Machine Learning with Applications 17 (2024) 100581

Contents lists available at ScienceDirect

Machine Learning with Applications

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

Navigating interpretability and alpha control in GF-KCSD testing with measurement error: A Kernel approach

Elham Afzali *, Saman Muthukumarana, Liqun Wang

Department of Statistics, University of Manitoba, Winnipeg, R3T 2N2, Manitoba, Canada

ARTICLE INFO

Keywords:
Gradient-free Kernel conditional stein discrepancy (GF-KCSD)
Bootstrap resampling
Measurement error
Goodness-of-fit testing
Brain MRI data analysis
Conditional distributions
Kernel methods

ABSTRACT

The Gradient-Free Kernel Conditional Stein Discrepancy (GF-KCSD), presented in our prior work, represents a significant advancement in goodness-of-fit testing for conditional distributions. This method offers a robust alternative to previous gradient-based techniques, specially when the gradient calculation is intractable or computationally expensive. In this study, we explore previously unexamined aspects of GF-KCSD, with a particular focus on critical values and test power—essential components for effective hypothesis testing. We also present novel investigation on the impact of measurement errors on the performance of GF-KCSD in comparison to established benchmarks, enhancing our understanding of its resilience to these errors. Through controlled experiments using synthetic data, we demonstrate GF-KCSD's superior ability to control type-I error rates and maintain high statistical power, even in the presence of measurement inaccuracies. Our empirical evaluation extends to real-world datasets, including brain MRI data. The findings confirm that GF-KCSD performs comparably to KCSD in hypothesis testing effectiveness while requiring significantly less computational time. This demonstrates GF-KCSD's capability as an efficient tool for analyzing complex data, enhancing its value for scenarios that demand rapid and robust statistical analysis.

1. Introduction

In the dynamic field of statistical and machine learning models, the rapid development of complex algorithms has outpaced the development of robust model validation and diagnostic methods. It is essential to bridge this gap to understand the strengths and limitations of complex models, enhance their interpretability, and identify areas for improvement. By assessing how well a model fits observed data, these techniques enable model validation, interpretability, and finding ways to make it better. Acknowledging that "all models are wrong, but some are useful" (Box, 1976; Box & Draper, 1987), the challenge lies not in validating the truth of a model but rather in quantifying how much it deviates from the observed data (Gelman & Shalizi, 2013). This continuous cycle of model criticism, revision, and reevaluation is a part of the data analysis process as illustrated by Blei's Box's loop (Blei, 2014).

A cornerstone in statistical model criticism is the goodness-of-fit test, with traditional tests such as the χ^2 test (Pearson, 1900), Kolmogorov–Smirnov test (Smirnov, 1939), and Anderson–Darling test (Anderson & Darling, 1952) being essential tools for practitioners to ensure their models match the real-world data. However, these traditional tests often demand a fully specified model distribution, a limitation in modern applications where distributions are frequently known only up to an intractable normalization constant. This challenge arises in scenarios such as statistical models for network data, large-scale graphical models, and deep generative models.

To address this issue, subsequent studies have developed a novel goodness-of-fit assessment strategy. This strategy applies Stein's technique (Stein, 1972), which works directly with model distributions that are non-normalized. The foundational groundwork laid by Gretton, Borgwardt, Rasch, Schölkopf, and Smola (2012) in integral probability metrics showcased the efficacy of utilizing kernel embeddings to compare distributions. Further developments, building on this, by Gorham and Mackey (2015) proposed a discrepancy measure based on Stein operators and functions in Sobolev space. In order to improve goodness-of-fit testing methods for complex and non-normalized model distributions, researchers (Chwialkowski, Strathmann, & Gretton, 2016; Liu, Lee, & Jordan, 2016) used the unit ball of a reproducing kernel Hilbert space (RKHS) as the designated test function class. After that, Jitkrittum, Kanagawa, and Schölkopf (2020) made further strides by developing the kernel conditional Stein discrepancy (KCSD) test tailored for conditional goodness-of-fit assessment. Notably, it is imperative to acknowledge that these methodologies necessitate gradient computations, a computational aspect that may introduce instability, as highlighted by Fisher, Oates, et al. (2022). They presented the gradient-free kernel

* Corresponding author.
E-mail addresses: afzalie@myumanitoba.ca (E. Afzali), Saman.Muthukumarana@umanitoba.ca (S. Muthukumarana), Liqun.Wang@umanitoba.ca (L. Wang).

https://doi.org/10.1016/j.mlwa.2024.100581
Received 25 June 2024; Received in revised form 23 July 2024; Accepted 1 August 2024
Available online 5 August 2024

Stein discrepancy (GF-KSD), a novel method for assessing goodness-of-fit that does not rely on gradients. It uses kernel-based algorithms to calculate the difference between model distributions and observed data, making it appropriate for complicated models and high-dimensional data when gradients are either unavailable or computationally costly. However, one shortcoming of the GF-KSD technique is that it focuses on analyzing marginal distributions rather than conditional distributions. Building on this foundation, Afzali and Muthukumarana (2023) further expanded the application of gradient-free methodologies to the realm of conditional distribution testing through the development of the Gradient-Free Kernel Conditional Stein Discrepancy (GF-KCSD). This advancement integrates several cutting-edge concepts, including kernel embedding of distributions, Stein operators, and the application of gradient-free discrepancies within a conditional framework. Such integration not only addresses the challenges of gradient instability but also enhances the adaptability and reliability of goodness-of-fit assessment in the context of complex modern data sets. The rest of this study rigorously investigates empirical performance of GF-KCSD.

The effect of measurement error on non-parametric test statistics critically undermines the accuracy and credibility of statistical interpretations. These methods, which avoid the assumption of a specific data distribution, are notably susceptible to errors in measurement. Such inaccuracies can cause improper data analysis, which is fundamental to analyses that rely on precise data representation. As a result, measurement error can substantially skew results, typically diminishing the perceived magnitude of effects and decreasing the ability to discern significant differences. These are important consequences that necessitate the implementation of robust approaches to mitigate these effects and ensure more reliable statistical analyses. This highlights the critical need to account for measurement error when designing and interpreting research employing non-parametric statistics (Carroll, Ruppert, Stefanski, & Crainiceanu, 2006; Delaigle & Van Keilegom, 2021; Fuller & Fuller, 1987).

In this work, we aim to (i) explore the critical values of the GF-KCSD test statistic, (ii) evaluate the statistical power of GF-KCSD against relevant benchmarks, and (iii) examine the effects of measurement error on the performance of the GF-KCSD and established benchmarks. Understanding the distribution and appropriate critical values will enable principled hypothesis testing using GF-KCSD. Comparisons with KCSD and other divergence tests will provide insight into the advantages of GF-KCSD. Furthermore, assessing the impact of measurement errors is critical for assuring the robustness of statistical inference, especially in actual applications with noisy data.

Paper Outline: This paper begins with Section 2, which lays the framework by explaining the key concepts and theoretical backgrounds required to understand the study's aims. In Section 3, we go into the methodology used to evaluate the performance of the GF-KCSD approach, including a thorough discussion of the features of measurement errors and their expected impact on GF-KCSD effectiveness. Section 4 provides experimental examples that highlights the practicality and effectiveness of the GF-KCSD approach for assessing statistical models. This section is separated into two major subsections: one for synthetic datasets and one for real-world brain MRI database. Finally, Section 5 summarizes our findings and offers a discussion of the implications for future research initiatives.

