Composite goodness-of-fit test with the Kernel Stein Discrepancy and a bootstrap for degenerate U-statistics with estimated parameters

Florian Brück[1], Veronika Reimoser[2], and Fabian Baier[2]

[1]Research Institute for Statistics and Information Science, University of Geneva
[2]Chair of Mathematical Finance, Technical University Munich

October 26, 2025

Abstract

This paper formally derives the asymptotic distribution of a goodness-of-fit test based on the Kernel Stein Discrepancy introduced in (Oscar Key et al. "Composite Goodness-of-fit Tests with Kernels", Journal of Machine Learning Research 26.51 (2025), pp. 1–60). The test enables the simultaneous estimation of the optimal parameter within a parametric family of candidate models. Its asymptotic distribution is shown to be a weighted sum of infinitely many χ^2-distributed random variables plus an additional disturbance term, which is due to the parameter estimation. Further, we provide a general framework to bootstrap degenerate parameter-dependent U-statistics and use it to derive a new Kernel Stein Discrepancy composite goodness-of-fit test.

Keywords: Kernel Stein Discrepancy, goodness-of-fit testing, bootstrap

1 Introduction

Assessing the goodness-of-fit of statistical models to observed data is crucial in both machine learning and statistics. More precisely, for a considered distribution P and observed data $\{X_i\}_{i\in[n]}$ from another distribution Q, goodness-of-fit tests compare the null hypothesis $H_0: P = Q$ against the alternative hypothesis $H_1: P \neq Q$. Traditional tests such as the Kolmogorov-Smirnov, Cramér-von-Mises and Anderson-Darling test compare cumulative distribution functions. However, they struggle with high dimensional data and models where the cumulative distribution functions or the exact likelihoods are intractable.

In case the likelihood is known up to normalizing constant, the so-called Kernel Stein Discrepancy (KSD) can be used for goodness-of-fit testing, see [9, 13, 7] and [1] for a comprehensive overview of its uses in statistics and machine learning. Under some regularity conditions, $\mathrm{KSD}(P, Q) = 0$ iff $P = Q$. A hypothesis test of H_0 is then based on an estimate of $\mathrm{KSD}(P, Q)$, rejecting H_0 when $\mathrm{KSD}(P, Q)$ is sufficiently large. Going beyond testing whether $P = Q$, we can also ask whether the data originates from any member of a parametric family $\{P_\theta\}_{\theta\in\Theta}$, leading to composite goodness-of-fit tests. Their null hypothesis and alternative are

$$H_0^C : \exists\, \theta_0 \in \Theta : Q = P_{\theta_0} \quad \text{against the alternative} \quad H_1^C : Q \notin \{P_\theta\}_{\theta\in\Theta}.$$

In [10], two composite goodness-of-fit tests were proposed, one based on the KSD and one based on the so-called Maximum Mean Discrepancy (MMD). The authors have rigorously derived a composite goodness-of-fit testing framework for the MMD, allowing θ_0 to be estimated simultaneously. Further, they suggested a similar framework for the KSD without rigorously proving its validity stating "*We also include encouraging empirical results for KSD but leave the extension of our theoretical framework to this test for future work*". In this paper, we fill this gap in the literature and formally derive the asymptotic distribution of the KSD estimator proposed in [10], under the null and the alternative.

Formally, the KSD is defined for two probability measures P and Q which have Lebesgue densities p and q, respectively. Omitting some regularity conditions, the KSD is given by

$$\mathrm{KSD}_q(p) := \sqrt{\mathbb{E}_{X,X'\sim q}[h_p(X,X')]},$$

where $h_p(X,X')$ is a function that depends only on p in terms of $\nabla \log(p)$. It is now easy to see that $\mathrm{KSD}_q^2(p)$ can be estimated by the corresponding U-statistic of order two given by

$$\widehat{\mathrm{KSD}}_q^2(p) = \frac{1}{n(n-1)} \sum_{i,j\in[n],i\neq j} h_p(X_i,X_j), \tag{1}$$

