For random forest, XGBoost, and linear support vector machine, we refer interested readers to Scornet, Biau, and Vert (2015), Peng, Wang, and Wu (2016), and Gao and Zhou (2020), respectively, for detailed rates of r_n under certain models.

For black-box learners, we adopt a nonparametric perspective. For the Hölder class, Stone (1982) established that the minimax optimal rate of convergence in the infinity norm for nonparametric regression is $r_n = (\log n/n)^{d/(2d+p)}$, where d is the smoothness index and the predictor space is p-dimensional. This minimax rate can be (nearly) attained by methods such as kernel regression (Hardle, Janssen, and Serfling 1988) and deep learning regression (Imaizumi 2023), among others. For a given r_n, if the rate $(\log n/n)^{d/(2d+p)}$ is faster, the model remains under the null hypothesis H_0. However, when the dimensionality p is high or the smoothness d is low, this rate may be slower than r_n, placing the model under the alternative hypothesis H_1, with $r_n^{(a)}$ denoting the corresponding rate. In high-dimensional settings, if the covariates $\mathbf{x}$ lie on a low-dimensional manifold, the convergence rate may be improved through effective dimension reduction (Bauer and Kohler 2019; Jiao et al. 2023).

2.2. Cumulative Covariance Revisited

Before presenting our testing framework for evaluating the goodness-of-fit of general learning procedures, we first revisit the concept of cumulative covariance (Zhou et al. 2020; Li et al. 2023).

According to the definition of the residual term in (2.2), when $m_{\mathcal{D}_n}(\mathbf{x}) = m(\mathbf{x})$, the error ϵ equals the residual ε, whose expectation satisfies $E(\varepsilon \mid \mathbf{x}) = E(\epsilon \mid \mathbf{x}) = 0$. This scenario corresponds to the original testing problem in Li et al. (2023). However, when only an estimate of $m(\mathbf{x})$ is available, under the null hypothesis H_0, the uniform bound $\sup_{\mathbf{x}\in\mathbb{S}} |m_{\mathcal{D}_n}(\mathbf{x}) - m(\mathbf{x})| = O_p(r_n)$ implies that $E(\varepsilon \mid \mathbf{x}) = O_p(r_n)$. The primary objective, in turn, becomes to test whether $E(\varepsilon \mid \mathbf{x})$ converges to zero at the rate r_n.

Let $\tilde{Z}$ be an independent copy of a random variable Z. The cumulative covariance $\mathrm{CCov}(\varepsilon \mid X_k)$ is defined as

$$\mathrm{CCov}(\varepsilon \mid X_k) = E[\mathrm{cov}^2\{\varepsilon, I(X_k < \tilde{X}_k) \mid \tilde{X}_k\}], \tag{2.4}$$

where $I(\cdot)$ denotes the indicator function.

When working on high-dimensional predictors, we opt to use a state-of-the-art test based on cumulative covariance proposed by Li et al. (2023), which is well-suited for assessing the effect of predictors on the response in high dimensions. Under the null hypothesis H_0, $\sup_{\mathbf{x}\in\mathbb{S}} |m_{\mathcal{D}_n}(\mathbf{x}) - m(\mathbf{x})| = O_p(r_n)$ holds, indicating that $\mathrm{CCov}(\varepsilon \mid X_k) = o_p(1)$. This naturally motivates us to construct the test statistic based on the sum of all marginal cumulative covariances:

$$\sum_{k=1}^{p} \mathrm{CCov}(\varepsilon \mid X_k). \tag{2.5}$$

To test the hypothesis in (2.3), we aim to construct an asymptotically normal estimate of (2.5) that could greatly ease the computational cost for general learning procedures.

Remark 3. In constructing the test statistic based on (2.5) to assess H_0, we aggregate the marginal effects measured by $\mathrm{CCov}(\varepsilon \mid X_k)$. This formulation ensures that the cumulative covariance converges to zero under the null hypothesis H_0 and diverges under the alternative hypothesis H_1, which we will show in the next section. The key reason underlying this construction is that $E(\varepsilon) = o(1)$ does not necessarily imply $E(\varepsilon \mid \mathbf{x}) = o(1)$ uniformly over $\mathbf{x}$, as unconditional expectations may obscure local deviations. This kind of localized idea was first introduced by Zhang, Ding, and Yang (2023) to test general classification procedures. Let $\{\mathbb{S}_t, t = 1, \ldots, T\}$ be a partition of the support of $\mathbf{x}$. Their method effectively tests the convergence of $E(\varepsilon \mid \mathbf{x} \in \mathbb{S}_t)$, for $t = 1, \ldots, T$. Such a localized

approach to testing conditional expectations provides a more general and powerful framework, as it leverages the relationship, $E(\varepsilon \mid \mathbf{x}) = O(r_n)$ uniformly over $\mathbf{x} \implies E(\varepsilon \mid \mathbf{x} \in \mathbb{S}_t) = O(r_n)$, uniformly over $\mathbf{x}$, $t = 1, \ldots, T \implies E(\varepsilon) = O(r_n)$.

2.3. Test Statistic

In this section, we develop our testing procedure and establish the asymptotic properties of the resulting test statistic.

To perform the test, we adopt a data-splitting strategy, partitioning the dataset $\mathcal{D}_n$ into training and test subsets, denoted as $\mathcal{D}_{n_1} = \{(\mathbf{x}_{1,i}, Y_{1,i})\}_{i=1}^{n_1}$ and $\mathcal{D}_{n_2} = \{(\mathbf{x}_{2,i}, Y_{2,i})\}_{i=1}^{n_2}$, respectively, where $n = n_1 + n_2$. Given a general learning procedure, we use the training set to estimate the conditional expectation function, denoted as $m_{\mathcal{D}_{n_1}}(\mathbf{x})$. The residuals computed from the test set, $\{\varepsilon_{2,i}\}_{i=1}^{n_2}$, are then used to estimate the cumulative covariance $\sum_{k=1}^{p} \mathrm{CCov}(\varepsilon \mid X_k)$.

First, we introduce some notation. We denote $(n)_m = n(n-1)\ldots(n-m+1)$ for $1 \le m \le n$, and

$$\sum_{(i,j)}^{n}, \quad \sum_{(i,j,s)}^{n}, \quad \sum_{(i,j,s,l)}^{n}, \quad \text{and} \quad \sum_{(i,j,s,l,r)}^{n},$$

denote summations that are taken over all possible permutations of distinct indices. Let $\mathbf{x}_{2,i} = (X_{2,i1}, \ldots, X_{2,ip})^\top$ and $\psi(X_1, X_2, X_3) = I(X_1 < X_3) - I(X_2 < X_3)$. Consider the following estimator for $\sum_{k=1}^{p} \mathrm{CCov}(\varepsilon \mid X_k)$,

$$T = \frac{1}{4(n_2)_5} \sum_{k=1}^{p} \sum_{(i,j,s,l,r)}^{n_2} (\varepsilon_{2,i} - \varepsilon_{2,j})(\varepsilon_{2,s} - \varepsilon_{2,l}) \psi(X_{2,ik}, X_{2,jk}, X_{2,rk}) \times \psi(X_{2,sk}, X_{2,lk}, X_{2,rk}), \tag{2.6}$$

and an estimator for the variance of T,

$$S^2 = \{4c_{n_2} n_2 (n_2 - 1)\}^{-1} \sum_{(i,j)}^{n_2} (\varepsilon_{2,i}\varepsilon_{2,j})^2 \times \Big[\sum_{k=1}^{p} K_1\{F_{n_2,k}(X_{2,ik}), F_{n_2,k}(X_{2,jk})\}\Big]^2, \tag{2.7}$$

where $F_{n_2,k}$ is the empirical distribution function of X_k, $K_1\{F_{n_2,k}(X_{1k}), F_{n_2,k}(X_{2k})\} = F_{n_2,k}^2(X_{1k}) + F_{n_2,k}^2(X_{2k}) - 2\max\{F_{n_2,k}(X_{1k}), F_{n_2,k}(X_{2k})\} + 2/3$, and $c_n = \{(1 - n^{-1})^2 + n^{-2}\}^2$ is a finite sample adjustment factor in reducing the bias of S^2. Our test statistic is $\{n_2(n_2-1)/2\}^{1/2} T/S$.

