Statistica Sinica Preprint No: SS-2025-0183

Title	A Conditionally Studentized Test for High-dimensional Parametric Regression via Sample Splitting
Manuscript ID	SS-2025-0183
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0183
Complete List of Authors	Feng Liang, Chuhan Wang, Jiaqi Huang and Lixing Zhu
Corresponding Authors	Jiaqi Huang
E-mails	jhuang@mail.bnu.edu.cn
Notice: Accepted author version.	

A Conditionally Studentized Test for High-Dimensional Parametric Regression via Sample Splitting

Feng Liang[1], Chuhan Wang[1], Jiaqi Huang[2], Lixing Zhu[1]

1 Department of Statistics, Beijing Normal University at Zhuhai, China

2 School of Statistics, University of Michigan, Ann Arbor, Michigan

Abstract:

This paper introduces a Conditionally Studentized Test (COST) for model checking in general parametric regression models, addressing this challenge without relying on dimension reduction or sparsity assumptions. COST is constructed from two disjoint sample partitions linked by a weight matrix and incorporates a conditional studentization with respect to one of the subsamples. It can achieve asymptotic normality under the null hypothesis, regardless of the form of the initial test statistic (global or local smoothing-based) and irrespective of the relationship between predictor dimension, sample size, and number of parameters (fixed or diverging under certain rate constraints). Under certain conditions on the regression functions, asymptotic normality can even hold when the predictor dimension exceeds the sample size, potentially enabling the analysis of high-dimensional problems. Furthermore, COST demonstrates a fast rate of detection for local alternatives. The paper explores sample partitioning and provides numerical studies showcasing COST's finite-sample performance, including scenarios where the predictor dimension equals the sample size.

Key words and phrases: Asymptotic model-free test, conditional studentization, high dimensions, model checking, sample-splitting

1. Introduction

Regression modeling is fundamental for analyzing complex relationships, with model diagnostics crucial for ensuring estimation accuracy and inference reliability, particularly in high-dimensional settings. This is exemplified by our case study of movie box office revenue data in Section 4. Traditional diagnostics, developed for fixed dimensions, require adaptation for diverging predictor dimensions. Consider the parametric regression model:

$$Y = g(\boldsymbol{X}; \boldsymbol{\theta}) + \varepsilon, \tag{1.1}$$

where $g : \mathbb{R}^q \times \Theta \to \mathbb{R}$ denotes a known functional form with parameter vector $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$, random predictors $\boldsymbol{X} \in \mathbb{R}^q$, and error term ε satisfying $\mathbb{E}(\varepsilon|\boldsymbol{X}) = 0$. We allow q to diverge with sample size n. The core model validation task is testing: for a subset $\boldsymbol{\Theta} \subset \mathbb{R}^p$,

$$H_0 : \mathrm{P}\{E(Y|\boldsymbol{X}) = g(\boldsymbol{\theta}_0, \boldsymbol{X})\} = 1 \text{ for some } \boldsymbol{\theta}_0 \in \boldsymbol{\Theta} \tag{1.2}$$

versus

$$H_1 : \mathrm{P}\{E(Y|\boldsymbol{X}) = g(\boldsymbol{\theta}, \boldsymbol{X})\} < 1 \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}. \tag{1.3}$$

Existing approaches, categorized into local smoothing methods (e.g., Härdle and Mammen (1993); Zheng (1996); Zhu and Li (1998)) and global smoothing methods (e.g., Stute and Zhu (2002); Zhu (2003); Escanciano (2006)), face limitations in high dimensions. Local smoothing methods suffer from the curse of dimensionality, while global methods have limiting distributions dependent on unknown quantities and complex stochastic process embeddings. Recent work by Shah and Bühlmann (2018) and Janková et al. (2020) addresses high-dimensional linear and generalized linear models under fixed design matrices and sparsity structures. Dimension-reduction paradigms (e.g., Tan and Zhu (2019, 2022); Guo et al. (2016)) rely on latent low-dimensional structures, failing when these assumptions are violated, as demonstrated by Tan and Zhu (2022).

In this paper, we propose the COnditionally Studentized Test (COST), a novel test for general parametric models without dimension reduction assumptions. COST uses an incomplete two-order U-statistic of residuals $\mathcal{E}_{\mathcal{N}_1}^{\top}\mathcal{W}_{\mathcal{N}_1,\mathcal{N}_2}\mathcal{E}_{\mathcal{N}_2}/(n_1 n_2)$ as an unstudentized test statistic, where $\mathcal{N}_1 = \{(\boldsymbol{X}_i, Y_i)\}_{i=1}^{n_1}$ and $\mathcal{N}_2 = \{(\boldsymbol{X}_j, Y_j)\}_{j=n_1+1}^{n}$ are two subsmaples of sizes n_1 and n_2, $\mathcal{W}_{\mathcal{N}_1,\mathcal{N}_2}$ is an $n_1 \times n_2$ weight matrix linking two residual vectors $\mathcal{E}_{\mathcal{N}_1}$

and $\mathcal{E}_{\mathcal{N}_2}$. We then derive the conditional standard deviation to construct the final test when $\mathcal{N}_1$ is given. COST offers several advantages: it is always asymptotically distribution-free under the null hypothesis regardless of whether the initial test statistic is local smoothing or global smoothing-based, eliminating the need for resampling and simplifying inference. It also synthesizes local and global smoothing strengths, achieving a standard normal limiting null distribution and near-optimal $O(n_1^{-1/2})$ detection rates where $n_1/n \to C > 0$, ensuring high sensitivity to deviations from the null hypothesis, even in complex, high-dimensional settings. Under some regularity conditions, it handles both high-dimensional parameters ($p \approx n^{1/3}$ and predictors ($q \geq p$ and even $q \geq n$), providing flexibility for modern datasets where predictor dimensions could exceed sample size. This shows the potential to tickle very high-dimensional problems. Furthermore, COST does not rely on sparsity or dimension reduction assumptions, making it robust to unstructured high-dimensional data. It connects to dimension-agnostic inference (e.g., Kim and Ramdas (2024); Gao et al. (2023); Shekhar et al. (2022, 2023)) but features a more complex U-statistic, correlated observations, dual dimensionality, and refined theoretical tools as it operates uniformly across fixed and high-dimensional regimes.

The paper is organized as follows: Section 2 details test statistic construction; Section 3 examines asymptotic properties and optimal splitting; Section 4 presents simulations and real data analysis; Section 5 discusses advantages and limitations; Section 6 outlines regularity conditions. All technical proofs are provided in the Supplementary Material.

2. Test statistic construction

As the test construction requires estimating the parameter in the model, we first briefly provide some relevant details.

2.1 Notation and parameter estimation

Write the underlying regression function as $m(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x})$, and define the error term as $\varepsilon = Y - m(\boldsymbol{X})$ and $\varepsilon_i = Y_i - m(\boldsymbol{X}_i)$ for $i = 1, \ldots, n$. Let $\dot{g}(\boldsymbol{\theta}, \boldsymbol{X}) = \partial g(\boldsymbol{\theta}, \boldsymbol{X})/\partial \boldsymbol{\theta}$ and $\ddot{g}(\boldsymbol{\theta}, \boldsymbol{X}) = \partial \dot{g}(\boldsymbol{\theta}, \boldsymbol{X})/\partial \boldsymbol{\theta}^\top$ denote the first- and second-order derivatives of $g(\boldsymbol{\theta}, \boldsymbol{X})$ with respect to $\boldsymbol{\theta}$, respectively. The L_2 norm is denoted by $\|\cdot\|$. For notational simplicity, we write the conditional expectation $E(A \mid B = b)$ as $E_B(A)$ for any random variable or vector A and B. Set

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\min} E\left\{Y - g(\boldsymbol{\theta}, \boldsymbol{X})\right\}^2 = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\min} E\left\{m(\boldsymbol{X}) - g(\boldsymbol{\theta}, \boldsymbol{X})\right\}^2.$$

Under the null hypothesis and the Condition 1 specified in Section 6, $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$. Under the alternatives, $\boldsymbol{\theta}^*$ typically depends on the distribution of $\boldsymbol{X}$.

Define $\boldsymbol{\Sigma}_* = E\left\{\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X})\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X})^\top\right\}$. Under the null hypothesis, $\boldsymbol{\Sigma}_*$ is reduced to $\boldsymbol{\Sigma} = E\left\{\dot{g}(\boldsymbol{\theta}_0, \boldsymbol{X})\dot{g}(\boldsymbol{\theta}_0, \boldsymbol{X})^\top\right\}$.

To estimate $\boldsymbol{\theta}_0$, we define the least squares estimator

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{n} \left\{Y_i - g(\boldsymbol{\theta}, \boldsymbol{X}_i)\right\}^2. \tag{2.4}$$

Under regularity conditions 1–5 in Section 6, $\hat{\boldsymbol{\theta}}$ admits a convergence rate and an asymptotically linear representation. These properties are fundamental for establishing the asymptotic behavior of the proposed test under both the null and alternative hypotheses. The corresponding results are provided in three lemmas in the Supplementary Material: the first two parallel Theorems 1 and 2 in Tan and Zhu (2022), and the third extends Theorem 4 therein.

2.2 Motivation for the test statistic

Let $e = Y - g(\boldsymbol{\theta}_0, \boldsymbol{X})$ and $e_i = Y_i - g(\boldsymbol{\theta}_0, \boldsymbol{X}_i)$. Under the null hypothesis, $m(\boldsymbol{X}) = g(\boldsymbol{\theta}_0, \boldsymbol{X})$, so that $e = \varepsilon$. Taking advantage of the idea of nonparametric estimation–based tests (see, e.g., Zheng (1996)), note that $E(e|\boldsymbol{X}) = 0$ under the null hypothesis. Consequently, $E\{eE(e|\boldsymbol{X})f(\boldsymbol{X})\} = E\{E^2(e|\boldsymbol{X})f(\boldsymbol{X})\} = 0$, and this expectation becomes positive under alternatives, where $f(\cdot)$ denotes the density of $\boldsymbol{X}$.