where $(X_i)_{i\in[n]}$ denotes an i.i.d. sample from Q, referring to [14] for an introduction to U-statistics. In particular, the KSD can be estimated using an unnormalized version of p, since $\nabla \log(p)$ is independent of the normalizing constant. This property makes the KSD especially useful when the normalizing constant is intractable or unknown. For our composite goodness-of-fit testing framework, we assume that $P_\theta(\mathrm{d}x) = p_\theta(x)\mathrm{d}x$ for all $\theta \in \Theta$ and that we have an estimator $\hat{\theta}_n$ of θ_0 at hand, which allows to estimate $\mathrm{KSD}_q^2(p_{\theta_0})$ via $\widehat{\mathrm{KSD}}_q^2(p_{\hat{\theta}_n})$. However, the main difficulty when working with the KSD is that, under H_0^C, $\widehat{\mathrm{KSD}}_q^2(p_{\theta_0})$ is a degenerate U-statistic with convergence rate n^{-1}, whereas, under H_1^C, $\widehat{\mathrm{KSD}}_q^2(p_{\theta_0})$ is typically a non-degenerate U-statistic with convergence rate $n^{-1/2}$. This significantly complicates the derivation of the asymptotic distribution of $\widehat{\mathrm{KSD}}_q^2(p_{\hat{\theta}_n})$ and even more so the estimation of critical values of the corresponding goodness-of-fit test. It should be noted that [15] and [11] also provided a framework for the derivation of the asymptotic distribution of degenerate U-statistics under parameter estimation, however, their frameworks are not applicable in our context as they require that either the U-statistics core follows a specific form or that it is degenerate w.r.t. the empirical measure of the data generating process, which is usually not satisfied in our setting.

To propose a theoretically valid hypothesis test, the quantiles of the asymptotic distribution also have to be estimated consistently. For the KSD, [10] proposed a parametric and a wild bootstrap scheme for composite goodness-of-fit tests based on the KSD, but did not provide any theoretical validation. Here, we propose an alternative framework and prove its theoretical validity, even when we estimate θ_0 through $\hat{\theta}_n$. Further, we show that the wild bootstrap of [10] does not yield a theoretically valid test, and illustrate that this issue arises more generally when the wild bootstrap is naively extended to degenerate U-statistics. Finally, we want to emphasize that our proposed bootstrap framework is not only valid for the particular case of the KSD. Instead, we provide a general approach to bootstrap degenerate parameter-dependent U-statistics, which is a result of independent interest beyond the scope of this paper.

The paper is structured as follows: We start with an introduction to the KSD. We then provide assumptions under which we derive the asymptotic distribution of the estimator of the KSD while simultaneously estimating a parameter. Then, we provide a general framework to bootstrap parameter-dependent degenerate U-statistic and apply it to the case of the KSD. Finally, we illustrate our method in a short simulation study. All proofs are deferred to the Appendix.

2 The KSD

In the following, we rigorously introduce the KSD. Let $\mathcal{X} \subseteq \mathbb{R}^d$ with a non-empty interior and let q be a strictly positive and continuously differentiable density of the probability measure Q on $\mathbb{R}^d$ with support $\mathcal{X}$. Further, for any function $f : \mathbb{R}^d \to \mathbb{R}$ we denote $\nabla_x f := \left(\frac{\partial}{\partial x_i} f(x)\right)_{1 \leq i \leq d}$ for the gradient of f, implicitly assuming existence, and denote $s_q := (s_{q,i})_{1 \leq i \leq d} = \nabla_x \log(q)$ for the score function.

Definition 1. *We say that a function $f : \mathcal{X} \to \mathbb{R}$ is in the Stein class of q if f is continuously differentiable and satisfies $\int_{\mathcal{X}} \nabla_x (f(x)q(x))\, dx = \mathbf{0}$. A positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be in the Stein class of q if k has continuous second-order partial derivatives, and $k(x, \cdot)$ is in the Stein class of q for any fixed $x \in \mathcal{X}$. Further, we say that k is integrally strictly positive definite if for any function g s.t. $0 < \int_{\mathcal{X}} g(x)^2 \mathrm{d}x < \infty$ we have that $\int_{\mathcal{X}} \int_{\mathcal{X}} g(x)k(x, x')g(x')\mathrm{d}x\mathrm{d}x' > 0$.*

For example, it is easy to check that the Gaussian kernel $k(x, x') = \exp\left(-b\|x - x'|_2^2\right)$ is integrally stricitly positive definite and in the Stein class of any continuously differentiable density with support $\mathbb{R}^d$.

In the following, we work with the definition of the KSD from [13, Definition 3.2, Proposition 3.3, Theorem 3.6], as this representation is suitable for estimation.

Definition 2. *Let k be in the Stein class of q, p denote a strictly positive and continuously differentiable density with support $\mathcal{X}$, $\boldsymbol{\delta}_{p,q}(X) = \boldsymbol{s}_p(X) - \boldsymbol{s}_q(X)$ denote the score difference between p and q and assume that $\int_{\mathcal{X}} \left(q(x)^\intercal \boldsymbol{\delta}_{p,q}(x)\right)^2 \mathrm{d}x < \infty$. Suppose p, q, and k are such that $\mathbb{E}_{X,X' \sim q}[\boldsymbol{\delta}_{p,q}(X)^T k(X, X')\boldsymbol{\delta}_{p,q}(X')] < \infty$. Then, we define the KSD as*

$$\mathrm{KSD}_q^2(p) = \mathbb{E}_{X,X' \sim q}[h_p(X, X')] \tag{2}$$

where X, X' are independent and $h_p : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined as

$$h_p(x, x') = \boldsymbol{s}_p(x)^T k(x, x')\boldsymbol{s}_p(x') + \boldsymbol{s}_p(x)^T \nabla_{x'} k(x, x') + \nabla_x k(x, x')^T \boldsymbol{s}_p(x') + \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i \partial x_i'} k(x, x') \tag{3}$$