We impose the following assumptions to establish the asymptotics of $\{n_2(n_2-1)/2\}^{1/2} T/S$ under the null hypothesis.

Assumption 1. Assume that

$$0 < c \le \mathrm{var}(\epsilon \mid \mathbf{x}) \le E^{1/2}[\{\epsilon - E(\epsilon \mid \mathbf{x})\}^4 \mid \mathbf{x}] \le C < \infty,$$

almost surely for some constants c and C.

Assumption 2. Define $V(\mathbf{x}_1, \mathbf{x}_2) = \sum_{k=1}^{p} K_1\{F_k(X_{1k}), F_k(X_{2k})\}$, where F_k is the population counterpart of $F_{n_2,k}$. As $p \to \infty$ and $n_2 \to \infty$,

$$E\{V(\mathbf{x}_1, \mathbf{x}_2) V(\mathbf{x}_2, \mathbf{x}_3) V(\mathbf{x}_3, \mathbf{x}_4) V(\mathbf{x}_4, \mathbf{x}_1)\} / E^2\{V(\mathbf{x}_1, \mathbf{x}_2)^2\} \to 0,$$
$$E\{V(\mathbf{x}_1, \mathbf{x}_2)^4\} / [n_2 E^2\{V(\mathbf{x}_1, \mathbf{x}_2)^2\}] \to 0.$$

Assumptions 1 and 2 are used by Li et al. (2023) to ensure the central limit theorem. Assumption 2 imposes an implicit requirement on the dimension p. As shown by Li et al. (2023), for d-dependent predictors $\mathbf{x}$, Assumption 2 in fact requires $d = o(p^{1/3})$ for divergent p. We impose Assumption 3 to ensure the convergence of the estimator $m_{\mathcal{D}_n}(\mathbf{x})$ under H_0.

Assumption 3. Under H_0, $\sup_{\mathbf{x}\in\mathbb{S}} |m_{\mathcal{D}_n}(\mathbf{x}) - m(\mathbf{x})| = O_p(r_n)$ as $n \to \infty$, with $r_n \to 0$ as $n \to \infty$.

The assumption $\sup_{\mathbf{x}\in\mathbb{S}} |m_{\mathcal{D}_n}(\mathbf{x}) - m(\mathbf{x})| = O_p(r_n)$ serves as a general criterion for evaluating the underlying learning procedure. The infinity norm, which underpins this assumption, is widely employed, including in the well-known Kolmogorov–Smirnov test and the pioneering goodness-of-fit test for binary-data learning procedures in Zhang, Ding, and Yang (2023). It quantifies the sensitivity of a learning procedure to deviations across the entire distribution of $\mathbf{x}$, ensuring that substantial discrepancies in specific regions are not overlooked, which is potentially critical in applications such as risk assessment and quality control. Unlike integrated norms used in other tests, the infinity norm prevents localized differences from being masked by averaging. The infinity norm convergence assumption commonly requires that the support $\mathbb{S}$ be compact in linear regression and that the density of $\mathbf{x}$ be bounded away from zero in kernel regression. For example, Stone (1982) used the assumption that the density of $\mathbf{x}$ is absolutely continuous and bounded away from zero to derive the minimax optimal convergence rates for estimating functions in the Hölder class.

Theorem 1 (Asymptotic behavior under H_0). Under Assumptions 1–3 and the null hypothesis H_0, if the training set size n_1 and test set size n_2 satisfy $n_2 = o(r_{n_1}^{-1})$ as $n \to \infty$, then the test statistic $\{n_2(n_2-1)/2\}^{1/2} T/S$ converges in distribution to $\mathcal{N}(0,1)$ as $n_2, p \to \infty$.

Theorem 1 indicates that the test statistic will converge to a normal distribution for a suitable splitting ratio. Next, we consider the convergence under the alternative.

Assumption 4. Under H_1, there exist another sequence $r_n^{(a)}$ which satisfies $r_n = o\{r_n^{(a)}\}$, and $\exists \mathbb{M}_n \subseteq \mathbb{S}$ with $\Pr(\mathbf{x} \in \mathbb{M}_n)$ bounded away from 0 such that $\inf_{\mathbf{x}\in\mathbb{M}_n} |m_{\mathcal{D}_n}(\mathbf{x}) - m(\mathbf{x})| = \Omega_p\{r_n^{(a)}\}$.

Theorem 2 (Asymptotic behavior under H_1). Under Assumptions 1, 2, and 4, and the alternative hypothesis H_1, if the training set size n_1 and test set size n_2 satisfy $\{r_{n_1}^{(a)}\}^{-1} = o(n_2)$ as $n \to \infty$, then the absolute test statistic $|\{n_2(n_2-1)/2\}^{1/2} T/S|$ diverges to infinity in probability as $n_2, p \to \infty$.

If $r_{n_1}^{(a)}$ fails to converge to zero, as $n_2 \to \infty$, the consistency can always hold. Otherwise, if $r_{n_1}^{(a)}$ is also converging to zero but with a slower rate compared with r_{n_1}, we can use a different splitting ratio to detect the true converging rate for a specific learning procedure. In Section 3, we will elaborate on the details of the implementation of our test. The following corollary indicates that we can control the Type-I error and achieve power simultaneously with a suitable splitting ratio.

Corollary 3 (Consistency under H_0 and H_1). Under Assumptions 1–4, if the sample is split such that the test set size n_2 satisfies $n_2 = \omega[\{r_{n_1}^{(a)}\}^{-1}]$ and $n_2 = o(r_{n_1}^{-1})$ as $n \to \infty$, then the test statistic $\{n_2(n_2-1)/2\}^{1/2} T/S$ converges in distribution to $\mathcal{N}(0,1)$ under H_0 as $n_2, p \to \infty$, while under H_1, its absolute value $|\{n_2(n_2-1)/2\}^{1/2} T/S|$ diverges to infinity in probability as $n_2, p \to \infty$.

3. Implementation of the Test Statistic

3.1. Different Splitting Ratios in Assessing General Learning Procedures

The convergence rate r_n in Theorem 1 may be unknown for the general learning procedures. However, we can use different splitting ratios to comprehensively diagnose the general learning procedure. Specifically, when we reject H_0 in (2.3) at a lower ratio, the more confident we are likely to regard the learning procedure as inappropriate. Following from Zhang, Ding, and Yang (2023), when evaluating general learning procedures, we consider three splitting ratios, where the training set size is 90%, 75%, and 50%, and four patterns of the assessment result are as follows.

Pattern 1. The p-value fails to reject H_0 under all splitting ratios, which means the learning procedure converges fast to the underlying $m(\mathbf{x})$

Pattern 2. The p-value rejects H_0 only under 50% splitting ratio, which means the general learning procedure converges to the true $m(\mathbf{x})$ with a fairly fast rate.

Pattern 3. The p-value rejects H_0 under both 50% and 75% splitting ratios, which means the general learning procedure converges to the true $m(\mathbf{x})$ with a relatively slow rate.

Pattern 4. The p-value rejects H_0 under all splitting ratios, which means the general learning procedure fails to converge to the true $m(\mathbf{x})$.

In practice, we can deploy more kinds of splitting ratios to evaluate the convergence of the learning procedure comprehensively.