To approximate $E(e|\boldsymbol{X})f(\boldsymbol{X})$, one may consider $E_{\boldsymbol{X}}\{e_1 W_n(\boldsymbol{X}, \boldsymbol{X}_1)\}$

with a weight function $W_n(\cdot,\cdot)$. For instance, we can take $W_n(\mathbf{X},\mathbf{X}_1)$ as $K_h(\boldsymbol{X}-\boldsymbol{X}_1)$, where $K_h(\cdot)=K(\cdot/h)/h^q$ denotes the kernel density function with bandwidth h. This leads to the approximation $E\{eE(e|\boldsymbol{X})f(\boldsymbol{X})\}\approx E\{e_1eW_n(\boldsymbol{X},\boldsymbol{X}_1)\}$, which, for notational convenience, can be written as $E\{e_1e_2W_n(\boldsymbol{X}_1,\boldsymbol{X}_2)\}$. Several other forms of $W_n(\cdot,\cdot)$ have been proposed, such as $\exp(-\|\boldsymbol{X}_1-\boldsymbol{X}_2\|^2/2)$ in Bierens (1982), or $1/\sqrt{\|\boldsymbol{X}_1-\boldsymbol{X}_2\|^2+1}$ in Li et al. (2019). More generally, Escanciano (2009) discussed a wide class of weight functions. Let $w_{ij}=W_n(\boldsymbol{X}_i,\boldsymbol{X}_j)$, which may depend on $\boldsymbol{X}_i$, $\boldsymbol{X}_j$, and n. A common practice is to construct U-statistics based on the empirical version of such expectations. However, these statistics often lack tractable limiting null distributions (see, e.g., Li et al. (2019)) and suffer severely from the curse of dimensionality (see, e.g., Tan and Zhu (2022)).

To overcome these issues, we motivate our test construction by the following idea. Split the sample into two disjoint parts: $\mathcal{N}_1=\{(\boldsymbol{X}_i,Y_i)\}_{i=1}^{n_1}$ and $\mathcal{N}_2=\{(\boldsymbol{X}_j,Y_j)\}_{j=n_1+1}^{n}$ with $n=n_1+n_2$. Define the incomplete U-type average:

$$U_n=\frac{1}{\sqrt{n_1n_2}}\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n}e_ie_jw_{ij}, \tag{2.5}$$

It is easy to verify that $E(U_n)/\sqrt{n_1n_2}=E\{e_1e_2W_n(\boldsymbol{X}_1,\boldsymbol{X}_2)\}$. In this formulation, the e_i's from the two parts are used without duplication, while the weight w_{ij} connects them. Importantly, conditional on $\mathcal{N}_2$, the terms

$\sum_{j=n_1+1}^{n} e_i e_j w_{ij}/\sqrt{n_1 n_2}, i = 1, \cdots, n_1$ are independent, forming the basis of our test construction.

To estimate the residuals e_i's by using plug-in estimator of $\boldsymbol{\theta}_0$ such that the two estimated sums keep "independence", we define the least squares estimators of $\boldsymbol{\theta}$ as

$$\hat{\boldsymbol{\theta}}_1 = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{n_1} \{Y_i - g(\boldsymbol{\theta}, \boldsymbol{X}_i)\}^2,$$
$$\hat{\boldsymbol{\theta}}_2 = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{j=n_1+1}^{n} \{Y_j - g(\boldsymbol{\theta}, \boldsymbol{X}_j)\}^2,$$

where $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ have the same asymptotic results as $\hat{\boldsymbol{\theta}}$ defined in (2.4) when n is replaced by n_1 and n_2 respectively. Write $\hat{e}_i = Y_i - g(\hat{\boldsymbol{\theta}}_1, \boldsymbol{X}_i),\ i = 1, \ldots, n_1$, and $\hat{e}_j = Y_j - g(\hat{\boldsymbol{\theta}}_2, \boldsymbol{X}_j),\ j = n_1+1, \ldots, n$. U_n is estimated by

$$\hat{U}_n = \frac{1}{\sqrt{n_1 n_2}} \sum_{i=1}^{n_1} \sum_{j=n_1+1}^{n} \hat{e}_i \hat{e}_j w_{ij}. \tag{2.6}$$

After plugging in the residuals, due to the correlation between $\hat{e}_i, i = 1, \cdots, n_1$, $\hat{U}_n$ is no longer a sum of independent variables given $\mathcal{N}_2$. This correlation makes it challenging to derive its limiting null distribution. To address this, we next investigate the asymptotic decomposition of $\hat{U}_n$ under the null hypothesis to motivate the construction.

By decomposing the residuals $\hat{e}_i$'s in the first part of the samples $\mathcal{N}_1$, the Corollary 1 in Supplementary Material yields that:

$$\begin{aligned}\hat{U}_n =& \frac{1}{\sqrt{n_1 n_2}} \sum_{i=1}^{n_1} \sum_{j=n_1+1}^{n} \hat{e}_i \hat{e}_j w_{ij} \\ =& \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} e_i \frac{1}{\sqrt{n_2}} \sum_{j=n_1+1}^{n} \hat{e}_j \left[w_{ij} - \dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_i)^\top \boldsymbol{\Sigma}_*^{-1} E_{\boldsymbol{X}_j} \left\{ \dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_1) w_{1j} \right\} \right] \\ &+ o_p(1) \\ =& \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} e_i \tilde{w}_i(\mathcal{N}_2) + o_p(1). \end{aligned} \tag{2.7}$$

Given $\mathcal{N}_2$, the random variables $e_1 \tilde{w}_1(\mathcal{N}_2), \ldots, e_{n_1} \tilde{w}_{n_1}(\mathcal{N}_2)$ are conditionally independent and identically distributed. Intuitively, under the null hypothesis, V_n, which is a conditionally studentized version of $\sum_{i=1}^{n_1} e_i \tilde{w}_i(\mathcal{N}_2)/\sqrt{n_1}$, converges weakly to a normal distribution by the conditional Central Limit Theorem.

$$V_n = \frac{\frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} e_i \tilde{w}_i(\mathcal{N}_2)}{\sqrt{\frac{1}{n_1} \sum_{i=1}^{n_1} \left\{ e_i \tilde{w}_i(\mathcal{N}_2) - \frac{1}{n_1} \sum_{i=1}^{n_1} e_i \tilde{w}_i(\mathcal{N}_2) \right\}^2}}. \tag{2.8}$$

Based on this result, we can construct a feasible version by replacing the unknown quantities in V_n. Specifically, we replace the numerator of V_n with $\sum_{i=1}^{n_1} \sum_{j=n_1+1}^{n} \hat{e}_i \hat{e}_j w_{ij} / \sqrt{n_1 n_2}$, and substitute $\hat{e}_i$, $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\Sigma}}$ and $\sum_{i=1}^{n_1} \dot{g}(\hat{\boldsymbol{\theta}}, \boldsymbol{X}_i) w_{ij} / n_1$ for e_i, $\boldsymbol{\theta}^*$, $\boldsymbol{\Sigma}_*$ and $E_{\boldsymbol{X}_j}\{\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_1) w_{1j}\}$ in the denominator. Therefore, the COnditionally Studentized Test (COST) statistic is defined as:

$$COST_n = \frac{\frac{1}{\sqrt{n_1 n_2}} \sum_{i=1}^{n_1} \sum_{j=n_1+1}^{n} \hat{e}_i \hat{e}_j w_{ij}}{\sqrt{\frac{1}{n_1} \sum_{i=1}^{n_1} \left\{ \hat{e}_i \tilde{\tilde{w}}_i(\mathcal{N}_2) - \frac{1}{n_1} \sum_{i=1}^{n_1} \hat{e}_i \tilde{\tilde{w}}_i(\mathcal{N}_2) \right\}^2}}, \tag{2.9}$$

where

$$\tilde{\tilde{w}}_i(\mathcal{N}_2) = \frac{1}{\sqrt{n_2}} \sum_{j=n_1+1}^{n} \hat{e}_j \left[w_{ij} - \dot{g}\left(\hat{\boldsymbol{\theta}}, \boldsymbol{X}_i\right)^\top \hat{\boldsymbol{\Sigma}}^{-1} \frac{1}{n_1} \sum_{i=1}^{n_1} \left\{ \dot{g}\left(\hat{\boldsymbol{\theta}}, \boldsymbol{X}_i\right) w_{ij} \right\} \right],$$

and $\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{n} \dot{g}(\hat{\boldsymbol{\theta}}, \boldsymbol{X}_i) \dot{g}(\hat{\boldsymbol{\theta}}, \boldsymbol{X}_i)^\top / n$. We can prove $\hat{\boldsymbol{\theta}}$, $\sum_{i=1}^{n_1} \dot{g}(\hat{\boldsymbol{\theta}}, \boldsymbol{X}_i) w_{ij} / n_1$ and $\hat{\boldsymbol{\Sigma}}$ are consistent estimators of $\boldsymbol{\theta}^*$, $E_{\boldsymbol{X}_j}\{\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_1) w_{1j}\}$ and $\boldsymbol{\Sigma}_*$ respectively, which ensures consistency of the estimated conditional standard deviation. Note that in $\tilde{\tilde{w}}_i(\mathcal{N}_2)$ we use the full-data estimator $\hat{\boldsymbol{\theta}}$ rather than $\hat{\boldsymbol{\theta}}_2$. Asymptotically, both choices yield the same limit, but using $\hat{\boldsymbol{\theta}}$ provides a faster convergence rate to $\boldsymbol{\theta}^*$.

We can see that although the original U-statistic, and even $\hat{U}_n$, do not have a tractable limiting null distribution, COST has normal weak limit under the null hypothesis. This constitutes a key theoretical advantage.

Remark 1. Regarding the weight function w_{ij}, we note that in local smoothing tests, it is typically linked to nonparametric estimation, such as kernel smoothing. When a kernel function $K_h(\boldsymbol{X}_i - \boldsymbol{X}_j)$ is used as the weight $W_n(\boldsymbol{X}_i, \boldsymbol{X}_j)$, its magnitude may diverge as the sample size increases. However, this does not affect our construction because the proposed studentized statistic is scale-invariant—the factor h^q in w_{ij} cancels out between the numerator and denominator. Therefore, we can safely consider the weight function without this scaling factor.