This definition ensures that the KSD is well-defined and that $\mathrm{KSD}_q(p) = 0 \Leftrightarrow q = p$. The finiteness of $\int_{\mathcal{X}} \left(q(x)^\intercal \boldsymbol{\delta}_{p,q}(x)\right)^2 \mathrm{d}x < \infty$ heavily depends on the tails of p and q and needs to be checked on a case-by-case basis. It is now straightforward to see that (1) is an estimator of $\mathrm{KSD}_q^2(p)$ and that, whenever $\hat{\theta}_n$ is an estimator of θ_0, $\widehat{\mathrm{KSD}}_q^2(p_{\hat{\theta}_n})$ is an estimator of $\mathrm{KSD}_q^2(p_{\theta_0})$.

3 Asymptotic distribution of the composite KSD estimator

3.1 Assumptions

We now state assumptions that ensure the KSD is always well-defined and enable us to derive the asymptotic distribution of $\widehat{\mathrm{KSD}}_q^2(p_{\hat{\theta}_n})$. As a high-level assumption that is employed throughout the paper we remark that all random variables appearing in the following are assumed to originate from an abstract probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Therefore, statements such $o_{\mathbb{P}}(1)$ and $O_{\mathbb{P}}(1)$ always refer to convergence to 0 in probability and boundedness in probability on this abstract probability space. Moreover, the expectation of a vector valued function $f = (f_1, \ldots, f_d)$ is interpreted componentwise, i.e. $(\mathbb{E}[f_1], \ldots, \mathbb{E}[f_d])$ and similarly for matrices. Moreover, for any norm $\|\cdot\|$ on $\mathbb{R}$ and any matrix $A = (A_{i,j})_{1 \leq i,j \leq d'}$ we define $\|A\| := (\|A_{i,j}\|)_{1 \leq i,j \leq d'}$.

Finally, $\nabla_x^m f$ denotes the collection of all m-th partial derivatives of a function f w.r.t. x, i.e. $\nabla_x f$ is the gradient and $\nabla_x^2 f$ is the Hesse matrix of f, where we use the convention $\nabla^0 f := f$.

Assumption 1. *The observations $(X_i)_{i\in\mathbb{N}}$ are i.i.d. with distribution Q which has a continuously differentiable density q on $\mathcal{X} \subset \mathbb{R}^d$, where $\mathcal{X}$ has non-empty interior.*

Assumption 2. *Θ is a compact and convex subset of $\mathbb{R}^p$ with non-empty interior. $\theta_0 \in \arg\min_{\theta\in\Theta} \mathrm{KSD}_q(p_\theta)$ belongs to the interior of Θ.*

Assumption 3. *The kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is bounded, strictly integrally positive definite, in the Stein class of q and has bounded first order derivatives.*

Not every kernel satisfies this assumption, but the Gaussian kernel, for example, has bounded first and second order derivatives.

Assumption 4. *$\{P_\theta\}_{\theta\in\Theta}$ is an identifiable parametric family of models on $\mathcal{X}$. Each element $P_\theta \in \{P_\theta\}_{\theta\in\Theta}$ has continuously differentiable density p_θ w.r.t. the Lebesgue measure and support $\mathcal{X}$ and satisfies:*

1. *For every $\theta \in \Theta$ the map $x \mapsto p_\theta(x)$ is strictly positive and continuously differentiable on $\mathcal{X}$.*
2. *For every $x \in \mathcal{X}$ we have $\theta \mapsto p_\theta(x) \in \mathcal{C}^4(\Theta)$.*
3. *$\mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla^m s_{p_\theta}(X)\|_1^4\right] < \infty$ for all $m \in \{0, 1\}$.*
4. *$\mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla^m s_{p_\theta}(X)\|_1^2\right] < \infty$ for all $m \in \{2, 3\}$.*
5. *$\int_{\mathcal{X}} q(x)^\intercal \left(s_{p_\theta}(x) - s_q(x)\right)^2 \mathrm{d}x < \infty$ for all $\theta \in \Theta$.*

Essentially, this assumption ensures that θ_0 is unique and that h_θ has high-enough moments such that the corresponding KSD is well-defined and its estimator $\widehat{\mathrm{KSD}}$ converges asymptotically to a non-degenerate distribution.