2. Foundational concepts and theoretical background

Having established the challenges of interpreting test statistics in complex models, we now turn our attention to an overview of the Gradient-Free Kernel Conditional Stein Discrepancy (GF-KCSD). We will present the theoretical background and fundamental principles, laying the groundwork for our investigation of GF-KCSD's critical values, type-I error and test power.

To grasp how GF-KCSD works, we first need to cover some background information. We begin with an exploration of Stein's method and its extensions, discussing the Stein operator and Stein identity, which are crucial concepts that allow us to apply GF-KCSD to estimating conditional probability densities. Moreover, the role of Reproducing Kernel Hilbert Space (RKHS) in kernel methods will be explored to highlight its value in embedding distributions and enabling gradient-free discrepancy calculations. Our objective is to create a strong basis for the debate in the next sections of this paper.

2.1. The stein operator

The Stein operator, originally introduced by Stein (1972), is a pivotal element in statistical analysis, particularly within the scope of the Kernel Stein Discrepancy (KSD) method and its extensions. This operator serves as a linear differential operator that maps functions from a certain function space (often a reproducing kernel Hilbert space) to the space of functions whose expectations under the target distribution are zero. Let $\mathcal{H}$ denote a Hilbert space of functions. The Stein operator $\mathcal{T}_p$ is mathematically represented as $\mathcal{T}_p : \mathcal{H} \to \mathcal{H}_0$, where $\mathcal{H}_0$ is the subspace of functions in $\mathcal{H}$ such that $\mathbb{E}_p[\mathcal{T}_p\mathbf{g}(x)] = 0$. This operator transforms a function $\mathbf{g}(x) \in \mathcal{H}$ into

$$\mathcal{T}_p\mathbf{g}(x) = \sum_{i=1}^{n} \left(g_i(x) \cdot \frac{\partial}{\partial x_i} \log p(x) + \frac{\partial}{\partial x_i} g_i(x) \right),$$

where $\frac{\partial}{\partial x_i} \log p(x)$ is the gradient of the log-density of the target distribution, and $\frac{\partial}{\partial x_i} g_i(x)$ is the gradient of the test function $\mathbf{g}(x)$ with respect to x. The Stein operator thus combines the original function $\mathbf{g}(x)$ with information from the target distribution. This allows it to capture how $\mathbf{g}(x)$ changes with respect to both its original values and the density of the target distribution, resulting in $\mathcal{T}_p\mathbf{g}(x) \in \mathcal{H}_0$.

A fundamental property of the Stein operator is Stein's identity, which states that the expectation of $\mathcal{T}_p\mathbf{g}(x)$ is zero under the target distribution. This identity is critical within the framework of kernel Stein discrepancy approaches, acting as a key property that supports their efficacy. It guarantees that the connection between the test function and the target distribution is correctly represented, giving a trustworthy basis for model evaluation:

$$\mathbb{E}_{x\sim p}[\mathcal{T}_p\mathbf{g}(x)] = 0 \tag{1}$$

Eq. (1), not only supports Stein's identity, serves as the theoretical foundation for the use of the Stein discrepancy techniques in statistical analysis.

Stein's identity is fundamentally linked to the framework of the Reproducing Kernel Hilbert Space (RKHS)—a Hilbert Space where functions are endowed with the capacity for inner product operations, ensuring closure. Within this context, we consider the class of test functions to reside in an RKHS, specifically, $g_i \in \mathcal{H}$ for all i. Here, $\mathcal{H}$ represents the RKHS wherein each scalar-valued function, g_i, is an element, as discussed in seminal works (Chwialkowski et al., 2016; Liu et al., 2016).

In the well-defined framework of a Reproducing Kernel Hilbert Space (RKHS), defined over a set $\mathcal{X}$ with a reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we delve into the analysis of function properties through RKHS's inherent operations:

1. **Inner Product:** In the context of a reproducing kernel Hilbert space (RKHS), the inner product between any two functions f and g within this space, denoted as $f, g \in \mathcal{H}$, is defined by the expected value of their product under a target probability distribution $p(x)$:

$$\langle f, g\rangle_{\mathcal{H}} = \mathbb{E}_p[f(x) \cdot g(x)] \tag{2}$$

This definition integrates statistical expectations with the geometric principles of RKHS, thereby enhancing the way interactions between functions are analyzed within this mathematical space.

2. **Norm:** The norm of a function $f \in \mathcal{H}$, reflects its magnitude and behavior within the RKHS. It is calculated as the square root of the function's self-inner product:

$$\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}} = \sqrt{\mathbb{E}_p[f(x)^2]} \tag{3}$$

where $\mathbb{E}_p$ denotes the expectation with respect to the target distribution $p(x)$. This norm serves as a critical measure, providing insights into the function's variance and its distributional characteristics across the space.

This exposition not only highlights the role of RKHS in the fusion of geometric and probabilistic principles in function analysis but also showcases its critical application in statistical inference and the theoretical underpinnings of Stein's method.

The mathematical foundation provided by the Stein Operator and the reproducing kernel Hilbert space (RKHS) ensures a rigorous and systematic approach to the analysis of functions within the context of kernel Stein discrepancy methods. Given an observed dataset $(x_i)_{i=1}^n$ with an unknown distribution p_0, and a proposed model with distribution p, we utilize a class of test functions $\mathbf{g} \in \mathcal{G}$ in RKHS to measure the Kernel Stein Discrepancy (KSD). Here, $\mathcal{G}$ represents the class of test functions within the RKHS. The KSD is defined as:

$$\begin{aligned} S_p(p_0, \mathcal{T}_p, \mathcal{G}) &= \sup_{\|\mathbf{g}\|_{\mathcal{G}} \le 1} \|\mathbb{E}_{x\sim p_0}[\mathcal{T}_p \mathbf{g}(x)] - \cancelto{0}{\mathbb{E}_{x\sim p}[\mathcal{T}_p \mathbf{g}(x)]}\| \\ &= \sup_{\|\mathbf{g}\|_{\mathcal{G}} \le 1} \|\mathbb{E}_{x\sim p_0}[\mathcal{T}_p \mathbf{g}(x)]\| \end{aligned} \tag{4}$$

The KSD is zero if and only if the two distributions, p and p_0, are identical in the sense that there is no statistical difference between them detectable by any test function in the chosen RKHS. This method allows us to systematically assess how well the proposed model p captures the true characteristics of the dataset $\{x_i\}_{i=1}^n$, which follows the distribution p_0.

2.2. Conceptual backgrounds of the GF-KCSD

The GF-KCSD framework emerges as an extension of the kernel Stein discrepancy paradigm, tailored to address the complexities of conditional distributions. It does not depend on gradients of the proposed distribution like some other KSD methods, so it is not bogged down by typical computational limitations. GF-KCSD leverages the gradient-free conditional Stein operator, as introduced by Afzali and Muthukumarana (2023), which enables a thorough examination of conditional distributions that are both precise and computationally effective. This operator is defined as follows, capturing the essence of its gradient-free approach:

$$\zeta_{q(\mathbf{y}|\mathbf{x})}(\mathbf{y}, \mathbf{y}') = \omega(\mathbf{y}|\mathbf{x})\omega(\mathbf{y}'|\mathbf{x}') \left[\mathbf{g}.\nabla_{\mathbf{y}} \log q(\mathbf{y}|\mathbf{x}) + \nabla_{\mathbf{y}} \mathbf{g}\right], \tag{5}$$

where $\omega(\mathbf{y}|\mathbf{x}) = \frac{q(\mathbf{y}|\mathbf{x})}{p(\mathbf{y}|\mathbf{x})}$ is an importance weight to correct the bias introduced by the surrogate, q denotes the surrogate distribution, p signifies the target distribution, $\mathbf{g}$ is the test function within the RKHS. This operator adeptly integrates the derivative of the log-density of q with respect to y, bypassing direct derivatives of p, thus embodying the gradient-free characteristic. Here, the function $\mathbf{g} \in \mathcal{G}^d$, where $\mathcal{G}_l^{d_y} = \prod_{j=1}^{d_y} \mathcal{G}_l$, and $\mathcal{G}_l$ represents the RKHS associated with a positive definite kernel $l : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y}$. This operator assumes that q is absolutely continuous with respect to p, signifying that events with probability zero under p also have probability zero under q. This scenario suggests that q provides more detailed information compared to p, with the density ratio $\frac{q}{p}$ being bounded and ∇q being Lipschitz, meaning it varies smoothly across the entire space and remains within a certain range.