3. Asymptotic properties

3.1 The limiting null distribution

To derive the limiting null distribution of $COST_n$, we first establish that V_n converges weakly to a normal distribution under certain regularity conditions. We then demonstrate the asymptotic equivalence between the numerator (and denominator) of $COST_n$ and those of V_n. Therefore, $COST_n$ has normal limiting null distribution. The result is summarized as follows.

Theorem 1. *Suppose that Conditions 1-9 in Section 6 hold. Under the null hypothesis in* (1.2), *if* $p^3 \log n_1/n_1 \to 0$ *and* $p^3 \log n_2/n_2 \to 0$,

$$COST_n \xrightarrow{d} \mathbf{N}(0,1), \tag{3.10}$$

where " $\xrightarrow{d}$ *" stands for convergence in distribution.*

Remark 2. In Theorem 1, we do not impose any specific constraint on the relationship between n_1 and n_2. Nevertheless, the power of the proposed test depends on the relative sizes of n_1 and n_2. The following Theorem 2 establishes the corresponding theoretical results.

Remark 3. In Theorem 1, we require the conditions $p^3 \log n_1/n_1 \to 0$ and $p^3 \log n_2/n_2 \to 0$. These ensure the validity of the asymptotically linear representation of the parameter estimator and the L_2-consistency of the sample covariance matrices. Note that we impose some restrictions on the divergence rate of the parameter dimension p but not on the predictor dimension q. In fact, q is allowed to diverge to infinity at a rate much faster than that of p, or even faster than the sample size n. Consider a regression model in which each underlying parameter appears only once within a nested structure of nonlinear transformations. Specifically, for each parameter θ_l, there are p associated predictor variables $X_{l1}, X_{l2}, \ldots, X_{lp}$, which are successively combined through quadratic operations and additions. That is,

$$\begin{aligned} Y =& (\cdots((\theta_1 + X_{11})^2 + X_{12})^2 + \cdots + X_{1p})^2 + \cdots \\ &+ (\cdots(\theta_p + X_{p1})^2 + X_{p2})^2 + \cdots + X_{pp})^2 + \varepsilon. \end{aligned}$$

As a result, although only p parameters are present, the total number of predictors is $q = p \times p = p^2$. More generally, the model can be extended to settings with $q = p^k$ or $q = pn$, depending on the complexity of transformations. This illustrates a regime where $q \gg p$ and q can grow arbitrarily faster than n.

3.2 Power study

To investigate the power performance, we consider a sequence of alternative hypotheses:

$$H_{1n} : Y = g(\boldsymbol{\theta}_0, \boldsymbol{X}) + \delta_n l(\boldsymbol{X}) + \varepsilon. \tag{3.11}$$

When δ_n is a fixed nonzero constant, it corresponds to the global alternative model $Y = m(\boldsymbol{X}) + \varepsilon$, where $m(\boldsymbol{X}) \neq g(\boldsymbol{\theta}, \boldsymbol{X})$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. When $\delta_n \to 0$ as $n \to \infty$, it represents a sequence of local alternatives. Define $\boldsymbol{\Sigma}_\varepsilon = E\left\{\varepsilon^2 \dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X})\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X})^\top\right\}$, $\boldsymbol{\Sigma}_l = E\left\{l^2(\boldsymbol{X})\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X})\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X})^\top\right\}$ and $\boldsymbol{\Sigma}^*_{cov} = E\left\{\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_1)\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_2)^\top w_{12}\right\}$. Moreover, we recall that $\varepsilon_1 = Y_1 - m(\boldsymbol{X}_1)$ and $\varepsilon_j = Y_j - m(\boldsymbol{X}_j)$ for $j = n_1 + 1, \cdots, n$. Theorem 2 establishes the asymptotic behavior of $COST_n$ under both local and global alternatives.

Theorem 2. *Suppose that Conditions 1 – 11 in Section 6 hold, and that* $p^3 \log n_1/n_1 \to 0$ *and* $p^3 \log n_2/n_2 \to 0$. *Under the alternative hypothesis in (3.11), we have:*

(a) When $\delta_n = 1$, *let* $\boldsymbol{\Sigma}_* = E\left\{\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X})\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X})^\top\right\}$, $l(\boldsymbol{X}) = m(\boldsymbol{X}) - g(\boldsymbol{\theta}^*, \boldsymbol{X})$ *and* $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_* - E\left[\{m(\boldsymbol{X}) - g(\boldsymbol{\theta}^*, \boldsymbol{X})\}\ddot{g}(\boldsymbol{\theta}^*, \boldsymbol{X})\right]$. *Then*

$$\frac{COST_n}{\sqrt{n_1}} \xrightarrow{p} \frac{E\left\{l(\boldsymbol{X}_1)l(\boldsymbol{X}_2)w_{12}\right\}}{\sqrt{V_{(0)}}}, \tag{3.12}$$

where " $\xrightarrow{p}$ *" stands for convergence in probability and*

$$V_{(0)} = E\left[\left\{\varepsilon_1^2 + l^2(\boldsymbol{X}_1)\right\} E^2_{\mathbf{X}_1}\left\{l(\boldsymbol{X}_2)w_{12}\right\}\right] - E^2\left\{l(\boldsymbol{X}_1)l(\boldsymbol{X}_2)w_{12}\right\}$$

$$
+ E\left\{l(\boldsymbol{X}_2)\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_1)w_{12}\right\}^\top \boldsymbol{\Sigma}_*^{-1}\left(\boldsymbol{\Sigma}_\varepsilon + \boldsymbol{\Sigma}_l\right)\boldsymbol{\Sigma}_*^{-1} E\left\{l(\boldsymbol{X}_2)\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_1)w_{12}\right\}
$$

$$
- 2E\left[\left\{\varepsilon_1^2 + l^2(\boldsymbol{X}_1)\right\} l(\boldsymbol{X}_2)\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_1)w_{12}\right]^\top \boldsymbol{\Sigma}_*^{-1} E\left\{l(\boldsymbol{X}_2)\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_1)w_{12}\right\}.
$$

In cases (b)-(d), let $\boldsymbol{\Sigma} = E\left\{\dot{g}(\boldsymbol{\theta}_0, \boldsymbol{X})\dot{g}(\boldsymbol{\theta}_0, \boldsymbol{X})^\top\right\}$ *and* $l(\boldsymbol{X}) = m(\boldsymbol{X}) - g(\boldsymbol{\theta}_0, \boldsymbol{X})$.

(b) When $\delta_n = 1/\sqrt{n}$, *we have*

$$
COST_n - \frac{\frac{\sqrt{n_1}}{\sqrt{n_2 n}} \sum_{j=n_1+1}^{n} \varepsilon_j H(\boldsymbol{X}_j) + \frac{\sqrt{n_1 n_2}}{n}\mu}{\sqrt{E_{\mathcal{N}_2}\left[\left\{\frac{1}{\sqrt{n_2}}\sum_{j=n_1+1}^{n} \varepsilon_1\varepsilon_j w'_{1j} + \frac{\sqrt{n_2}}{\sqrt{n}}\varepsilon_1 H(\boldsymbol{X}_1)\right\}^2\right]}} \xrightarrow{d} \mathbf{N}(0,1), \tag{3.13}
$$

where

$$
\begin{aligned}
w'_{1j} =& w_{1j} - \dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_j)^\top \boldsymbol{\Sigma}^{-1} E_{\boldsymbol{X}_1}\left\{\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_j)w_{1j}\right\} \\
& - \dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_1)^\top \boldsymbol{\Sigma}^{-1} E_{\boldsymbol{X}_j}\left\{\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_1)w_{1j}\right\} \\
& + \dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_1)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{cov}^{*}\boldsymbol{\Sigma}^{-1}\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_j), \\
\mu =& E\left\{l(\boldsymbol{X}_1)l(\boldsymbol{X}_2)w_{12}\right\} \\
& - 2E\left\{\dot{g}(\boldsymbol{\theta}_0, \boldsymbol{X}_1)l(\boldsymbol{X}_2)w_{12}\right\}^\top \boldsymbol{\Sigma}^{-1} E\left\{\dot{g}(\boldsymbol{\theta}_0, \boldsymbol{X}_1)l(\boldsymbol{X}_1)\right\} \\
& + E\left\{\dot{g}(\boldsymbol{\theta}_0, \boldsymbol{X}_1)l(\boldsymbol{X}_1)\right\}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{cov}^{*}\boldsymbol{\Sigma}^{-1} E\left\{\dot{g}(\boldsymbol{\theta}_0, \boldsymbol{X}_1)l(\boldsymbol{X}_1)\right\} \\
H(\boldsymbol{X}_1) =& E_{\boldsymbol{X}_1}\left\{l(\boldsymbol{X}_2)w_{12}\right\} - E_{\boldsymbol{X}_1}\left\{\dot{g}(\boldsymbol{\theta}_0, \boldsymbol{X}_2)w_{12}\right\}^\top \boldsymbol{\Sigma}^{-1} E\left\{\dot{g}(\boldsymbol{\theta}_0, \boldsymbol{X}_1)l(\boldsymbol{X}_1)\right\} \\
& - \dot{g}(\boldsymbol{\theta}_0, \boldsymbol{X}_1)^\top \boldsymbol{\Sigma}^{-1} E\left\{\dot{g}(\boldsymbol{\theta}_0, \boldsymbol{X}_1)l(\boldsymbol{X}_2)w_{12}\right\} \\
& + \dot{g}(\boldsymbol{\theta}_0, \boldsymbol{X}_1)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{cov}^{*}\boldsymbol{\Sigma}^{-1} E\left\{\dot{g}(\boldsymbol{\theta}_0, \boldsymbol{X}_1)l(\boldsymbol{X}_1)\right\},
\end{aligned}
$$

and $\sum_{j=n_1+1}^{n} \varepsilon_j H(\boldsymbol{X}_j)/\sqrt{n_2} \xrightarrow{d} \mathbf{N}\left(0, E\left\{\varepsilon_1^2 H^2(\boldsymbol{X}_1)\right\}\right)$. *Particularly, if* $n_1/n = o(1)$, *then* $COST_n \xrightarrow{d} \mathbf{N}(0,1)$.