The next assumption requires the existence of the joint asymptotic distribution of a $2p+1$-dimensional random vector, denoting $\nabla_\theta \widehat{\mathrm{KSD}}_q^2(p_\theta) := \frac{1}{n(n-1)} \sum_{i,j\in[n],i\neq j} \nabla_\theta h_\theta(X_i, X_j)$ for all $\theta \in \Theta$, where we use the short-hand notation $h_\theta := h_{p_\theta}$.

Assumption 5. *Under H_0^C, the random vector*

$$\left(n\,\widehat{\mathrm{KSD}}_q^2(p_{\theta_0}),\ \sqrt{n}\left(\nabla_\theta \widehat{\mathrm{KSD}}_q^2(p_{\theta_0}) - \mathbb{E}_{X,X'\sim Q}\left[\nabla_\theta h_{\theta_0}(X, X')\right]\right),\ \sqrt{n}(\hat{\theta}_n - \theta_0)\right)$$

converges weakly to a real-valued random vector (Z_1, Z_2, Z_3) when $n \to \infty$.

Note that, under H_0^C and Assumptions 1, 3 and 4, Corollary 2 in the appendix shows marginal converge of $\left(n\,\widehat{\mathrm{KSD}}_q^2(p_{\theta_0}), \sqrt{n}\nabla_\theta \widehat{\mathrm{KSD}}_q^2(p_\theta)\right) \to (Z_1, Z_2)$, where $Z_1 \sim \sum_{j=1}^\infty \lambda_j(T_j^2 - 1)$ with $(T_j)_{j\in\mathbb{N}} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ and the λ_j are the eigenvalues of the operator $\psi \mapsto \mathbb{E}_{X\sim Q}[h_{\theta_0}(x, X)\psi(X)]$, and $Z_2 \sim \mathcal{N}\left(0, 4\mathrm{Var}\left(\mathbb{E}_X\left[\nabla_\theta h_{\theta_0}(X, X')\right]\right)\right)$. Further, the joint convergence of $\left(n\,\widehat{\mathrm{KSD}}_q^2(p_{\theta_0}), \sqrt{n}\nabla_\theta \widehat{\mathrm{KSD}}_q^2(p_\theta)\right)$ follows from standard U-statistics theory, exploiting the classical orthogonal eigenfunction expansion of the core h_{θ_0}. Thus, Assumption 5 is solely an assumption on the estimator $\hat{\theta}_n$. Under some technical regularity conditions, the asymptotic normality of $\hat{\theta}_n = \arg\min_{\theta\in\Theta} \widehat{\mathrm{KSD}}_q^2(p_\theta)$ has been obtained in [3, Section 3]. As this estimator is closely-related to classical M-estimators the assumption of joint normality seems a very weak requirement. However, to keep the framework as general as possible, we want to remark that $\hat{\theta}_n$ can be any estimator of θ_0 that obeys the joint convergence assumption.

3.2 Asymptotic distribution under H_0^C and H_1^C

Theorem 1 (Convergence under Null Hypothesis)**.** *Under Assumptions 1-5 and under H_0^C, we have*

$$n\,\widehat{\mathrm{KSD}}_q^2(p_{\hat{\theta}_n}) \xrightarrow{d} Z_1 + Z_2^\intercal Z_3 + Z_3^\intercal H^* Z_3 =: Z.$$

where $H^* := \left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\,\mathrm{KSD}^2(P_{\theta_0}, Q)\right)_{i,j\in[p]}$.

The non-composite version of the preceding theorem has already been derived in Theorem 4.1 of [13], where it is shown that $\widehat{\mathrm{KSD}}_q^2(p_{\theta_0}) \to Z_1$. Comparing the results, we see that the asymptotic distributions differ by the term $Z_2^\intercal Z_3 + Z_3^\intercal H^* Z_3$, which corresponds to the additional "noise" from parameter estimation.

Theorem 2 (Consistency under Alternative Hypothesis)**.** *Under H_1^C and Assumptions 1-4, we have that almost surely*

$$\liminf_{n\to\infty} \widehat{\mathrm{KSD}}_q^2(p_{\hat{\theta}_n}) > 0.$$

This theorem implies that the test statistic $n\,\widehat{\mathrm{KSD}}_q^2(p_{\hat{\theta}_n})$ diverges under H_1^C as $n \to \infty$, allowing us to conduct a valid hypothesis test whenever we have access to the quantiles of Z under H_0^C, which will be the content of the next section.