Consider two random vectors, $\mathbf{x}$ and $\mathbf{y}$, in the space $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^{D_x} \times \mathbb{R}^{D_y}$. We have a joint sample of independent and identically distributed pairs denoted as $(x_i, y_i)_{i=1}^n \sim p_{0_{xy}}$. The joint density function is defined by Bayes rule as $p_{0_{xy}}(\mathbf{x}, \mathbf{y}) = p_0(\mathbf{y}|\mathbf{x}) p_{0_x}(\mathbf{x})$ with $p_0(\mathbf{y}|\mathbf{x})$ being unknown and only accessible through the joint sample. The goal is to test the proposed conditional density function $p(\mathbf{y}|\mathbf{x})$ as the true conditional distribution $p_0(\mathbf{y}|\mathbf{x})$. This assessment is undertaken through a hypothesis test aimed at gauging the goodness-of-fit of the proposed model with respect to an unknown marginal distribution, denoted as p_{0_x}. The hypothesis test is formulated as follows:

$$\mathbf{H}_0 : p(\mathbf{y}|\mathbf{x}) \overset{p_{0_x}}{=} p_0(\mathbf{y}|\mathbf{x}), \quad \mathbf{H}_1 : p(\mathbf{y}|\mathbf{x}) \overset{p_{0_x}}{\neq} p_0(\mathbf{y}|\mathbf{x}) \tag{6}$$

where $p \overset{p_{0_x}}{=} p_0$ is specified that $p(\mathbf{y}|\mathbf{x}) = p_0(\mathbf{y}|\mathbf{x})$, for p_{0_x}-almost all supporting points $\mathbf{x}$ and for every $\mathbf{y} \in \mathcal{Y}$. Incorporating the conditional component involves multiplying the Stein operator by the kernel function $K(\mathbf{x}, \mathbf{x}')$. As a result, the Kernelized Conditional Stein Operator, denoted as $G_q(\mathbf{x}, \mathbf{y}) = K(\mathbf{x}, \mathbf{x}') \times \zeta_{q(\mathbf{y}|\mathbf{x})}(\mathbf{y}, \mathbf{y}')$, exhibits injectiveness owing to the C_0-universality of $K(\mathbf{x}, \mathbf{x}')$. This injectiveness serves as a direct confirmation of the Stein identity property, showcasing the versatility and reliability of the Kernelized Stein Operator within the context of C_0-universal kernels. The key to the GF-KCSD is the G_q such that the expectation under the distribution p vanish, i.e., for any function $g \in \mathcal{G}_l$:

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{xy}} G_q(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{xy}}[K(\mathbf{x}, \mathbf{x}') \times \zeta_{q(\mathbf{y}|\mathbf{x})}(\mathbf{y}, \mathbf{y}')] = 0 \tag{7}$$

The Gradient-Free Kernel Conditional Stein Discrepancy (GF-KCSD) between conditional probability distributions P_0 and P is a discrepancy measure defined as the supremum (maximum value) of the difference between the expected value of the Kernelized Stein Operator under the distributions P_0 and P, with respect to the norm in the RKHS $\mathcal{G}_l$:

$$S_{p,q}(p_0) = \sup_{\|g\|_{\mathcal{G}_l} \le 1} \left\| \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{0_{xy}}}[G_q(\mathbf{x}, \mathbf{y})] - \cancelto{0}{\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{xy}}[G_q(\mathbf{x}, \mathbf{y})]} \right\|_{\mathcal{G}_l} \tag{8}$$

Having a joint sample $(\mathbf{x_i}, \mathbf{y_i})_{i=1}^n \overset{i.i.d.}{\sim} p_{0_{xy}}$, the estimator for the GF-KCSD is shown as:

$$\widehat{S}^2_{p,q}(p_0) = \frac{1}{n(n-1)} \sum_{i \neq j} G_q((\mathbf{x}_i, \mathbf{y}_i), (\mathbf{x}_j, \mathbf{y}_j)) \tag{9}$$

The explicit form of the GF-KCSD estimator for a joint sample $\{(\mathbf{x_i}, \mathbf{y_i})\}_{i=1}^n$ is presented as:

$$\widehat{S}_{p,q}(p_0) = E_{xy} E_{x'y'} K(\mathbf{x}, \mathbf{x}') \zeta_{q(\mathbf{y}|\mathbf{x})}((x, y), (x', y')) \tag{10}$$

where the explicit form of $\zeta_{q(\mathbf{y}|\mathbf{x})}((x, y), (x', y'))$ is as follows:

$$\begin{aligned} \zeta_{q(\mathbf{y}|\mathbf{x})}((x, y), (x', y')) = \omega(\mathbf{y}|\mathbf{x})\omega(\mathbf{y}'|\mathbf{x}') \Big\{ & l\left(\mathbf{y}, \mathbf{y}'\right) \mathbf{s}_q^T(\mathbf{y}|\mathbf{x}) \mathbf{s}_q\left(\mathbf{y}'|\mathbf{x}'\right) \\ & + \sum_{i=1}^{D_y} \frac{\partial^2}{\partial y_i \partial y_i'} l\left(\mathbf{y}, \mathbf{y}'\right) + \mathbf{s}_q^T(\mathbf{y}|\mathbf{x}) \nabla_{\mathbf{y}'} l\left(\mathbf{y}, \mathbf{y}'\right) \\ & + \mathbf{s}_q^T\left(\mathbf{y}'|\mathbf{x}'\right) \nabla_{\mathbf{y}} l\left(\mathbf{y}, \mathbf{y}'\right) \Big\}, \end{aligned} \tag{11}$$

and $\mathbf{s}_q(\mathbf{y}|\mathbf{x})$ is the gradient of the log-density of $q(\mathbf{y}|\mathbf{x})$. We provided a concise overview of this concept in this section. For more in-depth information and proofs about the GF-KCSD, please refer to Section 2.2 of Afzali and Muthukumarana (2023).

3. Methodology

In this section, we describe the methodology used to investigate the GF-KCSD's sampling distribution, critical values, and performance within our experimental framework. We utilize a modification of the bootstrap method (Efron, 1992) to evaluate the sample distribution and identify critical values for hypothesis testing., which provides insights into the robustness of the GF-KCSD estimator. Simultaneously, we address the complications presented by measurement error by incorporating it into our analysis, thus accounting for the inherent uncertainties in empirical data and enhancing the accuracy and reliability of our findings. This methodological framework lays the groundwork for our subsequent experiment, where we explore the performance of GF-KCSD under various conditions.