(c) When $\delta_n = n^{-\alpha}$, $0 < \alpha < 1/2$, $n^{-2\alpha}p^3 \log n \to 0$ *and* $n^{-\alpha}\sqrt{n_1} \to \infty$, *the asymptotic behavior of* $COST_n$ *depends on the limiting behavior of* $\sqrt{n_2}n^{-\alpha}$ *as* $n \to \infty$. *We consider the following three cases.*

(c.1) If $\sqrt{n_2}n^{-\alpha} \to \infty$ *as* $n \to \infty$, *then*

$$\frac{n^{\alpha} COST_n}{\sqrt{n_1}} \xrightarrow{p} \frac{\mu}{\sqrt{E\left[\{\varepsilon_1 H(\boldsymbol{X}_1)\}^2\right]}},$$

where $H(\boldsymbol{X}_1)$ *has been defined in case (b).*

(c.2) If $\sqrt{n_2}n^{-\alpha} \to K$ *as* $n \to \infty$, *where* K *is a positive constant, then*

$$\frac{n^{\alpha} COST_n}{\sqrt{n_1}} - \frac{K\mu + \frac{1}{\sqrt{n_2}}\sum_{j=n_1+1}^{n} \varepsilon_j H(\boldsymbol{X}_j)}{\sqrt{E_{\mathcal{N}_2}\left[\left\{\frac{1}{\sqrt{n_2}}\sum_{j=n_1+1}^{n} \varepsilon_1\varepsilon_j w'_{1j} + K\varepsilon_1 H(\boldsymbol{X}_1)\right\}^2\right]}} \xrightarrow{p} 0,$$

where $K\mu + \frac{1}{\sqrt{n_2}}\sum_{j=n_1+1}^{n} \varepsilon_j H(\boldsymbol{X}_j) \xrightarrow{d} \mathbf{N}\left(K\mu, E\left[\{\varepsilon_1 H(\boldsymbol{X}_1)\}^2\right]\right)$ *and all quantities are as defined in case (b).*

(c.3) If $\sqrt{n_2}n^{-\alpha} \to 0$ *as* $n \to \infty$, *then*

$$\frac{n^{\alpha} COST_n}{\sqrt{n_1}} - \frac{\frac{1}{\sqrt{n_2}}\sum_{j=n_1+1}^{n} \varepsilon_j H(\boldsymbol{X}_j)}{\sqrt{E_{\mathcal{N}_2}\left[\left\{\frac{1}{\sqrt{n_2}}\sum_{j=n_1+1}^{n} \varepsilon_1\varepsilon_j w'_{1j}\right\}^2\right]}} \xrightarrow{p} 0,$$

where $\frac{1}{\sqrt{n_2}}\sum_{j=n_1+1}^{n} \varepsilon_j H(\boldsymbol{X}_j) \xrightarrow{d} \mathbf{N}\left(0, E\left[\{\varepsilon_1 H(\boldsymbol{X}_1)\}^2\right]\right)$.

(d) When $\delta_n = n^{-\alpha}$, $\alpha > 1/2$, *then* $COST_n \xrightarrow{d} \mathbf{N}(0,1)$.

Remark 4. Theorem 2 highlights three key aspects regarding the test's power. First, it demonstrates that the test can detect local alternatives at the fastest rate of $1/\sqrt{n}$. Second, in case (b) where $\delta_n = 1/\sqrt{n}$, it is evident that the optimal sample splitting strategy is $n_1 = n_2 = n/2$, which balances estimation accuracy and test efficiency. Third, for global alternatives or slower local alternatives (cases (a) and (c)), increasing n_1 clearly improves the power, which grows at the rate of $\sqrt{n_1}$; however, if n_2 is too small, the estimation of $\tilde{\tilde{w}}(\mathcal{N}_2)$ may become unreliable, potentially degrading test performance. This trade-off is confirmed by numerical studies with $n_2 = 0.5n$, $0.25n$, and $0.1n$. Consequently, we adopt $n_2 = 0.25n$ in the simulations reported below.

4. Numerical Studies

4.1 Simulations

In this section, we conduct numerical studies to evaluate the performance of COST. We first design an appropriate weight function. The weight function $1/\sqrt{\|\boldsymbol{X}_i - \boldsymbol{X}_j\|^2 + 1}$ employed by Li et al. (2019) integrates the advantages of both local and global smoothing tests. However, it exhibits a theoretical limitation in high-dimensional settings, as it converges to zero when $q \to \infty$. To address this issue, we adopt a modified weight function that incorporates not only the above form but also an additional component, $\sum_{k=1}^{q} K_h(X_{ik} - X_{jk})$. Here, the indices i and j denote the i-th and j-th observations, respectively. In our setting, X_i belongs to $\mathcal{N}_1$ and X_j belongs to $\mathcal{N}_2$, with $i = 1, 2, \cdots, n_1$ and $j = n_1+1, \cdots, n$. The summation form ensures that the weight function remains effective as q grows. Consequently, the proposed weight function is defined as

$$W_n(\boldsymbol{X}_i, \boldsymbol{X}_j) = w \times \frac{1}{\sqrt{\|\boldsymbol{X}_i - \boldsymbol{X}_j\|^2 + 1}} + (1 - w) \times \sum_{k=1}^{q} K_h(X_{ik} - X_{jk}),$$

representing a hybrid combination of two weighting schemes.

After defining the weight function, we investigate the sensitivity of the test to the choice of tuning parameters: the bandwidth h and the weight parameter w. For the bandwidth, we set $h = c \cdot n^{-0.2}$ with $c \in \{0.5, 0.75, 1, 1.25, 1.5\}$, and find that the empirical sizes and powers of the test are nearly identical across these choices, indicating that the procedure is robust to the specification of h. Similarly, we consider five representative values of w, namely $w \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. To illustrate the robustness of the test with respect to both h and w, we present results for model H_{32} in Figure 1. The left panel displays the empirical size and power for different values of c, while the right panel shows the behavior of the empirical size and power under varying w. The results indicate that both the empirical size and power remain fairly stable for different values of h and w. Without loss of generality, we fix $c = 1$ and $w = 0.5$ in the subsequent analyses.

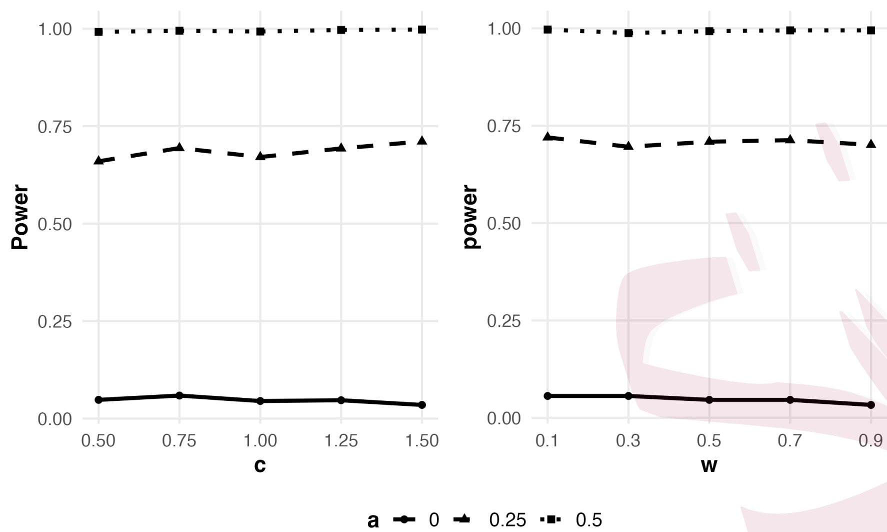

Figure 1: Empirical sizes and powers of $COST_n$ for model H_{32} in Study 3 ($n = 400$, $p = 16$, $q = 17$, $\mathbf{\Sigma} = \mathbf{\Sigma}_1$). Left panel: results for different bandwidths h with fixed weight $w = 0.5$. Right panel: results for different weight values w with fixed bandwidth $h = 1 \cdot n^{-0.2}$.

To assess the performance of our proposed test $COST_n$ relative to existing methods, we include a comparison with the $AICM_n$ in Tan and Zhu (2022), which has been shown to perform effectively in high-dimensional settings. While Janková et al. (2020) also addresses non-sparse models with random designs under our constraints, $AICM_n$ demonstrated superior performance in Tan and Zhu (2022)'s comparisons across various model settings. Consistent with these findings, our simulations confirm the effectiveness of the method in high-dimensional settings.

Furthermore, we evaluate $DrCOST_n$, a modified version of $COST_n$ that incorporates dimension reduction. If prior knowledge indicates that the model admits a dimension-reduction structure such that $Y \perp\!\!\!\perp E(Y|\boldsymbol{X}) \mid B^\top \boldsymbol{X}$, we can exploit all relevant information by replacing $\boldsymbol{X}$ with $B^\top \boldsymbol{X}$ in w_{ij}, say $w_{ij} = W_n(B^\top X_i, B^\top X_j)$. In this paper, we adopt a cumulative slicing estimation, as followed by Tan and Zhu (2022), to estimate the target matrix B.

For a comprehensive comparison, we conduct four numerical studies: models with dimension reduction, models with dimension reduction and diverging parameter dimensions, models without dimension reduction, and models without dimension reduction and diverging predictor dimensions. Studies 1 and 2 focus on multi-index model structures, representing scenarios where $AICM_n$ is expected to perform well. Study 3 investigates models without any underlying dimension-reduction structure. Finally, to demonstrate that our method can handle high-dimensional predictors q relative to the number of parameters p, Study 4 explores settings with $q = p^2$ and $q = n$. These four studies are designed to cover a wide range of scenarios, including those with and without dimension reduction, as well as varying predictor dimensionality. To assess the test's sensitivity to local deviations from the null, we vary both the drift coefficient a and the sample size n across simulations. Specifically, for a fixed n, increasing a strengthens the drift term, making the alternative easier to detect; for a fixed a, increasing n makes the effective drift larger, changing the problem from a local to a global alternative. This design allows us to examine how test power increases from local to global alternatives, as reflected in the results for different a and n values in subsequent tables.