3.2. Combining Results from Multiple Splitting

A single data split often results in unnecessary power loss, as only a subset of the data is used to construct the test statistic. To address this limitation, we advocate for a multiple-splitting approach via cross-validation, which enhances testing power while maintaining computational efficiency for complex learning procedures. Specifically, the data are randomly partitioned into $\mathbb{K}$ folds, with $\mathbb{K}-1$ folds used for training the general learner and the remaining fold reserved for testing. For the splitting ratios of 50%, 75%, and 90% discussed in Section 3.1, the data can be divided into 2, 4, and 10 folds, respectively. Let p_{0i}, $i = 1, 2, \ldots, m$ denote the individual p-values obtained from m multiple tests, where $m \leq \mathbb{K}$. To combine these p-values, we employ the Cauchy combination test proposed by Liu and Xie (2020). The test statistic is computed as follows:

$$T_c = \frac{1}{m}\sum_{i=1}^{m} \tan\left\{(0.5 - p_{0i})\pi\right\}. \tag{3.1}$$

Under the null hypothesis, each p_{0i} is uniformly distributed on the interval $[0,1]$. Consequently, each transformed component $\tan\{(0.5-p_{0i})\pi\}$ follows a standard Cauchy distribution. Remarkably, due to the stability property of the Cauchy distribution, the test statistic T_c retains a standard Cauchy distribution when the p-values are mutually independent or perfectly dependent. In cases where the p-values exhibit dependence, the combined p-value can be approximated by:

$$p_c = 0.5 - \arctan(T_c)/\pi.$$

The validity of the Cauchy combination test within our framework is further discussed in Appendix B.4.

Beyond the Cauchy combination, any valid p-value combination method can be applied. The problem of combining p-values to enhance testing power has a rich history and has recently garnered renewed interest. Classical methods, such as Fisher's, Pearson's, Stouffer's, and Tippett's approaches, typically require the assumption of independence among p-values. Recent advancements have focused on relaxing this stringent assumption; notable examples include the average-based method in DiCiccio, DiCiccio, and Romano (2020), the quantile-based method in Rüschendorf (1982), and the correlation-adjusted method proposed by Liu, Yu, and Li (2022).

4. Simulation Studies

In this section, we present extensive simulation studies to evaluate the finite-sample performance of our proposed method across five models. Specifically, Models 1–4 pertain to continuous responses, while Model 5 addresses binary outcomes.

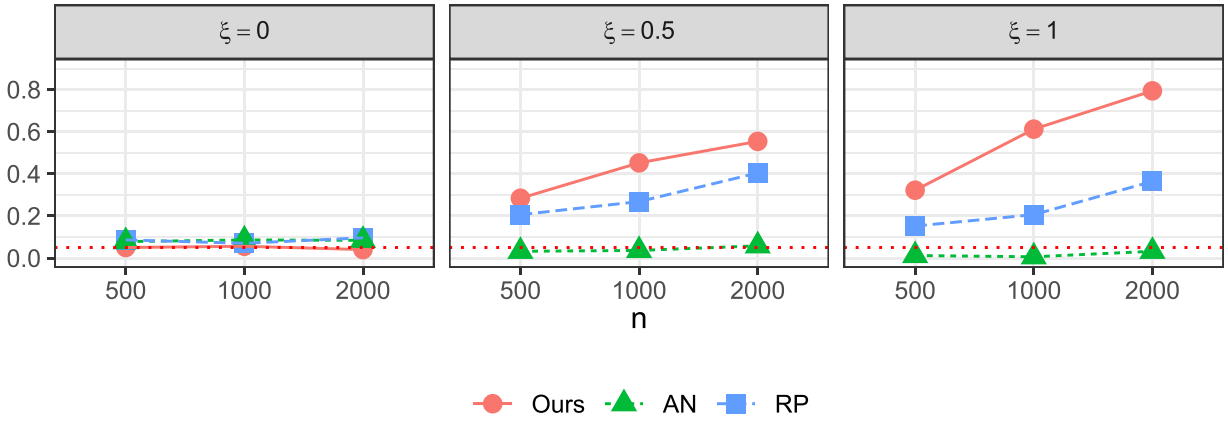
4.1. Comparison with Parametric Tests

We begin by demonstrating the competitive performance of our proposed test relative to two celebrated parametric goodness-of-fit tests for linear models: the adaptive Neyman (AN) test (Fan and Huang 2001) and the residual prediction (RP) test proposed by Shah and Bühlmann (2018).

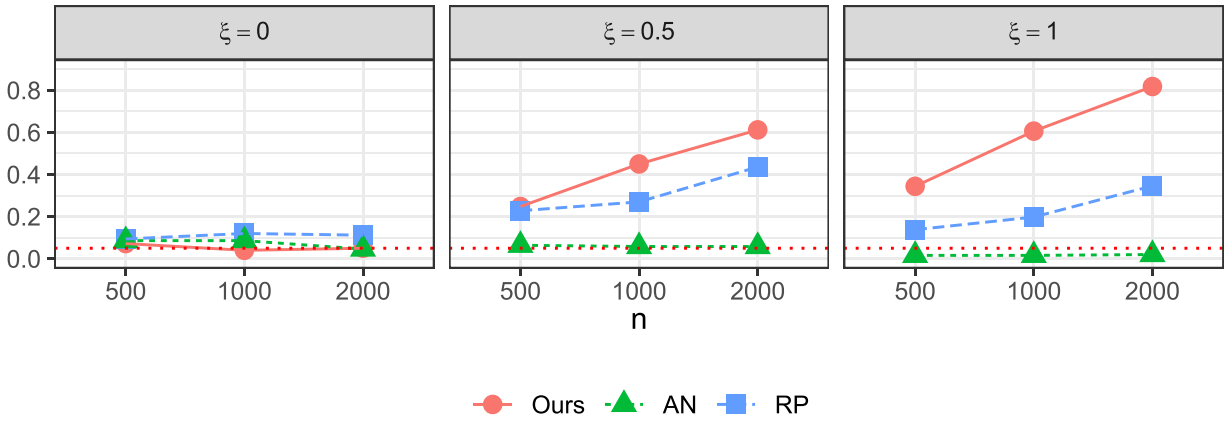
Model 1 (Type-I error and power comparison). Let $\mathbf{x} = (X_1, \ldots, X_p)^\top \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ denote a p-dimensional predictor. The response variable Y is generated as

$$Y = \beta_1 X_1 + \beta_2 X_2 + \exp\{\xi(X_3 + X_4^2 + |X_5|)\} + \epsilon,$$

where β_1 and β_2 are drawn from $\mathcal{N}(1,1)$, the error term $\epsilon \sim \mathcal{N}(0, 0.1^2)$, and the parameter $\xi \in \mathbb{R}$ regulates the data-generating process. When $\xi = 0$, the null hypothesis H_0 holds for all methods, whereas nonzero values of ξ correspond to the alternative hypothesis H_1. We consider two settings for the predictor covariance: independent predictors with $\mathbf{\Sigma} = \mathbf{I}_p$ and correlated predictors with $\mathbf{\Sigma} = (0.2^{|i-j|})_{p\times p}$.



(a) Independent predictors with $\mathbf{\Sigma} = \mathbf{I}_p$



(b) Correlated predictors with $\mathbf{\Sigma} = (0.2^{|i-j|})_{p\times p}$

Figure 1. Empirical Type-I error and power under Model 1. The horizontal axis denotes the sample size n, and the vertical axis shows the rejection rate. The red dotted line indicates the nominal significance level of 0.05. Results are based on 500 replications.