3.1. Bootstrap resampling

Our investigation into the GF-KCSD for goodness-of-fit testing embarks on a critical comparison between a specific non-normalized conditional model distribution $p(y|x)$ and a collection of independently and identically distributed (i.i.d.) joint samples $(x_i, y_i)_{i=1}^n$, drawn from an unspecified data-generating distribution $p_0(y|x)$. Given the complexity and non-parametric nature of the GF-KCSD, particularly the absence of a closed-form distribution under H_0, we deploy the bootstrap method. This essential resampling technique estimates an estimator's properties by sampling with replacement from the original dataset making it indispensable strategy for approximating the test threshold when traditional computational approaches are impractical (Arcones & Gine, 1992; Huskova & Janssen, 1993). This aligns our methodology with the forefront of statistical innovation as exemplified in Kernelized Stein Discrepancy (KSD) tests (Jitkrittum et al., 2020; Liu et al., 2016; Yang, Liu, Rao, & Neville, 2018).

Algorithm 1: Estimation of Test Threshold for Hypothesis Testing

1: **Input:** Dataset $(x_i, y_i)_{i=1}^n \sim p_{0_{xy}}$, Number of bootstrap samples $\tilde{n}$, Significance level α
2: **Objective:** To test $H_0 : p \overset{p_{0_x}}{=} p_0$ against $H_1 : p \overset{p_{0_x}}{\neq} p_0$
3: **Compute test statistic** $\widehat{S}^2_{p,q}(p_0) = \frac{1}{n(n-1)} \sum_{i \neq j} G_q((\mathbf{x}_i, \mathbf{y}_i), (\mathbf{x}_j, \mathbf{y}_j))$
4: **for** $b = 1, \cdots, \tilde{n}$ **do**
5: Sample with replacement $\{w_1, \ldots, w_n\} \sim \text{Multinomial}(n; \frac{1}{n}, \ldots, \frac{1}{n})$
6: Adjust weights: $\tilde{w}_i = \frac{w_i - 1}{n}$
7: Compute bootstrap statistic: $\hat{S}^*_{(p,q)b}(p_0) = \sum_{i=1}^n \sum_{j \neq i} \tilde{w}_i \tilde{w}_j G_q((\mathbf{x}_i, \mathbf{y}_i), (\mathbf{x}_j, \mathbf{y}_j))$
8: **end for**
9: Determine the $(1-\alpha)$-quantile $\eta_{1-\alpha}$ of the bootstrap statistics $\{\hat{S}^*_{(p,q)b}(p_0)\}_{b=1}^{\tilde{n}}$
10: **Output:** Test threshold $\eta_{1-\alpha}$ as the estimate for rejecting H_0

By setting the rejection threshold to $\eta_{1-\alpha}$, corresponding to the $(1-\alpha)$-quantile of the asymptotic null distribution, our test rigorously rejects the null hypothesis H_0 if the computed statistic $n\widehat{S}_{p,q}(p_0)$ surpasses this threshold. This technique not only tackles computational issues, but it also represents the evolving landscape of non-parametric hypothesis testing. The Algorithm 1 summarized our approach and outlines the procedures required for the estimation of the test threshold. This algorithm does more than automate computations; it represents a methodological evolution, ensuring that each step—from sampling to the final quantile determination—enhances the accuracy and interpretability of our goodness-of-fit testing.

3.2. Measurement errors

In many applied fields of study, our objective is to model the relationship between a response variable and one or more explanatory variables. However, in some cases, the covariates of interest are neither observable nor measurable. In such instances, we employ indirect measures of the true covariates, known as proxy variables. These proxy variables inevitably measure the actual covariates with some degree of error, leading to the designation of such models as measurement error models. Essentially, measurement error models are useful when the explanatory variables that genuinely govern the response behavior cannot be directly observed. The classical measurement error model is articulated as:

$$w_i = x_i + u_i, \tag{12}$$

where w_i is the observed proxy variable, x_i is the true unobserved covariate, and u_i is the random measurement error, assumed to be independent of x_i and characterized by a mean of zero ($\mathbb{E}[u_i] = 0$) and a variance of σ_u^2 ($\text{Var}[u_i] = \sigma_u^2$).

Two key assumptions underlie measurement error models: (i) unbiasedness, ensuring $\mathbb{E}[w_i|x_i] = x_i$, such that observed values reflect true values on average; and (ii) independence, requiring $\text{Cov}[x_i, u_i] = 0$, preventing systematic biases related to the magnitude of x_i. Violation of these assumptions can lead to biased and inefficient estimates.

This section delves deeply into classical measurement error models, with an emphasis on their consequences in statistical frameworks such as the Linear Gaussian Model and Heteroscedastic Gaussian Model, elucidating their theoretical underpinnings, estimation approaches, and implications when faced with measurement errors.

3.2.1. Linear Gaussian model with classical measurement error

In the analysis of simple linear regression models, it is crucial to account for measurement errors, especially when the true covariates cannot be directly observed. This section explores a simple linear regression model incorporating measurement error, highlighting its impact on the observed relationship between the response variable and covariates.

Consider the true relationship for the ith observation in a simple linear regression model expressed as:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

Here, β_0 and β_1 are the regression coefficients, and ϵ_i represents the random error term, which is assumed to be independent of x_i and normally distributed with mean zero and variance σ_ϵ^2, i.e., $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$. However, instead of observing the true covariate x_i, we observe a proxy variable w_i, which is related to x_i through the classical measurement error model $w_i = x_i + u_i$. where u_i is the measurement error, assumed to be independent of x_i and ϵ_i, and normally distributed with mean zero and variance σ_u^2, i.e., $u_i \sim \mathcal{N}(0, \sigma_u^2)$. Substituting $x_i = w_i - u_i$ into the linear Gaussian model, we obtain:

$$\begin{aligned} y_i &= \beta_0 + \beta_1(w_i - u_i) + \epsilon_i \\ &= \beta_0 + \beta_1 w_i - \beta_1 u_i + \epsilon_i \end{aligned}$$

Defining a new combined error term $\eta_i = \epsilon_i - \beta_1 u_i$, the observed model can be written as:

$$y_i = \beta_0 + \beta_1 w_i + \eta_i$$

The distribution of the combined error term η_i is also normal, as it is a linear combination of two independent normal variables:

$$\eta_i \sim \mathcal{N}(0, \sigma_\epsilon^2 + \beta_1^2 \sigma_u^2) \tag{13}$$

This formulation highlights the impact of measurement error in the explanatory variable x_i on the linear Gaussian model, causing the inflation of error variance and making standard methods like ordinary least squares (OLS) biased and inconsistent. To accurately estimate the regression coefficients β_0 and β_1, as well as the error variances σ_ϵ^2 and σ_u^2, alternative techniques such as errors-in-variables models or instrumental variables are required. These methods are crucial for adjusting the biases introduced by measurement error and ensuring reliable estimation of the true relationship between variables.

3.2.2. Heteroscedastic Gaussian model with classical measurement errors:

Heteroscedastic models play a fundamental role in statistical analyses, particularly when the assumption of homoscedasticity, or constant variance of error terms ϵ_i across observations, is violated. These models adeptly address the variability in uncertainty associated with predictions, a common occurrence in real-world data. The heteroscedastic Gaussian model can be simply expressed as:

$$y_i = f(x_i, \beta) + \epsilon_i, \tag{14}$$

where for the ith observation, y_i denotes the response variable, x_i represents the set of predictor variables, $f(x_i, \beta)$ is a function defining the relationship between predictors and response reliant on parameters β, and ϵ_i is the random error term. Unlike in homoscedastic models, ϵ_i

follows a normal distribution with a mean of zero and a variance of σ_i^2, where σ_i^2 is a function of predictor variables or other known quantities, expressed as $\sigma_i^2 = g(x_i;\theta)$, making the variance heteroscedastic.