Study 1. Generate data from the following double-index and triple-index models:

$$H_{11} : Y = \boldsymbol{\beta}_1^\top \boldsymbol{X} + a \left(\boldsymbol{\beta}_2^\top \boldsymbol{X}\right)^2 + \varepsilon,$$

$$H_{12} : Y = X_1 + \cos(2X_2) + a \exp(3X_2) + \varepsilon,$$

where $\varepsilon \sim \mathbf{N}(0, 1)$ and $\boldsymbol{X} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma})$, with covariance matrices $\boldsymbol{\Sigma}_1 = \mathbf{I}_p$ or $\boldsymbol{\Sigma}_2 = (0.5^{|i-j|})_{p\times p}$. We set $\boldsymbol{\beta}_1 = (\underbrace{1, \cdots, 1}_{q_1}, 0, \cdots, 0)^\top/\sqrt{q_1}$ and $\boldsymbol{\beta}_2 = (0, \cdots, 0, \underbrace{1, \cdots, 1}_{q_1})^\top/\sqrt{q_1}$ with $q_1 = \lfloor q/2 \rfloor$. Here X_i denotes the i-th component of $\boldsymbol{X}$. In the simulations, $a = 0$ corresponds to the null hypothesis, while $a \neq 0$ represents the alternative. All results are based on 1,000 replications and the significance level is set to 0.05. The empirical sizes and powers are reported in Tables 1-2.

Table 1 shows that $AICM_n$ performs best under model H_{11}. The relatively lower power of $COST_n$ in moderate samples can be attributed to the efficiency loss from sample splitting. In addition, since $COST_n$ determines critical values using its asymptotic null distribution rather than bootstrap resampling as in $AICM_n$, this difference may further contribute to its performance gap in smaller samples.

However, for model H_{12}, $COST_n$ outperforms both $DrCOST_n$ and $AICM_n$, even though the model structure is designed to favor the latter. A closer examination suggests that the structural dimension of the central subspace tends to be underestimated as one, which weakens the effectiveness of the dimension-reduction step and consequently lowers the power of $DrCOST_n$ and $AICM_n$ compared to $COST_n$.

Study 2. Generate data from the following multi-index models with $p = 0.1n$ and $p = 2\sqrt{n}$ respectively:

$$H_{21} : Y = \boldsymbol{\beta}_0^\top \boldsymbol{X} + a \exp\left(\boldsymbol{\beta}_0^\top \boldsymbol{X}\right) + \varepsilon,$$

$$H_{22} : Y = \boldsymbol{\beta}_1^\top \boldsymbol{X} + \exp\left(\boldsymbol{\beta}_2^\top \boldsymbol{X}\right) + a \exp\left(-\boldsymbol{\beta}_0^\top \boldsymbol{X}\right) + \varepsilon,$$

where $\boldsymbol{\beta}_0 = (1, 1, \cdots, 1)^\top / \sqrt{q}$ and other notations are the same as stated above. Study 2 focuses on scenarios where the parameter dimension p exceeds the rate of $n^{1/3}$, allowing us to examine how the tests perform when regularity conditions are violated. Specifically, models H_{21} and H_{22}

Table 1: Empirical sizes and powers for H_{11} in Study 1.

	a	$n=100$ $p=2$ $q=2$	$n=100$ $p=4$ $q=4$	$n=100$ $p=6$ $q=6$	$n=100$ $p=8$ $q=8$	$n=200$ $p=12$ $q=12$	$n=400$ $p=17$ $q=17$	$n=600$ $p=20$ $q=20$
$COST_n, \mathbf{\Sigma}_1$	0.00	0.040	0.052	0.044	0.051	0.061	0.050	0.051
	0.10	0.109	0.091	0.112	0.125	0.188	0.353	0.489
	0.25	0.408	0.385	0.385	0.406	0.726	0.956	0.993
$DrCOST_n, \mathbf{\Sigma}_1$	0.00	0.040	0.052	0.052	0.059	0.071	0.045	0.042
	0.10	0.101	0.101	0.106	0.091	0.167	0.290	0.436
	0.25	0.383	0.411	0.394	0.390	0.671	0.924	0.984
$AICM_n, \mathbf{\Sigma}_1$	0.00	0.059	0.053	0.050	0.065	0.056	0.052	0.046
	0.10	0.137	0.146	0.137	0.127	0.246	0.444	0.600
	0.25	0.552	0.563	0.572	0.556	0.870	0.993	0.999
$COST_n, \mathbf{\Sigma}_2$	0.00	0.046	0.046	0.039	0.052	0.052	0.058	0.046
	0.10	0.089	0.175	0.236	0.269	0.668	0.960	0.994
	0.25	0.427	0.678	0.811	0.840	0.988	1.000	1.000
$DrCOST_n, \mathbf{\Sigma}_2$	0.00	0.041	0.051	0.052	0.053	0.045	0.053	0.047
	0.10	0.096	0.162	0.229	0.237	0.576	0.924	0.984
	0.25	0.400	0.610	0.771	0.801	0.990	1.000	1.000
$AICM_n, \mathbf{\Sigma}_2$	0.00	0.046	0.065	0.059	0.050	0.056	0.059	0.056
	0.10	0.113	0.218	0.338	0.372	0.750	0.986	0.999
	0.25	0.528	0.800	0.917	0.947	1.000	1.000	1.000

Table 2: Empirical sizes and powers for H_{12} in Study 1.

	a	$n=100$ $p=2$ $q=2$	$n=100$ $p=2$ $q=4$	$n=100$ $p=2$ $q=6$	$n=100$ $p=2$ $q=8$	$n=200$ $p=2$ $q=12$	$n=400$ $p=2$ $q=17$	$n=600$ $p=2$ $q=20$
$COST_n, \mathbf{\Sigma}_1$	0.00	0.052	0.041	0.044	0.038	0.057	0.047	0.054
	0.05	0.611	0.576	0.567	0.557	0.800	0.913	0.913
	0.10	0.674	0.713	0.667	0.654	0.849	0.946	0.954
$DrCOST_n, \mathbf{\Sigma}_1$	0.00	0.046	0.043	0.051	0.045	0.057	0.057	0.051
	0.05	0.557	0.566	0.567	0.550	0.794	0.893	0.910
	0.10	0.569	0.639	0.621	0.647	0.811	0.921	0.943
$AICM_n, \mathbf{\Sigma}_1$	0.00	0.032	0.045	0.042	0.057	0.062	0.055	0.057
	0.05	0.152	0.146	0.164	0.156	0.187	0.274	0.402
	0.10	0.107	0.096	0.090	0.092	0.143	0.199	0.273
$COST_n, \mathbf{\Sigma}_2$	0.00	0.048	0.04	0.049	0.045	0.042	0.062	0.037
	0.05	0.600	0.581	0.582	0.580	0.806	0.885	0.935
	0.10	0.679	0.639	0.673	0.673	0.853	0.938	0.950
$DrCOST_n, \mathbf{\Sigma}_2$	0.00	0.041	0.047	0.048	0.044	0.052	0.071	0.052
	0.05	0.459	0.466	0.430	0.429	0.665	0.772	0.858
	0.10	0.494	0.451	0.442	0.453	0.655	0.816	0.830
$AICM_n, \mathbf{\Sigma}_2$	0.00	0.037	0.045	0.043	0.049	0.049	0.067	0.049
	0.05	0.122	0.122	0.137	0.113	0.135	0.110	0.112
	0.10	0.038	0.035	0.050	0.055	0.056	0.054	0.041

explore cases with $p = 0.1n$ and $p = 2\sqrt{n}$, respectively. The empirical sizes and powers are summarized in Tables 3–4.

The simulation results show that both $COST_n$ and $DrCOST_n$ remain valid even when the dimensionality assumptions are mildly violated. For model H_{21}, $AICM_n$ fails to maintain the nominal significance level, whereas our test effectively controls the size. For model H_{22}, our method exhibits slight size inflation in small samples, but achieves validity as the sample size increases. Overall, these findings demonstrate that our approach maintains reliable performance under dimensionality violations.

Table 3: Empirical sizes and powers for H_{21} in Study 2 with $p = 0.1n$.

	a	$n=50$, $p=5$, $q=5$	$n=100$, $p=10$, $q=10$	$n=500$, $p=50$, $q=50$	$n=1000$, $p=100$, $q=100$
$COST_n, \mathbf{\Sigma}_1$	0.00	0.058	0.062	0.059	0.061
	0.05	0.054	0.120	0.311	0.562
	0.10	0.121	0.245	0.822	0.956
$DrCOST_n, \mathbf{\Sigma}_1$	0.00	0.055	0.053	0.065	0.075
	0.05	0.043	0.097	0.205	0.360
	0.10	0.110	0.191	0.633	0.884
$AICM_n, \mathbf{\Sigma}_1$	0.00	0.053	0.048	0.081	0.080
	0.05	0.077	0.122	0.311	0.586
	0.10	0.162	0.271	0.865	0.996
$COST_n, \mathbf{\Sigma}_2$	0.00	0.059	0.052	0.054	0.074
	0.05	0.093	0.214	0.875	0.988
	0.10	0.219	0.524	0.997	0.999
$DrCOST_n, \mathbf{\Sigma}_2$	0.00	0.060	0.067	0.063	0.075
	0.05	0.068	0.173	0.790	0.964
	0.10	0.198	0.380	0.979	0.999
$AICM_n, \mathbf{\Sigma}_2$	0.00	0.074	0.072	0.068	0.107
	0.05	0.128	0.242	0.924	0.962
	0.10	0.260	0.561	0.925	0.956

Study 3. Generate data from the following models without dimension

Table 4: Empirical sizes and powers for H_{22} in Study 2 with $p = 2\sqrt{n}$.