Data are generated from Model 1 with $\xi = 0$ to examine Type-I error and with $\xi = 0.5$ and $\xi = 1$ to assess power. Under each scenario, we fix the dimension $p = 100$ and vary the sample size n over {500, 1000, 2000}. Because the AN test is designed for low-dimensional settings, we provide it with oracle information by fitting least squares with covariates $(X_1, \ldots, X_5)$. For the RP test, we use the default settings from its corresponding R package `RPtests`. To ensure a fair comparison, we adopt a single data-splitting scheme for our test. Specifically, we first fit the data using SCAD and subsequently evaluate the residuals on a held-out test set. The splitting ratios are set to 95%, 96%, and 97% for $n = 500$, 1000, and 2000, respectively. The choice of splitting ratios is guided by Theorem 1, which requires $n_2 = o(r_{n_1}^{-1})$ to ensure proper Type I error control. Under the null hypothesis, SCAD achieves a convergence rate of $(s/n_1)^{1/2}$ with the training set, leading us to set $r_{n_1} = (s/n_1)^{1/2}$. Consequently, the splitting ratio must satisfy $n_2^2/n_1 \lesssim 1$, a condition met by our chosen ratios across different sample sizes. Figure 1 displays empirical rejection rates of all three tests over 500 replications.

Figure 1 illustrates that our test maintains proper Type-I error control under H_0 while achieving the highest power under H_1. In contrast, the AN test exhibits minimal power, likely due to its limitations in high-dimensional settings.

4.2. Performance on General Learning Procedures

We next evaluate our method within the context of general regression models as described in Models 2–4. We consider both linear and nonlinear models and employ varying learning procedures. To the best of our knowledge, no alternative tests exist for direct comparison in these scenarios. Accordingly, we vary the data-splitting ratio as suggested in Section 3.1.

Model 2 (High-dimensional linear model). Let $\mathbf{x} = (X_1, \ldots, X_p)^\top \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ denote a p-dimensional predictor. The response Y is generated according to $Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$, where $\beta_j \sim \mathcal{N}(1, 0.1^2)$ for $j = 1, \ldots, 5$, and $\epsilon \sim \mathcal{N}(0, 1)$.

Model 3 (High-dimensional independent predictor and nonlinear relationship with response). Let $\mathbf{x} = (X_1, \ldots, X_p)^\top \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ denote a p-dimensional predictor. The response Y is defined as $Y = \{-3 + 6 \cdot I(|\beta_1 X_1| < 1)\} + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 +$

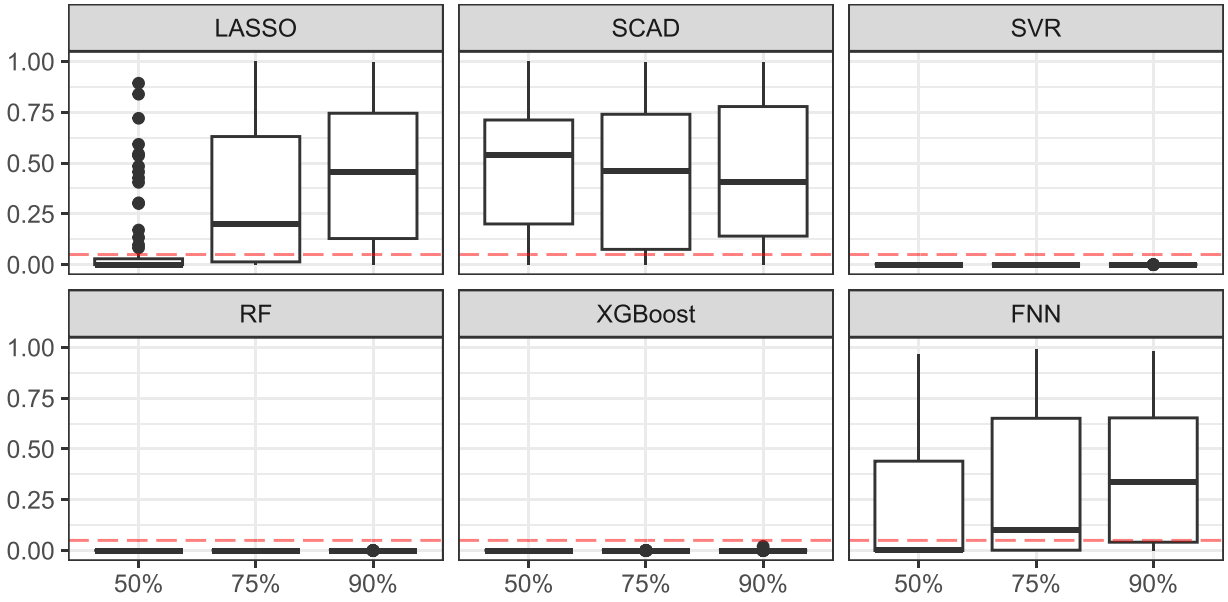


Figure 2. Boxplot of the p-value statistics from Model 2 with $p = 1000$. FNN denotes a feedforward neural network, RF represents random forest, and SVR stands for support vector regression. The sample size is $n = 500$. The horizontal axis indicates different splitting ratios, while the vertical axis displays the corresponding p-values. The red dashed line marks the significance level at 0.05. Results are based on 100 replications.

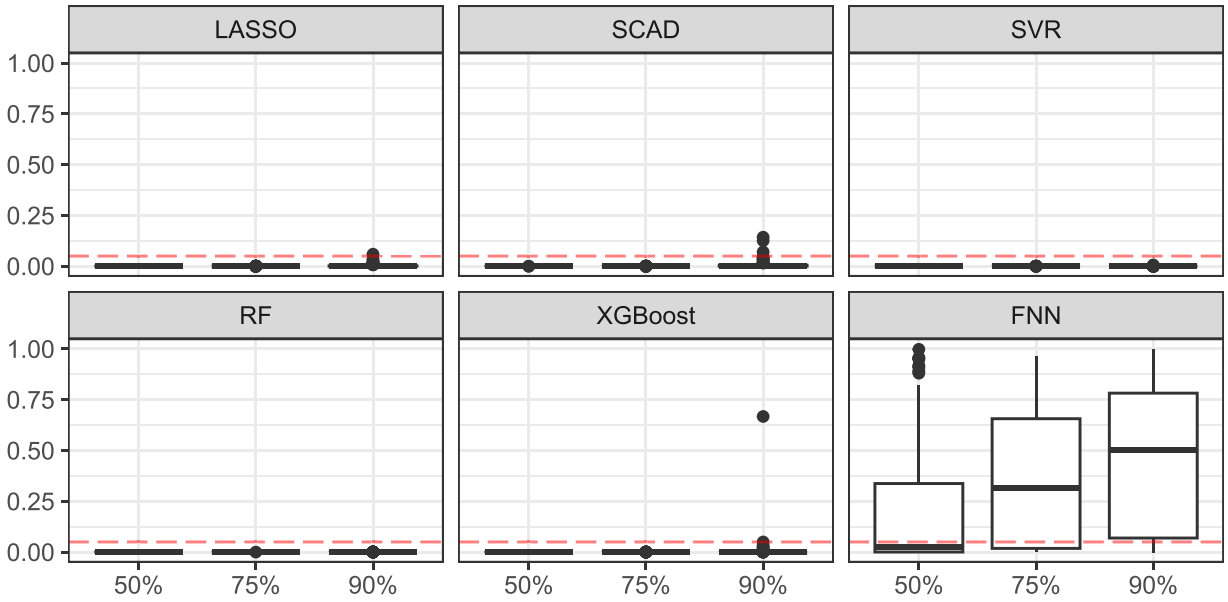


Figure 3. Boxplot of the p-value statistics from Model 3 with $p = 1000$. FNN denotes a feedforward neural network, RF represents random forest, and SVR stands for support vector regression. The sample size is $n = 500$. The horizontal axis indicates different splitting ratios, while the vertical axis displays the corresponding p-values. The red dashed line marks the significance level at 0.05. Results are based on 100 replications.