4 A bootstrap CLT for degenerate U-statistics with estimated parameters

It is a non-trivial problem to bootstrap a degenerate U-statistic, since the naive bootstrap may fail to provide the correct limiting law, as was first shown in [4]. Therefore, bootstrapping degenerate U-statistics requires extra care and several attempts have been made to solve this problem, see for example [2], who provided one of the first frameworks for bootstrapping degenerate U-statistics, and [12, 8] for more recent contributions. In our framework, the problem is even more difficult, as we also need to take into account the effect of the parameter estimation on the limiting distribution of the degenerate U-statistic. Apart from [10], the only works we are aware of that address the bootstrapping of degenerate U-statistics under concurrent parameter estimation are [15, 11]. However, their frameworks require certain properties of the U-statistic kernel which are often not verifiable in practice and a particular example of such a U-statistic kernel is given by (3). To obtain a general bootstrap scheme for degenerate, parameter-dependent U-statistics this section first shows that a naive extension of the bootstrap for non-degenerate U-statistics of [2] does not yield the right asymptotic distribution and we subsequently show how to correctly bootstrap degenerate, parameter-dependent U-statistics.

Let us start by introducing some notation. We use $\xrightarrow{\star}$ to denote weak convergence conditional on almost every sequence $(X_i)_{i\in\mathbb{N}}$. Moreover, for a sequence of random variables Z_n and a deterministic sequence a_n we write $Z_n = O^\star(a_n)$ (respectively, $Z_n = o^\star(a_n)$) to denote Z_n/a_n is bounded (respectively, converges to 0) in probability, conditional on almost every sequence $(X_i)_{i\in\mathbb{N}}$. For an arbitrary function f of two arguments define

$$U_n f := \frac{1}{n(n-1)} \sum_{i,j\in[n], i\neq j} f(X_i, X_j).$$

and define the empirically centered version of f as

$$f_n(\cdot,\cdot) := f(\cdot,\cdot) - \mathbb{E}_{X\sim\mathbb{Q}_n}\left[f(\cdot,X)\right] - \mathbb{E}_{X\sim\mathbb{Q}_n}\left[f(X,\cdot)\right] + \mathbb{E}_{X,X'\sim\mathbb{Q}_n}\left[f(X',X)\right],$$

where $\mathbb{Q}_n := n^{-1}\sum_{1\leq i\leq n}\delta_{X_i}$. It is important to observe that f_n is a degenerate U-statistic core w.r.t. $\mathbb{Q}_n$ for every $n\in\mathbb{N}$.

We are ready to describe the bootstrap scheme. We will use the U-statistic analogue of Efrons bootstrap introduced in [2], i.e., we use sampling with replacement from $(X_1,\ldots,X_n)$. We denote $(X_1^\star,\ldots,X_n^\star)$ as a sample of size n from $\mathbb{Q}_n$ and define the bootstrapped version of an arbitrary U-statistic with core f as

$$U_n^\star f := \frac{1}{n(n-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{n} f(X_i^\star, X_j^\star).$$

Let us specify our high-level assumptions to derive our bootstrap CLT. The assumptions are formulated for an arbitrary parameter-dependent core h_θ (not necessarily the KSD core) and ensure implicitly that we already have verified that the non-bootstrapped version of the test statistic $nU_n h_{\hat\theta_n}$ has a non-degenerate limit.

Assumption 6. *We assume Assumption 1 and the corresponding version of Assumption 5:*

$$\left(n\left(U_n h_{\theta_0} - \mathbb{E}\left[h_{\theta_o}\right]\right), \sqrt{n}\left(U_n\nabla h_{\theta_0} - \mathbb{E}\left[\nabla h_{\theta_0}\right]\right), \sqrt{n}(\hat\theta_n - \theta_0)\right) \to (Z_1,Z_2,Z_3)\in\mathbb{R}^{2p+1}.$$

Further, we assume

1. $\mathbb{E}\left[\sup_{\theta\in\Theta}\|\nabla^m h_\theta(Y_1,Y_2)\|_2^2\right] < \infty$ *for* $m\in\{0,1\}$ *and* $\mathbb{E}\left[\sup_{\theta\in\Theta}\|\nabla^m h_\theta(Y_1,Y_2)\|_1\right] < \infty$ *for* $m\in\{2,3\}$ *where* $Y_1,Y_2\in\{X,X'\}$ *for some i.i.d. copy* X' *of* X.

2. Θ *is compact and convex with non-empty interior and* θ_0 *is an interior point of* Θ. *The estimator of* θ_0 *is a functional of* $\mathbb{Q}_n$, *i.e.* $\hat\theta_n = \psi(X_1,\ldots,X_n)$.

First, we obtain a bootstrap CLT for $(nU_n^\star h_{\theta_0,n}, \sqrt{n}\left(U_n^\star\nabla h_{\theta_0,n} - U_n\nabla h_{\theta_0}\right))$.