	a	$n=100$ $p=20$ $q=10$	$n=400$ $p=40$ $q=20$	$n=900$ $p=60$ $q=30$
$COST_n, \mathbf{\Sigma}_1$	0.00	0.123	0.096	0.083
	0.25	0.063	0.644	0.985
	0.50	0.218	0.955	1.000
$DrCOST_n, \mathbf{\Sigma}_1$	0.00	0.133	0.085	0.096
	0.25	0.052	0.573	0.976
	0.50	0.168	0.944	1.000
$AICM_n, \mathbf{\Sigma}_1$	0.00	0.079	0.074	0.074
	0.25	0.519	0.998	1.000
	0.50	0.967	1.000	1.000
$COST_n, \mathbf{\Sigma}_2$	0.00	0.079	0.057	0.061
	0.25	0.099	0.683	0.882
	0.50	0.163	0.624	0.836
$DrCOST_n, \mathbf{\Sigma}_2$	0.00	0.085	0.071	0.055
	0.25	0.110	0.837	0.962
	0.50	0.195	0.805	0.932
$AICM_n, \mathbf{\Sigma}_2$	0.00	0.080	0.068	0.056
	0.25	0.763	0.866	0.937
	0.50	0.782	0.847	0.938

reduction structures:

$$H_{31} : Y = \sum_{i=1}^{q} \sin(\beta_{0i} X_i) + a \sum_{i=1}^{q} \text{expit}(-\beta_{0i} X_i) + \varepsilon,$$

$$H_{32} : Y = \sum_{i=1}^{q-1} X_i X_{i+1} + a \cos\left(\boldsymbol{\beta}_0^\top \boldsymbol{X}\right) + \varepsilon,$$

where β_{0i} denotes the i-th component of $\boldsymbol{\beta}_0$ and the rest of the notations are the same as stated above. This study investigates two models without dimension reduction. Tables 5–6 report that in high-dimensional settings—particularly for model H_{32} involving interaction terms—$AICM_n$ performs poorly, failing to exhibit nontrivial power as dimension grows, which highlights its reliance on dimension reduction. In contrast, $COST_n$ performs substantially better, effectively controlling the size while delivering satisfactory power.

Table 5: Empirical sizes and powers for H_{31} in Study 3.

	a	$n=100$ $p=2$ $q=2$	$n=100$ $p=4$ $q=4$	$n=100$ $p=6$ $q=6$	$n=100$ $p=8$ $q=8$	$n=200$ $p=12$ $q=12$	$n=400$ $p=17$ $q=17$	$n=600$ $p=20$ $q=20$
$COST_n, \mathbf{\Sigma}_1$	0.00	0.048	0.050	0.058	0.055	0.048	0.048	0.050
	0.05	0.058	0.123	0.205	0.369	0.906	0.999	1.000
	0.10	0.121	0.347	0.628	0.868	1.000	1.000	1.000
$AICM_n, \mathbf{\Sigma}_1$	0.00	0.046	0.043	0.009	0.001	0.000	0.000	0.000
	0.05	0.071	0.104	0.081	0.016	0.000	0.000	0.000
	0.10	0.123	0.298	0.361	0.152	0.003	0.000	0.000
$COST_n, \mathbf{\Sigma}_2$	0.00	0.039	0.040	0.056	0.073	0.060	0.054	0.046
	0.05	0.063	0.119	0.222	0.348	0.911	1.000	1.000
	0.10	0.117	0.355	0.630	0.843	1.000	1.000	1.000
$AICM_n, \mathbf{\Sigma}_2$	0.00	0.042	0.041	0.024	0.005	0.000	0.000	0.000
	0.05	0.056	0.107	0.103	0.042	0.001	0.000	0.000
	0.10	0.140	0.342	0.418	0.337	0.163	0.000	0.000

Table 6: Empirical sizes and powers for H_{32} in Study 3.

	a	$n=100$ $p=1$ $q=2$	$n=100$ $p=3$ $q=4$	$n=100$ $p=5$ $q=6$	$n=100$ $p=7$ $q=8$	$n=200$ $p=11$ $q=12$	$n=400$ $p=16$ $q=17$	$n=600$ $p=19$ $q=20$
$COST_n, \mathbf{\Sigma}_1$	0.00	0.047	0.048	0.041	0.049	0.066	0.057	0.045
	0.25	0.260	0.276	0.232	0.236	0.426	0.719	0.862
	0.50	0.753	0.711	0.696	0.677	0.917	0.996	1.000
$AICM_n, \mathbf{\Sigma}_1$	0.00	0.052	0.045	0.017	0.001	0.000	0.000	0.000
	0.25	0.357	0.258	0.098	0.012	0.000	0.000	0.000
	0.50	0.888	0.772	0.519	0.087	0.000	0.000	0.000
$COST_n, \mathbf{\Sigma}_2$	0.00	0.050	0.044	0.057	0.055	0.048	0.060	0.061
	0.25	0.259	0.200	0.178	0.168	0.206	0.327	0.363
	0.50	0.745	0.637	0.541	0.460	0.657	0.825	0.916
$AICM_n, \mathbf{\Sigma}_2$	0.00	0.054	0.050	0.021	0.003	0.000	0.000	0.000
	0.25	0.352	0.208	0.112	0.027	0.000	0.000	0.000
	0.50	0.882	0.701	0.490	0.166	0.003	0.000	0.000

Study 4. Generate data from the following models with $q = p^2$ and

$q = n$, respectively:

$$H_{41} : Y = \sum_{i=1}^{p} \left(\beta_{0i} \prod_{j=(i-1)r+1}^{\min(ir,q)} X_j \right)^{1/3} + a(\boldsymbol{\beta}_3^\top \boldsymbol{X}_{1:p})^2 + \varepsilon,$$

$$H_{42} : Y = \sum_{i=1}^{p} \sin \left(\beta_{3i} \sum_{j=(i-1)r+1}^{(i-1)r+r_1} X_j + \sum_{j=(i-1)r+r_1+1}^{\min(ir,q)} X_j \right) + a(\boldsymbol{\beta}_3^\top \boldsymbol{X}_{1:p})^2 + \varepsilon,$$

where $\boldsymbol{X}_{1:p} = (X_1, \ldots, X_p)^\top$, $\boldsymbol{\beta}_3 = (\underbrace{1, \ldots, 1}_{p_1}, 0, \ldots, 0)^\top / \sqrt{p_1}$ with $p_1 = \lfloor p/2 \rfloor$, and β_{3i} denotes the ith component of $\boldsymbol{\beta}_3$. Define $r = \lceil q/p \rceil$ and $r_1 = \lfloor r/2 \rfloor$. Other notations are consistent with previous studies.

Both models feature predictor dimension q much larger than the parameter dimension p. Tables 7–8 report the empirical sizes and powers. The results show that $COST_n$ maintains accurate size and high power even as the predictor dimension q increases substantially, indicating that our method remains valid without explicit restrictions on q. In contrast, $AICM_n$ performs poorly in these scenarios, confirming the superior robustness of $COST_n$.

4.2 A real data example

To illustrate our method, we use the CSM dataset (Ahmed et al. (2015), `https://archive.ics.uci.edu/dataset/424/csm+conventional+and+Social+media+movies+dataset+2014+and+2015`). After preprocessing, the dataset contains 187 observations with 11 predictors $\mathbf{X}$ (Rating, Genre, Budget, etc.) and Gross Income as the response. We initially examine whether the

Table 7: Empirical sizes and powers for H_{41} in Study 4.

	a	$n=100$ $p=2$	$n=100$ $p=4$	$n=100$ $p=6$	$n=100$ $p=8$	$n=200$ $p=12$	$n=400$ $p=17$	$n=600$ $p=20$
				$q=p^2$				
$COST_n, \mathbf{\Sigma}_1$	0.00	0.048	0.046	0.046	0.067	0.053	0.054	0.059
	0.10	0.097	0.121	0.141	0.131	0.213	0.390	0.562
	0.25	0.461	0.471	0.466	0.489	0.765	0.973	0.991
$AICM_n, \mathbf{\Sigma}_1$	0.00	0.040	0.000	0.000	0.000	0.000	0.000	0.000
	0.10	0.076	0.000	0.000	0.000	0.000	0.000	0.000
	0.25	0.236	0.000	0.000	0.000	0.000	0.000	0.000
$COST_n, \mathbf{\Sigma}_2$	0.00	0.048	0.056	0.050	0.067	0.054	0.082	0.043
	0.10	0.058	0.171	0.252	0.352	0.719	0.963	0.995
	0.25	0.178	0.583	0.840	0.887	0.996	1.000	1.000
$AICM_n, \mathbf{\Sigma}_2$	0.00	0.042	0.000	0.000	0.000	0.000	0.000	0.000
	0.10	0.050	0.000	0.000	0.000	0.000	0.000	0.000
	0.25	0.105	0.000	0.000	0.000	0.000	0.000	0.000
				$q=n$				
$COST_n, \mathbf{\Sigma}_1$	0.00	0.043	0.036	0.057	0.053	0.069	0.047	0.046
	0.10	0.135	0.140	0.126	0.123	0.228	0.393	0.569
	0.25	0.500	0.501	0.515	0.478	0.819	0.976	0.992
$AICM_n, \mathbf{\Sigma}_1$	0.00	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.10	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.25	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$COST_n, \mathbf{\Sigma}_2$	0.00	0.054	0.051	0.055	0.049	0.053	0.057	0.061
	0.10	0.127	0.218	0.317	0.361	0.740	0.966	0.996
	0.25	0.516	0.780	0.876	0.906	0.998	1.000	1.000
$AICM_n, \mathbf{\Sigma}_2$	0.00	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.10	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.25	0.000	0.000	0.000	0.000	0.000	0.000	0.000

dataset follows a linear regression model commonly used in practice. However, COST rejects the linear model (p-value $= 0.0412$), and the residual pattern in the Figure 2(a) further supports the conclusion that the linear specification is inadequate.