$\beta_5 X_5 + \epsilon$, where $\beta_j \sim \mathcal{N}(1, 0.1^2)$, for $j = 1, \ldots, 5$, and $\epsilon \sim \mathcal{N}(0, 1)$.

Model 4 (High-dimensional correlated predictor and nonlinear relationship with response). Let $\mathbf{x} = (X_1, \ldots, X_p)^\top \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ denote a p-dimensional predictor with $\boldsymbol{\Sigma} = (0.5^{|i-j|})_{p\times p}$. The response Y is generated as $Y = \beta_1 X_1^2 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$, where $\beta_j \sim \mathcal{N}(1, 0.1^2)$ for $j = 1, \ldots, 5$, and $\epsilon \sim \mathcal{N}(0, 1)$.

We vary the dimension p in $\{500, 1000\}$ and apply a range of learning procedures to the data generated from Models 2–4: random forest with 100 trees (Breiman 2001a), XGBoost with default settings from the `xgboost` R package (Chen and Guestrin 2016), a feedforward neural network with one hidden layer containing 80 neurons (Schmidhuber 2015), support vector regression with a radial basis function kernel (Smola and Schölkopf 2004), and linear regression regularized via the Least Absolute Shrinkage and Selection Operator (LASSO, Tibshirani 1996) and Smoothly Clipped Absolute Deviation (SCAD, Fan and Li 2001). For both LASSO and SCAD, the penalty parameter λ is chosen by cross-validation.

Figures 2–4 show the boxplots of the p-values for the case $p = 1000$; similar results for $p = 500$ are provided in Appendix A (Figures S.1–S.3). In the high-dimensional linear model (Model 2), the SCAD estimator exhibits superior performance—

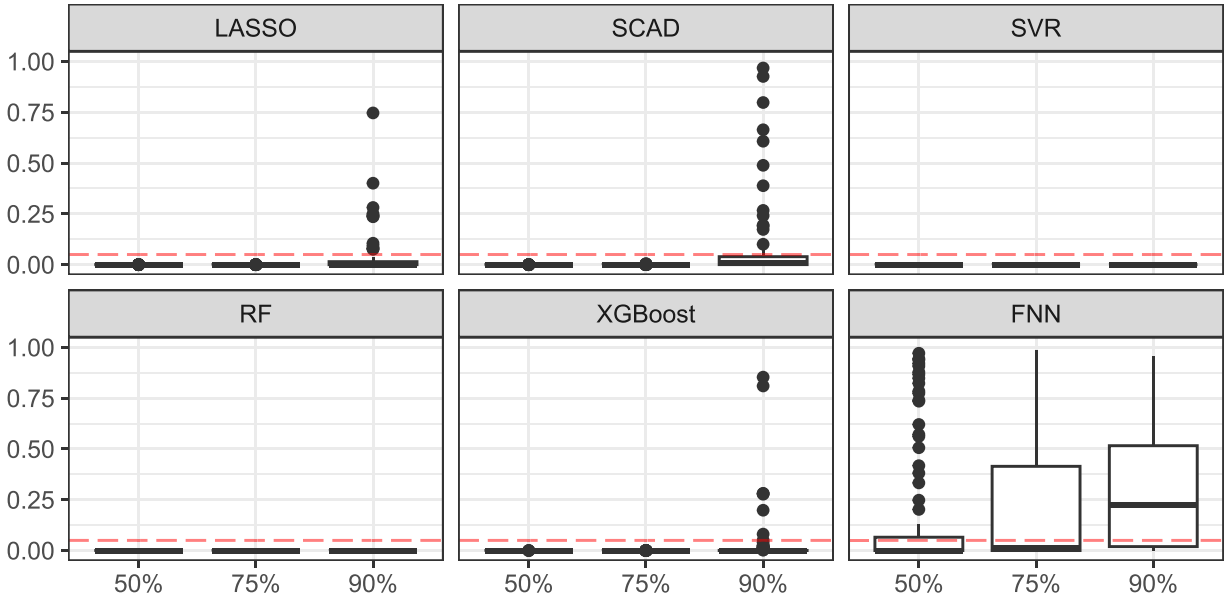


Figure 4. Boxplot of the p-value statistics from Model 4 with $p = 1000$. FNN denotes a feedforward neural network, RF represents random forest, and SVR stands for support vector regression. The sample size is $n = 500$. The horizontal axis indicates different splitting ratios, while the vertical axis displays the corresponding p-values. The red dashed line marks the significance level at 0.05. Results are based on 100 replications.

classified as *Pattern 1* in Section 3.1—owing to its enhanced variable selection accuracy. Although LASSO performs slightly worse than SCAD, its results remain satisfactory. For the general nonlinear models (Models 3 and 4), only the feedforward neural network converges toward the true distribution, conforming to either *Pattern 2* or *Pattern 3* as described in Section 3.1. The remaining methods, which fall into *Pattern 4*, display relatively inferior performance in these high-dimensional nonlinear settings. This convergence property of feedforward neural networks is partly due to their universal approximation capability; as shown by Hornik, Stinchcombe, and White (1989), such networks can approximate any Borel measurable function given a sufficient number of hidden neurons. Recent works (Bauer and Kohler 2019; Jiao et al. 2023) further demonstrate their capacity to model high-dimensional data under low-dimensional manifold assumptions.

4.3. Performance on Binary Responses

We now consider a binary outcome example in Model 5 and compare the Type-I error and power of our test with those of BAGofT (Zhang, Ding, and Yang 2023) and GRASP (Javanmard and Mehrabi 2024).

Model 5 (Binary outcome). Let $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be a p-dimensional predictor with $p \in \{100, 200\}$, and consider sample sizes $n \in \{500, 1000, 2000\}$. First, we generate an intermediate response $\zeta = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + 1/|\xi X_4 + 3| + \xi X_5^2 + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 0.1^2)$. The binary response is then defined as $Y = I\{\zeta > \text{median}(\zeta)\}$, where median$(\zeta)$ denotes the median of ζ.

In Model 5, we assess binary responses and compare our proposed test with two state-of-the-art goodness-of-fit procedures designed for binary learning settings: BAGofT (Zhang, Ding, and Yang 2023) and GRASP (Javanmard and Mehrabi 2024). For Type-I error and power evaluations, a fixed data-splitting ratio is maintained: 95% for $n = 500$, 96% for $n = 1000$, and 97% for $n = 2000$. Both our test and the BAGofT test use a single split per replication, with the original p-values reported. The BAGofT test is implemented with the default settings of its R package `BAGofT`, while the GRASP test is executed according to Algorithm 2 in Javanmard and Mehrabi (2024) with parameters $\tau = 0$ and $L = 10$.

When $\xi = 0$, the linear support vector machine (SVM) is correctly specified, corresponding to the null hypothesis H_0. Conversely, for $\xi \neq 0$, the linear SVM is misspecified, representing the alternative hypothesis H_1. We fit the model using a linear SVM while varying ξ, and the results are summarized in Figure 5, with Figure 5(a) showing the case $p = 100$ and Figure 5(b) corresponding to $p = 200$.