Lemma 1. *Under Assumption 6 we have* $(nU_n^\star h_{\theta_0,n}, \sqrt{n}\left(U_n^\star\nabla h_{\theta_0} - U_n\nabla h_{\theta_0}\right)) \xrightarrow{\star} (Z_1,Z_2)$.

Let $\theta_n^\star$ denotes the estimator of θ_0 which is computed from the bootstrap sample, i.e., $\theta_n^\star = \psi(X_1^\star\ldots,X_n^\star)$. Similarly to the previous section, we need to assume that our bootstrapped estimator satisfies a bootstrap CLT jointly with $(nU_n^\star h_{\theta_0,n}, \sqrt{n}\left(U_n^\star\nabla h_{\theta_0} - U_n\nabla h_{\theta_0}\right))$.

Assumption 7. *We assume that*

$$\left(nU_n^\star h_{\theta_0,n}, \sqrt{n}\left(U_n^\star\nabla_\theta h_{\theta_0} - U_n\nabla_\theta h_{\theta_0}\right), \sqrt{n}(\theta_n^\star - \hat\theta_n)\right) \xrightarrow{\star} (Z_1,Z_2,Z_3).$$

This is a high-level assumption which should be valid whenever $\hat\theta_n$ is based on an i.i.d. expansion. For example, if $\hat\theta_n$ is a non-degenerate U-statistic, [2] provides the bootstrap CLT $\sqrt{n}(\theta_n^\star - \hat\theta_n) \xrightarrow{\star} Z_3$ and [2, Remark 2.10 ii] implies the required joint convergence.

A consequence of Lemma 1 is that the naive extension of the bootstrap for degenerate U-statistics fails whenever $Z_2, Z_3 \neq 0$.

Proposition 1. *Under Assumptions 6 and 7 we have that* $nU_n^\star h_{\hat\theta_n^\star,n} \xrightarrow{\star} Z_1$.

The proposition illustrates the we cannot naively apply the bootstrap for degenerate U-statistics when the parameter estimation influences its limiting law, since the limiting law of the naively bootstrapped degenerate U-statistic is equal to the limiting law without parameter estimation. To solve this issue, we suggest using a correction term that increases the variability of the bootstrapped degenerate U-statistic in just the right way to obtain the correct limiting law.

Theorem 3. *Under Assumptions 6 and 7*

$$nU_n^\star h_{\theta_n^\star,n} + n(\theta_n^\star - \hat{\theta}_n)^\intercal \left(U_n^\star \nabla_\theta h_{\theta_n^\star} - U_n \nabla_\theta h_{\hat{\theta}_n}\right) \xrightarrow{\star} Z_1 + Z_2^\intercal Z_3 + Z_3^\intercal \mathbb{E}_{X,X'\sim Q}\left[\nabla_\theta^2 h_{\theta_0}(X,X')\right] Z_3$$

To the best of our knowledge, the theorem provides the first general valid bootstrap scheme for degenerate U-statistics in the presence of parameter estimation. The increased generality however does come at the price of requiring access to ∇h_θ, which might not always be the case.

Remark 1. *Note that $\mathbb{E}[h_{\theta_0}] = 0$ must not be satisfied for Theorem 3 to hold. Therefore, under our assumptions for Theorem 3, we always obtain a finite limit for the bootstrap of a parameter-dependent degenerate U-statistic, even when its mean is non-zero. This is particularly useful in a hypothesis testing framework, where often H_0 is of the form $\mathbb{E}[h_{\theta_0}] = 0$ and one usually needs to ensure that the bootstrapped test statistic is $O^\star(1)$ under H_1^C.*

4.1 Bootstrapping the KSD

It remains to show that the framework for bootstrapping parameter-dependent degenerate U-statistics can be applied to the KSD. Denote $\widetilde{\mathrm{KSD}}_q^2(p_{\theta_n^\star}) := U_n^\star h_{\theta_n^\star,n} + (\theta_n^\star - \hat{\theta}_n)\left(U_n^\star \nabla_\theta h_{\theta_n^\star} - U_n \nabla_\theta h_{\hat{\theta}_n}\right)$, where h_θ denotes the core of the KSD test statistic, and recall that $n\,\widehat{\mathrm{KSD}}_q^2(p_{\hat{\theta}_n}) \to Z$ as specified in Theorem 1. Note that Lemma 2 in the appendix verifies all moment conditions that are required to apply Theorem 3 for the KSD. Therefore, we have the following corollary:

Corollary 1. *Under Assumptions 1-7 and under H_0^C, we have that $n\widetilde{\mathrm{KSD}}_q^2(p_{\theta_n^\star}) \xrightarrow{\star} Z$. Moreover, under H_1^C, $n\widetilde{\mathrm{KSD}}_q^2(p_{\theta_n^\star}) = O^\star(1)$.*

The result can be used to formulate a goodness-of-fit test based on the KSD.