Applying cumulative slicing estimation, we identify the projection direction

$$\hat{\boldsymbol{\beta}} = (0.3186, 0.0748, 0.5351, 0.5384, 0.0733, -0.0548,$$

Table 8: Empirical sizes and powers for H_{42} in Study 4.

	a	$n=100$ $p=2$	$n=100$ $p=4$	$n=100$ $p=6$	$n=100$ $p=8$	$n=200$ $p=12$	$n=400$ $p=17$	$n=600$ $p=20$
				$q=p^2$				
$COST_n, \boldsymbol{\Sigma}_1$	0.00	0.061	0.093	0.076	0.049	0.055	0.065	0.049
	0.10	0.110	0.142	0.097	0.072	0.071	0.069	0.073
	0.25	0.437	0.348	0.199	0.119	0.156	0.198	0.225
$AICM_n, \boldsymbol{\Sigma}_1$	0.00	0.037	0.000	0.000	0.000	0.000	0.000	0.000
	0.10	0.058	0.000	0.000	0.000	0.000	0.000	0.000
	0.25	0.265	0.000	0.000	0.000	0.000	0.000	0.000
$COST_n, \boldsymbol{\Sigma}_2$	0.00	0.044	0.108	0.077	0.039	0.060	0.045	0.046
	0.10	0.103	0.217	0.144	0.101	0.137	0.187	0.245
	0.25	0.429	0.612	0.425	0.343	0.535	0.747	0.852
$AICM_n, \boldsymbol{\Sigma}_2$	0.00	0.041	0.000	0.000	0.000	0.000	0.000	0.000
	0.10	0.048	0.000	0.000	0.000	0.000	0.000	0.000
	0.25	0.285	0.000	0.000	0.000	0.000	0.000	0.000
				$q=n$				
$COST_n, \boldsymbol{\Sigma}_1$	0.00	0.063	0.045	0.058	0.055	0.057	0.058	0.056
	0.10	0.079	0.063	0.066	0.076	0.073	0.080	0.073
	0.25	0.201	0.175	0.143	0.113	0.148	0.203	0.227
$AICM_n, \boldsymbol{\Sigma}_1$	0.00	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.10	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.25	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$COST_n, \boldsymbol{\Sigma}_2$	0.00	0.066	0.035	0.054	0.054	0.041	0.047	0.048
	0.10	0.076	0.075	0.086	0.110	0.138	0.204	0.260
	0.25	0.195	0.281	0.306	0.325	0.507	0.722	0.871
$AICM_n, \boldsymbol{\Sigma}_2$	0.00	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.10	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.25	0.000	0.000	0.000	0.000	0.000	0.000	0.000

$$-0.2498, 0.2082, 0.3812, -0.0534, 0.2332)^\top.$$

Based on this, we then consider a polynomial regression model,

$$Y = \theta_1 + \theta_2 \left(\boldsymbol{\beta}^\top \boldsymbol{X}\right) + \theta_3 \left(\boldsymbol{\beta}^\top \boldsymbol{X}\right)^2 + \varepsilon, \tag{4.14}$$

which provides a substantially improved fit (p-value $= 0.4746$), indicating no evidence of model misspecification. Figure 2(b) shows that the residual patterns are markedly reduced, consistent with the findings of Tan and Zhu (2022).

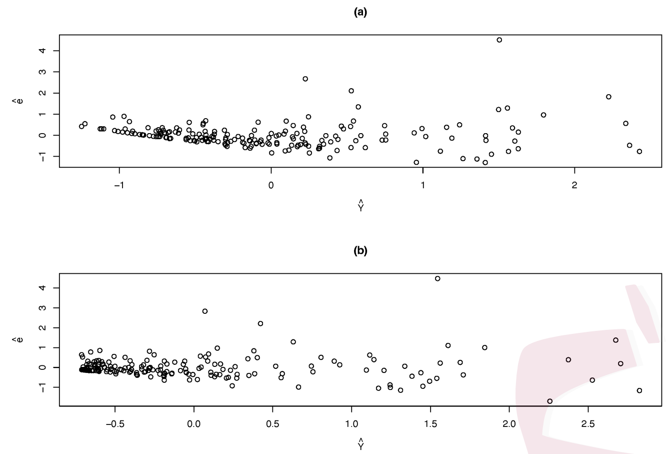

Figure 2: (a) Scatter plot of residuals generated from the linear regression model versus the fitted values, (b) Scatter plot of residuals generated from the model (4.14) versus the fitted values.

5. Discussions

This paper proposes a new test statistic for assessing general parametric regression models in high-dimensional settings. The test combines sample splitting with conditional studentization, leading to a tractable normal limiting null distribution. It is easy to implement and performs well in maintaining the nominal level and achieving good power in simulations.

A limitation of the sample-splitting approach is the inevitable loss of power since only $\mathcal{N}_1$ is used to construct statistic. Although cross-fitting is often adopted to improve power, it is not applicable here. Swapping the role of $\mathcal{N}_1$ and $\mathcal{N}_2$ would yield identical numerators but highly correlated denominators, making the joint asymptotic distribution intractable and destroying the key advantage of our method—its explicit null distribution.

Regarding dimensionality, without sparsity assumptions, the proposed test can handle a diverging parameter dimension up to order $n^{1/3}$, which is near optimal (Tan and Zhu, 2022). For higher-dimensional cases, sparsity and penalized estimation techniques are required, as in Shah and Bühlmann (2018) and Janková et al. (2020). Our simulations also show that the test remains stable when the predictor dimension q exceeds p, demonstrating its potential in ultra–high-dimensional settings.

6. Regularity Conditions

Condition 1. There exists a unique minimizer $\boldsymbol{\theta}^* \in \mathbb{R}^p$ of the squared loss function, lying in the interior of the compact parameter set $\boldsymbol{\Theta}$.

Condition 2. Let $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)$. The regression function $g(\boldsymbol{\theta}, \boldsymbol{x})$ is three times continuously differentiable with respect to $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. For any $j, k = 1, 2, \cdots, p$, define

$$\dot{g}_j(\boldsymbol{\theta}, \boldsymbol{x}) = \frac{\partial g(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \theta_j} \text{ and } \ddot{g}_{jk}(\boldsymbol{\theta}, \boldsymbol{x}) = \frac{\partial^2 g(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \theta_j \partial \theta_k}.$$

Let $U(\boldsymbol{\theta}^*)$ be a subset consisting of all $\boldsymbol{\theta}$ satisfying $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \le r_0$ in the interior of $\boldsymbol{\Theta}$, where r_0 is a positive constant. Furthermore, for all $j, k = 1, 2, \cdots, p$ and for all $\boldsymbol{\theta} \in U(\boldsymbol{\theta}^*)$, there exists a function $F(\cdot)$ such that for any $\boldsymbol{x}$, $|g(\boldsymbol{\theta}, \boldsymbol{x})| \leq F(\boldsymbol{x}), |\dot{g}_j(\boldsymbol{\theta}, \boldsymbol{x})| \leq F(\boldsymbol{x}), |\ddot{g}_{jk}(\boldsymbol{\theta}, \boldsymbol{x})| \leq F(\boldsymbol{x})$ and $E\{F^4(\boldsymbol{X})\} = O(1)$. Moreover, $E(\varepsilon^4)$ is bounded for the regression model.

Condition 3. Define $\psi_j(\boldsymbol{\theta}, \boldsymbol{x}) = \{m(\boldsymbol{x}) - g(\boldsymbol{\theta}, \boldsymbol{x})\}\dot{g}_j(\boldsymbol{\theta}, \boldsymbol{x})$ and let $P\ddot{\psi}_{j\boldsymbol{\theta}} = E\left\{\frac{\partial^2 \psi_j(\boldsymbol{\theta}, \boldsymbol{X})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top}\right\}$. Let $\lambda_i(P\ddot{\psi}_{j\boldsymbol{\theta}})$ denote the i-th eigenvalue of $P\ddot{\psi}_{j\boldsymbol{\theta}}$ for $j = 1, 2, \cdots, p$. Then $\max_{1 \leq i,j \leq p} \lambda_i(P\ddot{\psi}_{j\boldsymbol{\theta}}) \leq c$, for any $\boldsymbol{\theta} \in U(\boldsymbol{\theta}^*)$, where c is a positive constant independent of n and p.

For any matrix $\boldsymbol{A}$, let $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ denote its smallest and largest eigenvalues, respectively.

Condition 4. (1) There exist two constants a and b, independent of n and p, such that $0 < a \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq b < \infty$.

(2) Assume that a^*, b^*, a_ε, and b_ε are four constants independent of n and p such that $0 < a^* \leq \lambda_{\min}(\boldsymbol{\Sigma}_*) \leq \lambda_{\max}(\boldsymbol{\Sigma}_*) \leq b^* < \infty$ and $0 < a_\varepsilon \leq \lambda_{\min}(\boldsymbol{\Sigma}_\varepsilon) \leq \lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon) \leq b_\varepsilon < \infty$.

(3) Assume $\|\boldsymbol{\Sigma}^*_{\text{cov}}\| \leq b_{\text{cov}}$, where b_{cov} is a constant independent of n and p.

Condition 5. There exist two measurable functions $F_1(\cdot)$ and $F_2(\cdot)$ satisfying $E\{F_1^4(\boldsymbol{X})\} = O(1)$ and $E\{F_2^8(\boldsymbol{X})\} = O(1)$, such that $|\ddot{\psi}_{jkl}(\boldsymbol{\theta}_1, \boldsymbol{x}) -$

$\ddot{\psi}_{jkl}(\boldsymbol{\theta}_2, \boldsymbol{x})\big| \leq \sqrt{p}\,\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|\, F_1(\boldsymbol{x})$ and $\big|g^{(3)}_{jkl}(\boldsymbol{\theta}_1, \boldsymbol{x}) - g^{(3)}_{jkl}(\boldsymbol{\theta}_2, \boldsymbol{x})\big| \leq \sqrt{p}\,\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|\, F_2(\boldsymbol{x})$ for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in U(\boldsymbol{\theta}^*)$, where $g^{(3)}_{jkl}(\boldsymbol{\theta}, \boldsymbol{x}) = \frac{\partial^3 g(\boldsymbol{\theta}, \boldsymbol{x})}{\partial\theta_i \partial\theta_k \partial\theta_l}$.