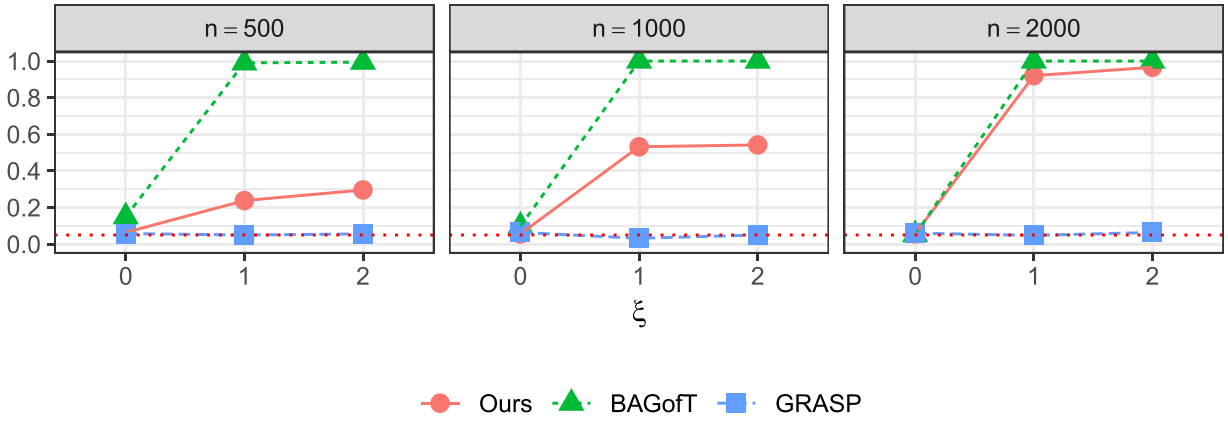
Figure 5 demonstrates that our method maintains Type-I error control when $\xi = 0$ and exhibits increasing power as ξ deviates from zero (e.g., $\xi = 1, 2$). In contrast, the GRASP test shows limited power under Model 5. Although the BAGofT test is more powerful than our approach, it fails to control the Type-I error adequately, particularly for small sample sizes (e.g., $n = 500$) or high-dimensional settings (e.g., $p = 200$).

5. Real Data Examples

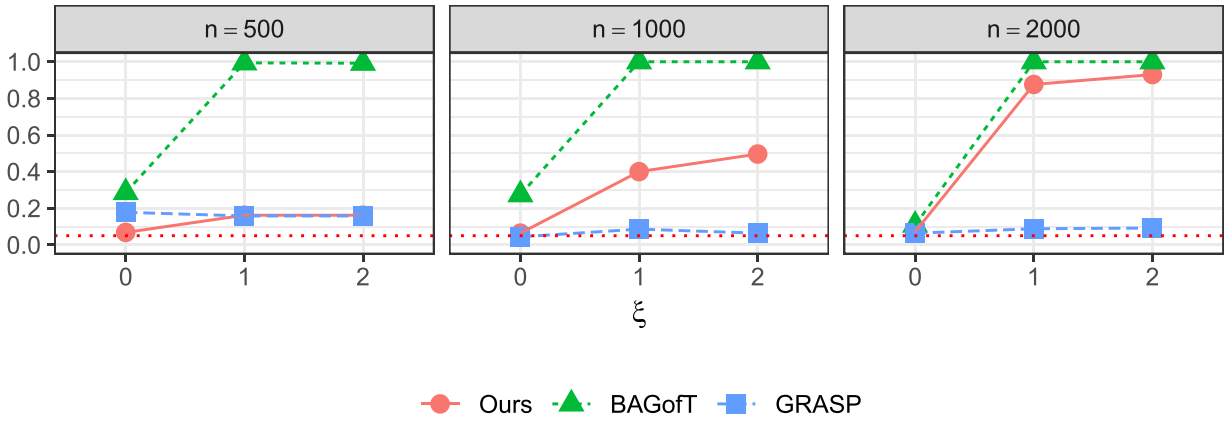
We illustrate the practical applicability of our methodology through three real-world datasets. Sections 5.1–5.2 focus on regression problems, while Section 5.3 examines a classification task.

5.1. Cortisol Stress Reactivity Dataset

The cortisol stress reactivity dataset is a well-established high-dimensional dataset widely used in mediation analysis to examine the role of DNA methylation in mediating the relationship between childhood trauma and cortisol stress reactivity (Houtepen et al. 2016; van Kesteren and Oberski



(a) $p = 100$



(b) $p = 200$

Figure 5. Empirical Type-I error and power under Model 5. The horizontal axis denotes signal strength ξ, and the vertical axis shows the rejection rate. The red dotted line indicates the significance level of 0.05. Results are based on 500 replications.

2019; Guo et al. 2022; He, He, and Xu 2025). Beyond mediation analysis, the dataset serves as a valuable resource for general regression studies. It consists of $n = 85$ samples, encompassing 385,882 DNA methylation loci, childhood trauma status, cortisol stress reactivity, and six immune cell proportions (B cells, CD4 T cells, CD8 T cells, monocytes, granulocytes, and natural killer cells), alongside confounding variables such as age and sex. The dataset is publicly available at *https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-GEOD-77445*.

In our analysis, we investigate the association between DNA methylation loci and the CD8 T cell proportion, a key immune component. Given the ultrahigh dimensionality of the predictors, we first apply a marginal screening procedure based on Pearson correlation (Fan and Lv 2008) to reduce the number of features to $p = 1000$. We then apply our test to assess the performance of the Average of the response, LASSO, SCAD, SVR, RF, XGBoost, and FNN in predicting CD8 T cell proportions.

Due to the limited sample size, we employ a fixed data split, allocating $n_1 = 70$ samples for training and $n_2 = 15$ for testing. Table 2 reports the p-values from our test and the mean squared errors (MSEs) obtained over 500 replications.

Table 2. Test p-values and mean squared errors (MSEs) for various learning procedures.

	Average	LASSO	SCAD	SVR	RF	XGBoost	FNN
p-value	0.010	0.108	0.247	0.002	0.000	0.001	0.001
MSE (Training)	1.113	0.369	0.390	0.011	0.227	0.000	2.011
MSE (Test)	1.026	0.731	0.728	1.121	0.967	1.036	2.065

NOTE: Training and test set sizes are fixed at $n_1 = 70$ and $n_2 = 15$, respectively. MSEs based on 500 replications are reported as 1000 times their actual values.

The results indicate that a linear model is a reasonable assumption for the relationship between DNA methylation and CD8 T cell proportion, as LASSO and SCAD are the only methods yielding p-values above 0.05. Furthermore, the MSEs over 500 replications reinforce the validity of our test, with the test set MSEs for LASSO and SCAD outperforming other approaches.

5.2. Wine Quality Dataset

The wine quality dataset, first introduced by Cortez et al. (2009), has been extensively analyzed using multiple regression, neural

Table 3. Test results and mean squared errors (MSEs) for the wine quality dataset.

	Splitting ratio	Average	LASSO	SCAD	SVR	RF	XGBoost	FNN
p-value	50%	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	75%	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	90%	0.000	0.000	0.000	0.000	0.000	**0.763**	0.000
MSE (Training)	50%	0.783	0.563	0.559	0.008	0.172	0.069	0.134
	75%	0.784	0.568	0.557	0.008	0.171	0.116	0.136
	90%	0.785	0.561	0.557	0.008	0.171	0.137	0.118
MSE (Test)	50%	0.785	0.584	0.574	0.786	0.765	0.591	0.891
	75%	0.785	0.579	0.573	0.786	0.762	0.559	0.878
	90%	0.779	0.571	0.570	0.782	0.757	**0.542**	0.847

NOTE: *p*-values are obtained from $m = 2$ multiple splitting procedures, and MSEs are computed over 100 replications. The bold values indicate the only scenario that passes the test, along with its corresponding MSE on the test set.

networks, and support vector regression, with SVR identified as the best-performing model in terms of mean absolute deviation and accuracy. The dataset is available at *https://archive.ics.uci.edu/dataset/186/wine+quality*.