Algorithm 1: A KSD goodness-of-fit test

Requirements: I.i.d. sample $(X_i)_{1\le i\le n}$ from Q, estimator $\hat{\theta}_n$ of $\arg\min_{\theta\in\Theta} \mathrm{KSD}_q(p_\theta)$, number B of bootstrap replications and confidence level γ.

Calculate $\hat{\theta}_n$ and $\widehat{\mathrm{KSD}}_q^2(p_{\hat{\theta}_n})$;
for $1 \le b \le B$ **do**
  Draw a sample $(X_i^\star)_{1\le i\le n}$ of size n from $(X_i)_{1\le i\le n}$;
  Calculate $\theta_n^\star((X_i^\star)_{1\le i\le n})$ and $T^{(b)} := \widetilde{\mathrm{KSD}}_q^2(p_{\theta_n^\star})$;
end

Decision: Reject H_0^C when $\widehat{\mathrm{KSD}}_q^2(p_{\hat{\theta}_n}) > \mathrm{Quantile}\left(1-\gamma; T^{(b)}_{1\le b\le B}\right)$; otherwise accept H_0^C.

Remark 2 (Wild bootstrap for the KSD)**.** *[10] proposed a wild bootstrap to mimic the asymptotic distribution of the V-statistic $\widehat{\mathrm{KSD}}_q^2(p_{\hat{\theta}_n}) + ((n-1)n)^{-1}\sum_{i=1}^n h_{\hat{\theta}_n}(X_i,X_i)$, without providing theoretical guarantees. It*

turns out that, by similar arguments as above, one can show that the asymptotic distribution of their bootstrap procedure is solely $Z_1 + \mathbb{E}\left[h_{\theta_0}(X, X)\right]$*, see Appendix B for the proof of this statement. Therefore, their bootstrap procedure does not yield a theoretically valid testing procedure as it ignores the influence of parameter estimation on the asymptotic distribution.*

5 Simulations

The simulation study analyses the finite sample performance of our composite KSD test. Since the theoretical results from the previous sections are asymptotic, a Monte Carlo study is useful to verify how well the approximations hold for moderate sample sizes. We focus on the empirical level and power of the test described in Algorithm 1. For the kernel k, we choose the Gaussian kernel

$$k(x, y) \;=\; \exp\Big(-\tfrac{1}{2\ell^2}\,\|x - y\|^2\Big), \qquad x, y \in \mathbb{R}^d,$$

with a dimension-dependent bandwidth $\ell(d) = c\sqrt{d}$, for a fixed tuning constant $c > 0$, which was tuned on a separate dataset.

5.1 Monte Carlo study under the null hypothesis

To the best of our knowledge, no existing KSD-based test is capable of handling composite hypotheses while asymptotically controlling the test level under the null. Nevertheless, to assess the finite sample performance, we compare our test with the wild bootstrap-based KSD test of [10], which is shown to not keep its level in Appendix B.

To assess the performance of the two tests under the null, let $X \in \mathbb{R}^d$ denote a d-dimensional multivariate standard normal random vector and set

$$Q \;=\; \mathcal{N}(0, I_d), \qquad X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} Q.$$

The family of models $(P_\theta)_{\theta\in\Theta}$ is the d-variate Gaussian family with unknown mean and covariance and the optimal parameter is given by $\theta_0 = (0, I_d)$, which is estimated via the empirical mean and covariance matrix. We conducted 500 independent Monte Carlo replications and set the number of bootstrap replications for both our bootstrap and the wild bootstrap to 200. Figure 1 shows the empirical rejection probabilities for an asymptotic level of 5% for dimensions $d \in \{1, 4\}$ and sample sizes $n \in \{200, 300, 400, 500, 600\}$, where additional plots can be found in Appendix C.

One can see that in dimension $d = 1$, both tests keep their level equally well. However, for dimension $d = 4$, we can see that the wild bootstrap based test from [10] does not keep its level, whereas our proposed test keeps its level reasonably well, in line with the theoretical findings from the previous sections.