Incorporating classical measurement error into a heteroscedastic Gaussian model requires adjusting both the observed relationship and the variance to account for the error introduced by using measured predictors w_i instead of the true covariates x_i. Assuming that the measurement error u_i is normally distributed with mean zero and variance σ_u^2, substituting $x_i = w_i - u_i$ into the heteroscedastic Gaussian model, we obtain:

$$\begin{aligned} y_i &= \beta_0 + \beta_1(w_i - u_i) + \epsilon_i \\ &= \beta_0 + \beta_1 w_i - \beta_1 u_i + \epsilon_i \end{aligned}$$

Defining a new combined error term $\eta_i = \epsilon_i - \beta_1 u_i$, the observed model can be written as:

$$y_i = \beta_0 + \beta_1 w_i + \eta_i$$

The distribution of the combined error term η_i is more complex than in the linear Gaussian case, as its variance now depends on both the true covariate x_i and the measurement error variance σ_u^2:

$$\mathrm{Var}[\eta_i] = \sigma^2(x_i;\theta) + \beta_1^2\sigma_u^2$$

This formulation highlights the additional challenges posed by measurement error in the heteroscedastic Gaussian model. Not only does the measurement error lead to an inflation of the error variance, but it also complicates the estimation of the variance function $\sigma^2(x_i;\theta)$ since the true covariate x_i is unobserved.

To obtain consistent estimates of the regression coefficients β_0 and β_1, the parameter vector θ governing the variance function, and the measurement error variance σ_u^2, specialized estimation techniques are required. These techniques must account for both the heteroscedasticity in the error term and the presence of measurement error in the explanatory variable.

Potential approaches may include extensions of errors-in-variables models, generalized method of moments (GMM) estimators, or other methods specifically designed to handle heteroscedasticity and measurement error simultaneously. The choice of estimation method will depend on the specific assumptions and characteristics of the heteroscedastic Gaussian model under consideration.

4. Experiment design

This section presents an empirical study that investigates the interpretability of GF-KCSD test statistics in conditional models under a variety of settings. Our experiments are separated into two parts: analyses using synthetic data and evaluations using real-world data, each designed to evaluate different aspects of the GF-KCSD method.

4.1. Synthetic data experiments

In the first part, we focus on synthetic data experiments using the experimental setup provided by Jitkrittum et al. (2020), who proposed the Kernel Conditional Stein Discrepancy (KCSD) approach, serving as a benchmark for our analysis. Our work is unique in that it takes into account both symmetric and asymmetric distributions of measurement errors, acknowledging their considerable influence on statistical method performance—a factor often omitted in previous studies including the foundational KCSD paper. We evaluate the statistical power of GF-KCSD by comparing it to existing gradient-based techniques, focusing on how various distributions of measurement errors affect method performance. Our analysis has two main objectives: first, to assess GF-KCSD's effectiveness relative to other methods; and second, to demonstrate how measurement errors influence the efficacy of GF-KCSD and these comparative techniques. This approach allows us to provide a comprehensive evaluation of the performance across methods. We focus on the performance of the subsequent methods.

GF-KCSD: Introduced by Afzali and Muthukumarana (2023), GF-KCSD assesses the fit of conditional distributions without computing gradients, using Gaussian kernels for both x and y. The kernel function is $k(w,w') = \exp\left(-\frac{|w-w'|^2}{2\sigma_w^2}\right)$, with bandwidth $\sigma_w = \text{median}\{|w_i - w_j|\}_{i,j=1}^n$ which is determined by the median heuristic.

KCSD: The Kernel Conditional Stein Discrepancy (KCSD) established by Jitkrittum et al. (2020), evaluates conditional density models using a kernel-based approach, facilitating a reliable assessment of model fit.

MMD: This method, adapting the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) for conditional goodness-of-fit testing, which was extended by Jitkrittum et al. (2020) to evaluate conditional distributions. It partitions the dataset into two subsets, generating new samples from the model $p(\cdot|x^{(2)})$.

Zheng: An enhancement of Zheng (2012)'s method by Jitkrittum et al. (2020) utilizes Gaussian kernels for smoothing, with bandwidth parameter heuristic $h_j = \hat{s}_j n^{-1/(4+d_x)}$.

We explore the performance of these methods on three conditional modeling problems: Linear Gaussian Model (LGM), Heteroscedastic Gaussian Model (HGM), and Quadratic Gaussian Model (QGM). In all cases, we assume the conditional densities $p(y|x)$ and $p_0(y|x)$ are non-normalized. The LGM serves as a baseline, accurately reflecting the true data-generating process, while the HGM and QGM introduce heteroscedastic behavior and a subtle deviation from linearity, respectively, challenging the goodness-of-fit tests. Details of each model are provided below.

Quadratic Gaussian Model (QGM): The Quadratic Gaussian Model (QGM) is essential for assessing the performance of goodness-of-fit testing methods, especially when the suggested conditional model's assumption do not align perfectly with the underlying true distribution. We denote data points by $(x,y) \in \mathbb{R}\times\mathbb{R}$. The proposed model hypothesizes a linear relationship, $p(y|x) \sim \mathcal{N}(x+1,1)$, whereas the true distribution, $p_0(y|x)$, has a quadratic relationship, $\mathcal{N}(0.1x^2 + x + 1, 1)$, and a marginal distribution $p_{0_x}(x) \sim Uniform(-2,2)$.

This setting demonstrates a clear comparison between the basic linear assumption of $p(y|x)$ and the quadratic dynamic of $p_0(y|x)$, highlighting a scenario where the simplicity of the model's assumption might not fully encapsulate the complexity of the data's inherent structure. The inclusion of a slight quadratic factor, weighted at 0.1, introduces a nuanced deviation challenging for goodness-of-fit tests to detect. This reflects real-world situations where theoretical models might oversimplify the actual complexity of the data they aim to represent. This scenario suggests that the alternative hypothesis H_1 is valid, indicating a noticeable discrepancy between the modeled and true distributions. This situation emphasizes the importance of scrutinizing model fit, especially in instances where conventional assumptions may inadequately capture the complexity of the underlying distributions.

Heteroscedastic Gaussian Model (HGM): The Heteroscedastic Gaussian Model (HGM) involves data points $(\mathbf{x},y) \in \mathbb{R}^3\times\mathbb{R}$, where the non-normalized proposed conditional model is specified as $p(y|\mathbf{x}) \sim \mathcal{N}\left(\sum_{i=1}^3 x_i, 1 + 10e^{-|\mathbf{x}-\mathbf{c}|^2}\right)$. This model exhibits heteroscedastic behavior, with the conditional variance being a function of $\mathbf{x}$, centered at the point $\mathbf{c} = \frac{2}{3}\mathbf{1}$, where $\mathbf{1}$ is a vector of ones.

Conversely, The true target conditional distribution, $p_0(y|\mathbf{x}) \sim \mathcal{N}\left(\sum_{i=1}^3 x_i, 1\right)$, is a homoscedastic Gaussian model with a constant unit variance. This sets a distinct contrast between the proposed model's heteroscedastic nature and the homoscedastic true distribution, highlighting the challenge for goodness-of-fit tests to discern the heteroscedastic anomaly localized within specific covariate regions.