The response variable is the quality rating of white wine, while predictors include fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, density, and pH. The dataset consists of $n = 4898$ observations. To align with our method's suitability for high-dimensional settings, we augment the dataset by creating noise predictors. Specifically, we add 500 independent standard Gaussian noise variables as additional predictors. The practice of augmenting data by adding noise draws inspiration from computer vision and natural language processing, where techniques like adding noise to images and synonym replacement are commonly employed (Feng et al. 2021; Abayomi-Alli et al. 2022).

We evaluate the performance of the Average of the response, LASSO, SCAD, SVR, RF, XGBoost, and FNN under different data splitting ratios, as detailed in Section 3.1. Table 3 presents the corresponding *p*-values and MSEs, where *p*-values are obtained using $m = 2$ multiple splitting procedures, and MSEs are computed over 100 replications.

The test results indicate that, after data augmentation, only XGBoost remains a viable model for this task, corresponding to *Pattern 3*, while all other methods fall under *Pattern 4*. The MSE results further confirm this conclusion.

5.3. MNIST Dataset

To illustrate the efficacy of our method on binary data, we consider a classification task using the MNIST dataset, which comprises a vast collection of handwritten digits (0–9). For this analysis, we focus on two visually similar digits—"4" and "9"—and randomly select a subset of $n = 1000$ images. We then apply several classification algorithms, namely XGBoost (with default settings from the R package `xgboost`) and five distinct configurations of feedforward neural networks (FNNs): (i) FNN-1: An FNN with one hidden layer containing a single neuron and employing the ReLU activation function; (ii) FNN-2: An FNN with one hidden layer containing a single neuron and using the sigmoid activation function; (iii) FNN-3: An FNN with two hidden layers consisting of 64 and 16 neurons, respectively, with ReLU activation; (iv) FNN-4: An FNN with three hidden layers with 128, 64, and 16 neurons, respectively, using the ReLU activation function; (v) FNN-5: An FNN with three hidden layers with 128, 64, and 16 neurons, respectively, using the sigmoid activation function.

Table 4. Test results and prediction accuracies for the MNIST dataset.

	Splitting ratio	XGBoost	FNN-1	FNN-2	FNN-3	FNN-4	FNN-5
p-value (Our test)	50%	0.014	0.000	0.000	0.775	0.926	0.004
	75%	0.762	0.000	0.000	0.513	0.570	0.658
	90%	0.529	0.000	0.000	0.363	0.511	0.177
p-value (BAGofT)	50%	0.353	0.000	0.000	0.000	0.783	0.003
	75%	0.206	0.000	0.000	0.001	0.266	0.001
	90%	0.340	0.000	0.003	0.000	0.319	0.010
p-value (GRASP)	50%	0.069	0.001	0.000	0.228	0.161	0.010
	75%	0.510	0.093	0.000	0.498	0.237	0.086
	90%	0.112	0.106	0.010	0.073	0.052	0.262
Accuracy (Training)	50%	1.000	0.632	0.822	0.992	0.989	0.983
	75%	1.000	0.631	0.887	0.990	0.990	0.979
	90%	1.000	0.660	0.893	0.991	0.988	0.977
Accuracy (Test)	50%	0.947	0.609	0.799	0.946	**0.949**	0.948
	75%	0.954	0.615	0.866	0.955	**0.959**	0.952
	90%	0.961	0.638	0.881	0.962	**0.964**	0.957

NOTE: Prediction accuracies are averaged over 100 replications, while *p*-values are derived from a single repetition ($m = 1$). FNN-1 through FNN-5 correspond to the five neural network architectures described in the main text. *p*-values that exceed the significance level $\alpha = 0.05$ are underlined, and the highest test-set prediction accuracy is highlighted in bold.

The results presented in Table 4 summarize the performance of the various models under different splitting ratios. For comparative purposes, we also report prediction accuracies (averaged over 100 replications) and contrast our findings with those obtained from the BAGofT (Zhang, Ding, and Yang 2023) and GRASP (Javanmard and Mehrabi 2024) goodness-of-fit tests, both specifically developed for binary classification tasks.

Notably, the FNN-4 configuration—with three hidden layers and neuron counts of 128, 64, and 16, respectively, using the ReLU activation function—attains the highest prediction accuracy on the test set, as highlighted in bold in Table 4. All three tests uniformly classify FNN-4 as *Pattern 1*, thereby failing to reject H_0 at the significance level $\alpha = 0.05$ across all considered splitting ratios.

Furthermore, XGBoost, FNN-3, and FNN-5 exhibit prediction accuracies comparable to FNN-4. Our proposed test identifies FNN-3 as *Pattern 1*, consistently not rejecting H_0 regardless of the splitting ratio. In contrast, both XGBoost and FNN-5 are categorized as *Pattern 2*, with H_0 not rejected when the splitting ratios are set at 75% and 90%.

In line with the simulation findings in Model 5, the BAGofT test demonstrates a propensity to reject H_0, classifying FNN-3 and FNN-5 as *Pattern 4* despite their high prediction accuracies, while the GRASP test tends to favor non-rejection. Specifically, GRASP categorizes FNN-1 as *Pattern 2* by rejecting it only at a splitting ratio of 50%; however, its prediction accuracy remains notably lower than that of the alternative methods.

6. Conclusion

This article introduces a novel hypothesis test to evaluate the efficacy of general learning procedures, including classical linear regression models, machine learning regression techniques, etc. The proposed testing procedure involves a data-splitting proto-

col where the training set establishes the regression model, and the test set is used to evaluate the goodness-of-fit. Simulation studies are used to demonstrate that this test can effectively identify good regression fits and exhibits a reasonable correlation with commonly used assessments, such as mean squared error. Additionally, the applications of our proposed test yield reasonable results. Our proposal is also applicable to binary data. In further studies, one can consider the goodness-of-fit for general multinomial learning procedures or generative learning procedures.

Supplementary Materials

Appendix: Proofs and additional simulation results are all relegated to the Appendix (.pdf file).

Acknowledgments

We are grateful to the Editor, Associate Editor, and two anonymous reviewers for their thoughtful and constructive comments, which have helped us significantly improve this article.

Disclosure Statement

The authors declare no conflicts of interest.

Funding

The research was supported by the National Key R&D Program of China (2023YFA1008702), the National Natural Science Foundation of China (12225113 and 12171477), and the Public Computing Cloud, Renmin University of China.

Data Availability Statement

All numerical studies were conducted by using R code. The data and R code are available at the following GitHub link *https://github.com/chenxuan-he/GoFHD2*.

References

Abayomi-Alli, O. O., Damaševičius, R., Qazi, A., Adedoyin-Olowe, M., and Misra, S. (2022), "Data Augmentation and Deep Learning Methods in Sound Classification: A Systematic Review," *Electronics*, 11, 3795. [10]

Bauer, B., and Kohler, M. (2019), "On Deep Learning as a Remedy for the Curse of Dimensionality in Nonparametric Regression," *The Annals of Statistics*, 47, 2261–2285. [3,8]

Breiman, L. (2001a), "Random Forests," *Machine Learning*, 45, 5–32. [1,7]

——— (2001b), "Statistical Modeling: The Two Cultures," *Statistical Science*, 16, 199–231. [1]

Chen, T., and Guestrin, C. (2016), "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', KDD '16*, pp. 785–794, New York: Association for Computing Machinery. [1,7]

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009), "Modeling Wine Preferences by Data Mining from Physicochemical Properties," *Decision Support Systems*, 47, 547–553. [9]

DiCiccio, C. J., DiCiccio, T. J., and Romano, J. P. (2020), "Exact Tests via Multiple Data Splitting," *Statistics & Probability Letters*, 166, 108865. [5]

Fan, J., Guo, S., and Hao, N. (2012), "Variance Estimation Using Refitted Cross-Validation in Ultrahigh Dimensional Regression," *Journal of the Royal Statistical Society*, Series B, 74, 37–65. [1]