Condition 6. There exists a measurable function $L(\cdot)$ with $E\{L^4(\boldsymbol{X})\} = O(1)$, such that $\big\|\dot{g}(\boldsymbol{\theta}_1, \boldsymbol{x}) - \dot{g}(\boldsymbol{\theta}_2, \boldsymbol{x})\big\| \leq \sqrt{p}\,\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|\, L(\boldsymbol{x})$ for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in U(\boldsymbol{\theta}^*)$.

Remark 5. Conditions 1, 2, 3, and 5 are standard assumptions in high-dimensional model checking, as discussed in Tan and Zhu (2022). Condition 4 resembles regularity condition (A2) in Tan and Zhu (2022), while we additionally require boundedness of the extreme eigenvalues for the other three matrices. Condition 6 is a general Lipschitz condition. Notably, Conditions 1–6 do not explicitly constrain the predictor dimension q; rather, the dimensionality is implicitly controlled by the boundedness of related functions and their derivatives. Consequently, although these conditions may become more stringent as q rapidly diverges to infinity, they remain applicable to certain function classes, allowing the test to perform effectively with high-dimensional predictor vectors.

Condition 7. $\inf_{\boldsymbol{x}_i, \boldsymbol{x}_j} W_n(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0$ and $E\{W_n^4(\boldsymbol{X}_1, \boldsymbol{X}_2)\} = E(w_{12}^4) \leq C_1$ for some positive constant C_1.

Remark 6. For Condition 7, the weight function w_{12} of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ can satisfy $E(w_{12}^4) \leq C_1$ in many forms, such as $w_{12} = 1/\sqrt{\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|^2/q + 1}$ or $w_{12} = \exp\{-\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|^2/(2q)\}$. In both cases, $w_{12} \leq 1$ always holds. For forms such as $w_{12} = \sum_{k=1}^{q} \exp\{-(X_{1k} - X_{2k})^2/2\}$, although $E(w_{12}^4)$ may diverge to infinity as $q \to \infty$ in some situations, we can normalize it by q and use $w_{12}^* = w_{12}/q$ instead, ensuring that $E({w_{12}^*}^4) \leq 1$.

In fact, under Conditions 4 and 7, we can infer that $\|E\{\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_1)\}\|$ and $\|E_{\boldsymbol{X}_1}\{\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_2)w_{12}\}\|$ are both bounded. Further details demonstrating that this condition is satisfied are provided in the Supplementary Material.

Condition 8. $\max_{1\leq k,l\leq p} E\{\varepsilon_1^4 \dot{g}_k^2(\boldsymbol{\theta}_0, \boldsymbol{X}_1)\dot{g}_l^2(\boldsymbol{\theta}_0, \boldsymbol{X}_1)\} < \infty$.

Condition 9. Define $w'_{ij} = w_{ij} - \dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_j)^\top \boldsymbol{\Sigma}^{-1} E_{\boldsymbol{X}_i}\{\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_j)w_{ij}\} - \dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_i)^\top \boldsymbol{\Sigma}^{-1} E_{\boldsymbol{X}_j}\{\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_i)w_{ij}\} + \dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_i)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\text{cov}}^*\boldsymbol{\Sigma}^{-1}\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_j)$. Assume that $E(\varepsilon_1^2\varepsilon_2^2 {w'_{12}}^2) \geq C_3$ and $E(\varepsilon_1^4\varepsilon_2^4 {w'_{12}}^4) \leq C_4$, where C_3 and C_4 are positive constants. Furthermore, $E\left[\left\{\dot{g}(\boldsymbol{\theta}_0, \boldsymbol{X}_1)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\text{cov}}^*\boldsymbol{\Sigma}^{-1}\dot{g}(\boldsymbol{\theta}_0, \boldsymbol{X}_2)\right\}^4\right] = O(1)$.

Remark 7. Condition 9 is a sufficient condition for the Berry–Esseen bound, which is essential to ensure the asymptotic normality of our test statistic under the null hypothesis. To analyze w'_{ij}, note that the linear projection of w_{ij} on $\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_i)$ is $\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_i)^\top \boldsymbol{\Sigma}^{-1} E_{\boldsymbol{X}_j}\{\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_i)w_{ij}\}$. The linear projection of the remainder term $w_{ij} - \dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_i)^\top \boldsymbol{\Sigma}^{-1} E_{\boldsymbol{X}_j}\{\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_i)w_{ij}\}$ on $\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_i)$ is $\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_i)^\top \boldsymbol{\Sigma}^{-1}\big[E_{\boldsymbol{X}_i}\{\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_j)w_{ij}\} - E\{\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_1)\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_2)^\top w_{12}\}$

$\boldsymbol{\Sigma}^{-1}\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_j)\big]$. Hence, w'_{ij} represents the residual component after w_{ij} is projected onto $\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_i)$ and $\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_j)$. Intuitively, w'_{ij} can be viewed as the estimation error when w_{ij} is linearly approximated by $\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_i)$ and $\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_j)$. Since the structures of w_{ij} and $\dot{g}(\boldsymbol{\theta}^*, \boldsymbol{X}_i)$ differ, it is reasonable to assume that $E(\varepsilon_1^2\varepsilon_2^2 {w'_{12}}^2)$ has a positive lower bound as $p \to \infty$. Given that w_{ij} is bounded, assuming the boundedness of $E(\varepsilon_1^4\varepsilon_2^4 {w'_{12}}^4)$ is also reasonable.

Condition 10. Under the global alternative hypothesis, assume $E\{l^4(\boldsymbol{X})\} = O(1)$. Furthermore, there exist two constants a_l and b_l, independent of n and p, such that $0 < a_l \le \lambda_{\min}(\boldsymbol{\Sigma}_l) \le \lambda_{\max}(\boldsymbol{\Sigma}_l) \le b_l < \infty$.

Condition 11. Under the local alternative hypothesis, suppose $H(\boldsymbol{X}_1) \neq 0$ almost surely, $E\{|\varepsilon_1 H(\boldsymbol{X}_1)|^2\} \ge C_5 > 0$ and $E\{|\varepsilon_1 H(\boldsymbol{X}_1)|^4\} \le C_6 < \infty$, where C_5 and C_6 are constants independent of n and p.

Remark 8. The asymptotic distribution of our test statistic under alternative hypotheses is established using Conditions 10 and 11. These conditions represent a trade-off: they facilitate the derivation of limiting properties, particularly under local alternatives, but are not strictly necessary for achieving high power. In fact, relaxing the moment or eigenvalue constraints on $l(\boldsymbol{X})$, $H(\boldsymbol{X})$, or $\boldsymbol{\Sigma}_l$ may allow the test statistic to diverge more rapidly, thereby increasing power. However, for a more tractable analysis of limiting behavior under local alternatives, Conditions 10 and 11 are particularly useful.

Supplementary Materials

The technical proofs are provided in the Supplementary Materials.

Acknowledgements

Equal contributions were made by all authors to this research. The authors are listed in order of seniority, and Jiaqi Huang is the corresponding author. The research was supported by the grants (NSFC12131006, NSFC12471276) from the National Natural Scientific Foundation of China and the grant (CI2023C063YLL) from the Scientific and Technological Innovation Project of China Academy of Chinese Medical Science.

References

Ahmed, M., M. Jahangir, H. Afzal, A. Majeed, and I. Siddiqi (2015). Using crowd-source based features from social media and conventional features to predict the movies popularity. In *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, pp. 273–278. IEEE.

Bierens, H. J. (1982). Consistent model specification tests. *Journal of Econometrics* ***20***, 105–134.

Escanciano, J. C. (2006). A consistent diagnostic test for regression models using projections. *Econometric Theory* ***22***, 1030–1051.

Escanciano, J. C. (2009). On the lack of power of omnibus specification tests. *Econometric Theory* ***25***, 162–194.

Gao, H., R. Wang, and X. Shao (2023). Dimension-agnostic change point detection. *arXiv preprint arXiv:2303.10808*.

Guo, X., T. Wang, and L. Zhu (2016). Model checking for parametric single-index models: a dimension reduction model-adaptive approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* ***78***, 1013–1035.

Härdle, W. and E. Mammen (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics* ***21***, 1926–1947.

Janková, J., R. D. Shah, P. Bühlmann, and R. J. Samworth (2020). Goodness-of-fit testing in high dimensional generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* ***82***, 773–795.

Kim, I. and A. Ramdas (2024). Dimension-agnostic inference using cross u-statistics. *Bernoulli* *30*(1), 683–711.

Li, L., S. N. Chiu, and L. Zhu (2019). Model checking for regressions: An approach bridging between local smoothing and global smoothing methods. *Computational Statistics & Data Analysis* ***138***, 64–82.

Shah, R. D. and P. Bühlmann (2018). Goodness-of-fit tests for high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* ***80***, 113–135.

Shekhar, S., I. Kim, and A. Ramdas (2022). A permutation-free kernel two-sample test. *Advances in Neural Information Processing Systems 35*, 18168–18180.

Shekhar, S., I. Kim, and A. Ramdas (2023). A permutation-free kernel independence test. *Journal of Machine Learning Research 24*(369), 1–68.

Stute, W. and L. Zhu (2002). Model checks for generalized linear models. *Scandinavian Journal of Statistics* ***29***, 535–545.

Tan, F. and L. Zhu (2019). Adaptive-to-model checking for regressions with diverging number of predictors. *The Annals of Statistics* ***47***, 1960–1994.

Tan, F. and L. Zhu (2022). Integrated conditional moment test and beyond: when the number of covariates is divergent. *Biometrika* ***109***, 103–122.

Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics* ***75***, 263–289.

Zhu, L. (2003). Model checking of dimension-reduction type for regression. *Statistica Sinica* ***13***, 283–296.

Zhu, L. and R. Li (1998). Dimension-reduction type test for linearity of a stochastic regression model. *Acta Mathematicae Applicatae Sinica* ***14***, 165–175.

Department of Statistics, Beijing Normal University at Zhuhai, Zhuhai, China

E-mail: 202331011012@mail.bnu.edu.cn, c.h.wang@mail.bnu.edu.cn, lzhu@bnu.edu.cn

School of Statistics, University of Michigan, Ann Arbor, MI, USA

E-mail: jiaqihua@umich.edu