The marginal distribution for $\mathbf{x}$ is assumed to be a multivariate Gaussian distribution with a mean vector of zeros and an identity covariance matrix, denoted as $p_{0_x}(\mathbf{x}) \sim \mathcal{N}(\mathbf{0},\mathbf{I})$.

This scenario challenges the capacity of goodness-of-fit tests to detect subtle deviations from a constant variance assumption, thereby

illuminating the nuanced dynamics of real-world data modeling where such uniformity in variance may not always be present.

Linear Gaussian Model (LGM): In this study, we evaluate the Linear Gaussian Model (LGM), where the data points are denoted as $(\mathbf{x}, y) \in \mathbb{R}^5 \times \mathbb{R}$, with $\mathbf{x}$ being a 5-dimensional vector of covariates, and y a scalar response variable. The conditional distribution $p(y|\mathbf{x})$ is assumed to follow a normal distribution with a mean $\sum_{i=1}^{5} i x_i$ and a constant variance of 1, i.e., $p(y|\mathbf{x}) = \mathcal{N}\left(\sum_{i=1}^{5} i x_i, 1\right)$. The predictor variables $\mathbf{x}$ are drawn from a multivariate normal distribution with mean zero and an identity covariance matrix, $\mathcal{N}(\mathbf{0}, \mathbf{I})$. This model specification ensures that the true conditional distribution $p_0(y|\mathbf{x})$ is perfectly aligned with our proposed model $p(y|\mathbf{x})$, satisfying the null hypothesis $H_0 : p(y|\mathbf{x}) = p_0(y|\mathbf{x})$. This perfect alignment is critical in our evaluation of the Type I error rates of various goodness-of-fit tests. Through this alignment, the model serves as an ideal framework to assess whether statistical tests appropriately maintain their nominal error rates under conditions of correct model specification.

Models with measurement errors:

In the context of these three scenarios, we add complexity by including measurement errors in each case. Each data observation w_i is modified by an error term u_i. To realistically reflect the variety of errors found in real-world data, we use two different distributions for these errors. This strategy allows us to more accurately portray the irregularities that are common in real-world data. Here's how we implement this:

- **Symmetric Errors:** For errors following a normal distribution, we define $w_i = x_i + u_i$, with $u_i \sim \mathcal{N}(0, \sigma_u^2)$ where $\sigma_u^2 = [0.5^2, 1^2, 2^2]$. This setup enables an exploration of the models' robustness against random measurement noise at different intensity levels.
- **Asymmetric Errors:** To examine the impact of errors that skew in one direction, showing possible systematic biases, we model $w_i = x_i + u_i$, where $u_i \sim \chi^2(5)$, follows a Chi-square distribution to simulate asymmetric measurement errors.

This structured approach allows us to thoroughly assess how different types of measurement errors affect the accuracy and reliability of our statistical models.

Central to our evaluation is the comparison of GF-KCSD's test power against several established conditional distribution tests: KCSD, FSCD, MMD, and Zheng, across 300 trials with a bootstrap sample size of 500 at a significance level of $\alpha = 0.05$, utilizing Gaussian kernels. To ensure a fair comparison across similar situations in the experiments and all models, the hyperparameters remain consistent throughout all simulations. This consistency allowed for reliable and comparable results across different simulations and models. It is crucial for accurately assessing the performance of the GF-KCSD method relative to other established techniques. This analysis focuses on the method's efficacy in hypothesis testing, specifically in distinguishing the modeled distribution $p(y|x)$ from the true distribution $p_0(y|x)$, with rejection rates serving as a barometer for both Type I error control and the test's power. In this study, we utilized an empirical Gaussian approximation as a surrogate model for the target model, fitting Gaussian parameters (mean and variance) directly from the data.

Fig. 1 illustrates the Type-I error rates and the power of four established tests across three distinct scenarios. The leftmost column of the figure shows that under the true null hypothesis (H_0) scenario, as the sample size increases, the false rejection rates for all tests remain within the specified significance level ($\alpha = 0.05$), accounting for predetermined sampling noise. Notably, the GF-KCSD test maintains error rates below this threshold even in small sample sizes, distinguishing its performance from the other tests when sample size is limited. Additionally, by incorporating measurement errors into the LGM setup, the performance of GF-KCSD remained stable, maintaining control at or below 0.05. In contrast, the other methods exhibited fluctuations and nuances in their performance.

In the scenario involving QGM, presented in the middle column of Fig. 1, an increase in test power with larger sample sizes is observed for all tests, despite the minimal weight assigned to the quadratic component in p_0. Evidently, the GF-KCSD test demonstrates superior performance compared to the MMD test.

In the HGM context, depicted in the third column of Fig. 1, where the disparities between p and p_0 are localized within the X domain, the GF-KCSD test is demonstrated to significantly outperform all other tests in identifying differences in X, particularly the KCSD test. Unlike the KCSD, which shows decreased performance with an increase in the standard deviation of measurement errors or under asymmetric error conditions, the GF-KCSD is adept at handling measurement errors and remains robust against them, maintaining its efficacy even when the measurement errors are significant or asymmetric. Meanwhile, the performance of both the KCSD and MMD tests decreases, especially under conditions of asymmetric measurement errors and when the standard deviation of measurement errors increases in symmetric scenarios.

4.2. Real-world data experiments (brain magnetic resonance imaging (MRI))

In this section, we conduct a comparative evaluation of the Gradient-Free Kernel Conditional Stein Discrepancy (GF-KCSD) and the Kernel Conditional Stein Discrepancy (KCSD) methodologies for hypothesis testing on a real-world dataset. The fundamental goal is to determine whether GF-KCSD performs similarly to the established KCSD test in rejecting or failing to reject the null hypothesis, which posits that the proposed model accurately captures the distribution of the observed data. Our analysis critically examines the consistency of outcomes between GF-KCSD and KCSD, assessing whether both methods lead to similar conclusions about the model's adequacy.

The dataset under investigation comprises preprocessed T1-weighted Magnetic Resonance Imaging (MRI) data from the study by Beheshti et al. (2021). Specifically, the analysis focuses on gray matter (GM) images, where voxel-level brain features are extracted, resulting in 3,747 voxels per volume. The dataset includes 876 Healthy Controls (HC), 70 individuals with Mild Cognitive Impairment (MCI), and 30 patients diagnosed with Alzheimer's Disease (AD), totaling 976 samples. The goal is to predict the age of the samples using these voxel-level features as covariates in the proposed model.

To propose the $p(y|\mathbf{x})$ model for hypothesis test, we implemented a Mixture Density Network (MDN) as described by Bishop (2006), trained on the preprocessed brain MRI dataset. The MDN aims to predict the conditional probability distribution of participants' ages, denoted y, based on their MRI-derived features, $\mathbf{x}$. Each dataset pair $(\mathbf{x}, y)$ belongs to $(\mathbb{R}^{3748} \times \mathbb{R})$, where 3747 features are extracted from MRI data, along with the category of the disease group. The architecture of the network comprises multiple layers, each dedicated to generating parameters for a Gaussian mixture model that characterizes the conditional distribution $p(y|\mathbf{x})$. The probability density function for a Gaussian Mixture Model, given the condition x, can be expressed as:

$$p(y|\mathbf{x}) = \sum_{i=1}^{K} \frac{\exp(a_i(\mathbf{x}))}{\sum_{j=1}^{K} \exp(a_j(\mathbf{x}))} \cdot \mathcal{N}(y|b_i(\mathbf{x}), \exp(c_i(\mathbf{x})))$$

where K is the number of Gaussian components, determining the model's complexity and flexibility. The parameters $a_i(\mathbf{x})$ are the raw network outputs for the mixing coefficients, influencing the weight of each Gaussian component. The mixing coefficients $\pi_i(\mathbf{x}) = \frac{\exp(a_i(\mathbf{x}))}{\sum_{j=1}^{K} \exp(a_j(\mathbf{x}))}$ ensure they sum to 1, representing the probability of each Gaussian component. The parameters $b_i(\mathbf{x})$ are the means, and $c_i(\mathbf{x})$ are the log-variances, with the exponential function ensuring positive variances. The Gaussian distribution $\mathcal{N}(y|b_i(\mathbf{x}), \exp(c_i(\mathbf{x})))$ models the conditional density of y given $\mathbf{x}$. This approach allows the MDN to capture complex, multimodal relationships between $\mathbf{x}$ and y.