Fan, J., and Huang, L.-S. (2001), "Goodness-of-Fit Tests for Parametric Regression Models," *Journal of the American Statistical Association*, 96, 640–652. [1,5]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [1,7]

Fan, J., Li, R., Zhang, C.-H., and Zou, H. (2020), *Statistical Foundations of Data Science* (1st ed.), Boca Raton: Chapman and Hall/CRC. [1]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of the Royal Statistical Society*, Series B, 70, 849–911. [9]

Fan, J., and Zhou, W.-X. (2016), "Guarding Against Spurious Discoveries in High Dimensions," *Journal of Machine Learning Research*, 17, 1–34. [1]

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., and Hovy, E. (2021), "A Survey of Data Augmentation Approaches for NLP," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, eds. C. Zong, F. Xia, W. Li and R. Navigli, pp. 968–988, Association for Computational Linguistics, online. [10]

Friedman, J. H. (2001), "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, 29, 1189–1232. [1]

Gao, W., and Zhou, Z.-H. (2020), "Towards Convergence Rate Analysis of Random Forests for Classification," in *Advances in Neural Information Processing Systems* (Vol. 33), Curran Associates, Inc., pp. 9300–9311. [3]

Girosi, F., Jones, M., and Poggio, T. (1995), "Regularization Theory and Neural Networks Architectures," *Neural Computation*, 7, 219–269. [1]

González-Manteiga, W., and Crujeiras, R. M. (2013), "An Updated Review of Goodness-of-Fit Tests for Regression Models," *TEST*, 22, 361–411. [2]

Guo, X., Li, R., Liu, J., and Zeng, M. (2022), "High-Dimensional Mediation Analysis for Selecting DNA Methylation loci Mediating Childhood Trauma and Cortisol Stress Reactivity," *Journal of the American Statistical Association*, 117, 1110–1121. [9]

Hardle, W., Janssen, P., and Serfling, R. (1988), "Strong Uniform Consistency Rates for Estimators of Conditional Functionals," *The Annals of Statistics*, 16, 1428–1449. [3]

He, C., He, Y., and Xu, W. (2025), "A Dual-Penalized Approach to Hypothesis Testing in High-Dimensional Linear Mediation Models," *Computational Statistics & Data Analysis*, 202, 108064. [9]

Hornik, K., Stinchcombe, M., and White, H. (1989), "Multilayer Feedforward Networks are Universal Approximators," *Neural Networks*, 2, 359–366. [8]

Hosmer, D. W., and Lemesbow, S. (1980), "Goodness of Fit Tests for the Multiple Logistic Regression Model," *Communications in Statistics - Theory and Methods*, 9, 1043–1069. [1]

Houtepen, L. C., Vinkers, C. H., Carrillo-Roa, T., Hiemstra, M., van Lier, P. A., Meeus, W., Branje, S., Heim, C. M., Nemeroff, C. B., Mill, J., Schalkwyk, L. C., Creyghton, M. P., Kahn, R. S., Joëls, M., Binder, E. B., and Boks, M. P. M. (2016), "Genome-Wide DNA Methylation Levels and Altered Cortisol Stress Reactivity Following Childhood Trauma in Humans," *Nature Communications*, 7, 10967. [8]

Imaizumi, M. (2023), "Sup-Norm Convergence of Deep Neural Network Estimator for Nonparametric Regression by Adversarial Training." arXiv:2307.04042. [3]

Janková, J., Shah, R. D., Bühlmann, P., and Samworth, R. J. (2020), "Goodness-of-Fit Testing in High Dimensional Generalized Linear Models," *Journal of the Royal Statistical Society*, Series B, 82, 773–795. [1,2]

Javanmard, A., and Mehrabi, M. (2024), "GRASP: A Goodness-of-Fit Test for Classification Learning," *Journal of the Royal Statistical Society*, Series B, 86, 215–245. [2,8,10]

Jiao, Y., Shen, G., Lin, Y., and Huang, J. (2023), "Deep Nonparametric Regression on Approximate Manifolds: Nonasymptotic Error Bounds with Polynomial Prefactors," *The Annals of Statistics*, 51, 691–716. [3,8]

le Cessie, S., and van Houwelingen, J. C. (1991), "A Goodness-of-Fit Test for Binary Regression Models, based on Smoothing Methods," *Biometrics*, 47, 1267–1282. [1]

LeCun, Y., Bengio, Y., and Hinton, G. (2015), "Deep Learning," *Nature*, 521, 436–444. [1]

Li, R., Xu, K., Zhou, Y., and Zhu, L. (2023), "Testing the Effects of High-Dimensional Covariates via Aggregating Cumulative Covariances," *Journal of the American Statistical Association*, 118, 2184–2194. [3,4]

Liu, W., Yu, X., and Li, R. (2022), "Multiple-Splitting Projection Test for High-Dimensional Mean Vectors," *Journal of Machine Learning Research*, 23, 1–27. [5]

Liu, Y., and Xie, J. (2020), "Cauchy Combination Test: A Powerful Test with Analytic p-Value Calculation Under Arbitrary Dependency Structures," *Journal of the American Statistical Association*, 115, 393–402. [5]

McCullagh, P. (1985), "On the Asymptotic Distribution of Pearson's Statistic in Linear Exponential-Family Models," *International Statistical Review / Revue Internationale de Statistique*, 53, 61–67. [1]

Peng, B., Wang, L., and Wu, Y. (2016), "An Error Bound for L_1-norm Support Vector Machine Coefficients in Ultra-High Dimension," *Journal of Machine Learning Research*, 17, 1–26. [3]

Rüschendorf, L. (1982), "Random Variables with Maximum Sums," *Advances in Applied Probability*, 14, 623–632. [5]

Schmidhuber, J. (2015), "Deep Learning in Neural Networks: An Overview," *Neural Networks*, 61, 85–117. [1,7]

Scornet, E., Biau, G., and Vert, J.-P. (2015), "Consistency of Random Forests," *The Annals of Statistics*, 43, 1716–1741. [3]

Shah, R. D., and Bühlmann, P. (2018), "Goodness-of-Fit Tests for High Dimensional Linear Models," *Journal of the Royal Statistical Society*, Series B, 80, 113–135. [1,2,5]

Shi, C., Song, R., Chen, Z., and Li, R. (2019), "Linear Hypothesis Testing for High Dimensional Generalized Linear Models," *The Annals of Statistics*, 47, 2671–2703. [3]

Smola, A. J., and Schölkopf, B. (2004), "A Tutorial on Support Vector Regression," *Statistics and Computing*, 14, 199–222. [1,7]

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014), "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, 15, 1929–1958. [1]

Stone, C. J. (1982), "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10, 1040–1053. [3,4]

Stute, W. (1997), "Nonparametric Model Checks for Regression," *The Annals of Statistics*, 25, 613–641. [1]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288. [1,3,7]

van Kesteren, E.-J., and Oberski, D. L. (2019), "Exploratory Mediation Analysis with Many Potential Mediators," *Structural Equation Modeling: A Multidisciplinary Journal*, 26, 710–723. [9]

Wang, L., Peng, B., Bradic, J., Li, R., and Wu, Y. (2020), "A Tuning-Free Robust and Efficient Approach to High-Dimensional Regression," *Journal of the American Statistical Association*, 115, 1700–1714. [3]

Zhang, J., Ding, J., and Yang, Y. (2023), "Is A Classification Procedure Good Enough?—A Goodness-of-Fit Assessment Tool for Classification Learning," *Journal of the American Statistical Association*, 118, 1115–1125. [2,3,4,5,8,10]

Zhou, T., Zhu, L., Xu, C., and Li, R. (2020), "Model-Free Forward Screening via Cumulative Divergence," *Journal of the American Statistical Association*, 115, 1393–1405. [3]