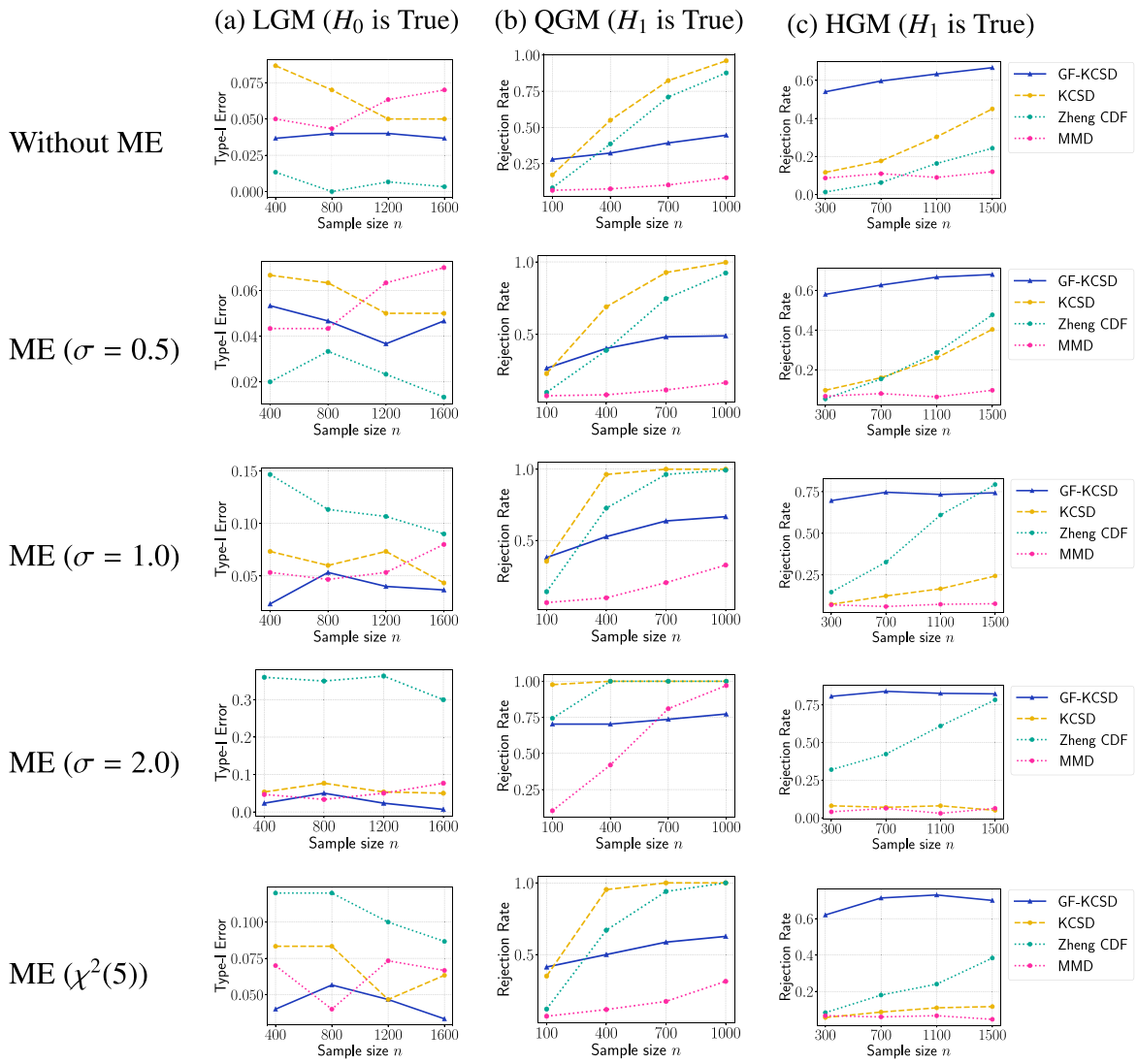

Fig. 1. This figure presents a comparative analysis of Type-I error rates and rejection rates across four distinct tests, all conducted at a significance level of $\alpha = 0.05$. The first row depicts scenarios devoid of measurement errors, while subsequent rows introduce measurement errors. From the second to the fourth row, these errors follow a normal distribution with mean 0 and standard deviations of 0.5, 1.0, and 2.0, respectively. The final row involves measurement errors modeled by a $\chi^2(5)$. **(a):** In these contexts, when H_0 is true, the GF-KCSD method demonstrates robustness in maintaining control over Type-I errors, effectively managing fluctuations attributable to sampling noise. **(b):** Under the alternative hypothesis (H_1) the efficacy of the GF-KCSD in discerning global differences is suboptimal. **(c):** Independent of the presence or absence of measurement errors, GF-KCSD consistently shows a highly effective in identifying local differences.

In this experiment, the network is configured with $K = 20$ Gaussian components and employs a ReLU-based framework to estimate the variances, means, and mixing coefficients. The network is trained to minimize the negative log-likelihood of the observed age relative to the model's predictions, using the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 1×10^{-4} and a batch size of 32. We used grid search to determine the optimal number of Gaussian components (K) and batch size, ensuring that the chosen parameters were well-suited to the data, thereby enhancing the test's effectiveness. Fig. 2 illustrates the comparison of the distribution of the real age and the sampled ages from the trained MDN model.

Next, we regard the trained MDN model as the proposed model $p(y|\mathbf{x})$. Our objective is to compare the performance of the Gradient-Free Kernel Conditional Stein Discrepancy (GF-KCSD) and the Kernel Conditional Stein Discrepancy (KCSD) in determining if p represents the true distribution of the dataset. Given that the MDN model has been trained on this dataset, we expect both GF-KCSD and KCSD to fail to reject the null hypothesis, indicating high model performance.

We calculate test statistics using GF-KCSD and KCSD to measure the discrepancy between the model-predicted and empirically observed distributions. For this purpose, Gaussian kernels with $\sigma = 1$ are employed for both the kernel for x and the kernel for y in the computation of these statistics, using a bootstrap sample size of 500. A significant discrepancy suggests that the model may not adequately represent the data. We compute p-values for each method using bootstrap resampling techniques to quantify the evidence against the null hypothesis

Table 1
Comparison of GF-KCSD and KCSD Test Results for Brain MRI Data.

Metric	GF-KCSD	KCSD
α	0.05	0.05
p-Value	0.81	0.98
Test Statistic	-9.59×10^{-8}	-5.0×10^{-4}
H0 Rejected	False	False
Simulations	500	500
Time (s)	0.57	1.99

The results are presented in Table 1. Both the GF-KCSD and KCSD methods failed to reject the null hypothesis. The p-values associated with each method are 0.81 for GF-KCSD and 0.98 for KCSD, indicating weak evidence against the null hypothesis. The corresponding

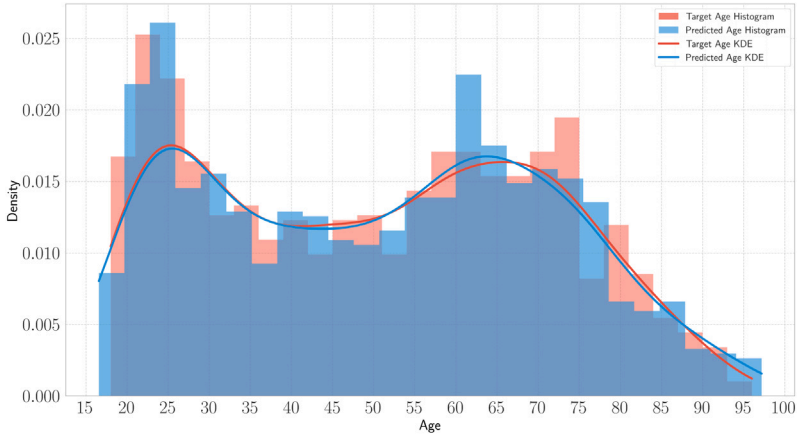

Fig. 2. Comparison of Target and Predicted Age Distributions. The predicted age distribution is generated using a Mixture Density Network (MDN) with 20 Gaussian components trained on preprocessed brain MRI data.

test statistics, representing the magnitude of deviation from the null hypothesis, are -9.59×10^{-8} and -5×10^{-4} for GF-KCSD and KCSD, respectively. These results highlight subtle distinctions in performance between the two methodologies.

To assess computational efficiency, we measured the mean execution time for both tests. Each test was executed 10 times, and the total time taken was recorded. The mean execution time per test was then calculated by dividing the total time by the number of executions. This process provides an estimate of the computational cost of performing the tests on the given dataset. The results indicated an average duration of 0.57 s for GF-KCSD and 1.99 s for KCSD per execution.

Notably, while both methods perform comparably in terms of p-values and test statistics, GF-KCSD requires significantly less computational time. This efficiency arises despite GF-KCSD involving an additional step to compute the surrogate model q. GF-KCSD computes the score function by deriving it from a simpler distribution compared to the more complex approach taken by KCSD. Although using the variational inference estimation of the proposed model p in GF-KCSD might lead to some loss of information, it achieves comparable results to the gradient-based KCSD in less time. The proposed model p could exhibit complex characteristics like multimodality or possess various minima; however, by utilizing an auxiliary distribution in GF-KCSD, which typically represents a unimodal distribution without such complexities, the derivation process becomes significantly more straightforward and time-efficient.

5. Conclusion

In this paper, we have explored the capabilities of the (GF-KCSD, a method we previously introduced). Our investigation delved into synthetic analyses that focused on challenging scenarios like models with measurement errors, as well as complex real-world datasets, with special emphasis on demanding situations such as brain MRI data analysis. The GF-KCSD has proven effective in managing Type-I errors and maintaining statistical power in the face of measurement errors, highlighting its robustness and computational efficiency in a variety of testing conditions. Looking ahead, several promising avenues for future research emerge. The GF-KCSD method could be adapted to discrete domains by modifying the Stein operator, as explored by Yang et al. (2018). Additionally, drawing inspiration from Matsubara, Knoblauch, Briol, and Oates (2022), the capacity of gradient-free methods to conduct Bayesian inference without complex gradient calculations presents a compelling opportunity to develop a generalized framework for Bayesian inference in models with intractable gradients. This framework aims to leverage the robust and efficient properties of GF-KCSD across a broader spectrum of Bayesian applications, particularly in scenarios where gradient computation is challenging or unfeasible. Beyond these advancements, tuning hyperparameters can significantly enhance model performance. Such developments encourage further exploration into kernel-based methods for statistical inference. We leave these and other potential developments for future exploration.

CRediT authorship contribution statement

Elham Afzali: Conceptualization, Methodology, Software, Visualization, Writing – original draft, Data curation. **Saman Muthukumarana:** Reviewing, Supervision, Validation. **Liqun Wang:** Reviewing, Supervision, Validation.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve readability and polish the manuscript's language. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgement

The authors gratefully acknowledge the partial support provided to Dr. Wang by the Natural Sciences and Engineering Research Council of Canada (NSERC) through Grant ID RGPIN-2023-04924 and to Dr. Muthukumarana through Grant ID RGPIN-2024-05460.

References

Afzali, E., & Muthukumarana, S. (2023). Gradient-free kernel conditional stein discrepancy goodness of fit testing. *Machine Learning with Applications*, Article 100463.

Anderson, T. W., & Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 193–212.

Arcones, M. A., & Gine, E. (1992). On the bootstrap of U and V statistics. *The Annals of Statistics*, 655–674.

Beheshti, I., Ganaie, M., Paliwal, V., Rastogi, A., Razzak, I., & Tanveer, M. (2021). Predicting brain age using machine learning algorithms: A comprehensive evaluation. *IEEE Journal of Biomedical and Health Informatics*, *26*(4), 1432–1440.

Bishop, C. M. (2006). *Pattern recognition and machine learning, vol. 2* (pp. 645–678). Springer google schola.

Blei, D. M. (2014). Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, *1*, 203–232.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799.

Box, G. E., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. John Wiley & Sons.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC.

Chwialkowski, K., Strathmann, H., & Gretton, A. (2016). A kernel test of goodness of fit. In *International conference on machine learning* (pp. 2606–2615). PMLR.

Delaigle, A., & Van Keilegom, I. (2021). Deconvolution with unknown error distribution. In *Handbook of measurement error models* (pp. 245–270). Chapman and Hall/CRC.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution* (pp. 569–593). Springer.

Fisher, M. A., Oates, C., et al. (2022). Gradient-free kernel stein discrepancy. arXiv preprint arXiv:2207.02636.

Fuller, W. A., & Fuller, W. A. (1987). *Measurement error models*. WILEY.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 8–38.

Gorham, J., & Mackey, L. (2015). Measuring sample quality with stein's method. *Advances in Neural Information Processing Systems*, *28*.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, *13*(1), 723–773.

Huskova, M., & Janssen, P. (1993). Consistency of the generalized bootstrap for degenerate U-statistics. *The Annals of Statistics*, 1811–1823.

Jitkrittum, W., Kanagawa, H., & Schölkopf, B. (2020). Testing goodness of fit of conditional density models with kernels. In *Conference on uncertainty in artificial intelligence* (pp. 221–230). PMLR.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Liu, Q., Lee, J., & Jordan, M. (2016). A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning* (pp. 276–284). PMLR.

Matsubara, T., Knoblauch, J., Briol, F.-X., & Oates, C. J. (2022). Robust generalised Bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *84*(3), 997–1022.

Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *50*(302), 157–175.

Smirnov, N. V. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin of the Moscow University. Mathematics*, *2*(2), 3–14.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. sixth Berkeley symp. math. stat. prob.* (pp. 583–602).

Yang, J., Liu, Q., Rao, V., & Neville, J. (2018). Goodness-of-fit testing for discrete distributions via Stein discrepancy. In *International conference on machine learning* (pp. 5561–5570). PMLR.

Zheng, X. (2012). Testing parametric conditional distributions using the nonparametric smoothing method. *Metrika*, *75*, 455–469.