5.2 Monte Carlo study under the alternative hypothesis

Under the alternative hypothesis, we again benchmark our test against the wild bootstrap-based KSD test from [10] as well as the MMD-based composite tests proposed in [5] and [10]. Our framework to assess the test under the alternative is as follows: We draw n samples from a symmetric two-component Gaussian mixture,

$$Q_\mu \;=\; \tfrac{1}{2}\,\mathcal{N}(e_1\mu, I_d) \;+\; \tfrac{1}{2}\,\mathcal{N}(e_1\mu, I_d),$$

with mixing weights $1/2$ and separation parameter $\mu \geq 0$, where e_1 denotes the first unit vector $(1, 0, \ldots, 0) \in \mathbb{R}^d$. Note that $Q_\mu \notin (P_\theta)_{\theta\in\Theta}$ for $\mu > 0$, hence $H_1 :\; Q_\mu \neq P_\theta \;\forall\theta \in \Theta$ holds. We conducted 300 independent

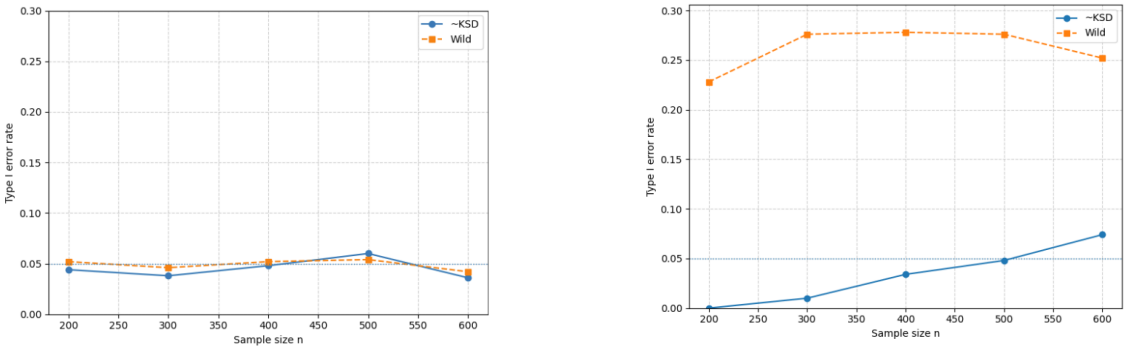

Figure 1: Simulation under the null hypothesis for dimension $d = 1$ (left) and dimension $d = 4$ (right) with $c = 0.2$. KSD denotes the test proposed in this paper whereas Wild denotes the KSD test from [10].

Monte Carlo replications and set the number of bootstrap replications B to 200. Figure 2 reports the power of the tests for varying dimensions $d \in \{1, 2, 4\}$, sample sizes $n \in \{100, 200, 300, 400, 500\}$ and $\mu \in \{1, 2\}$. Additional plots can be found in Appendix C.

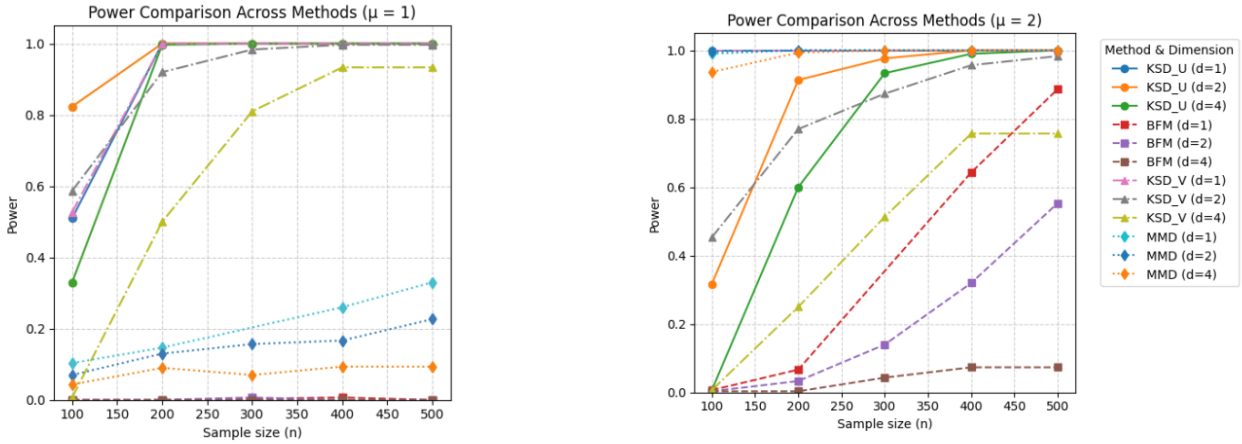

Figure 2: Empirical rejection probabilities under the alternative for dimensions $d \in \{1, 2, 4\}$, $\mu = 1$ (left) and $\mu = 2$ (right) with $c = 1$. KSD_U denotes the test proposed in this paper, KSD_V (resp. MMD) denotes the wild bootstrapped KSD (resp. MMD) tests from [10] and BFM denotes the MMD test from [5]

.

Figure 2 shows that all MMD based tests have a significantly less power than the KSD based tests, independently of the dimension and separation parameter. Furthermore, the KSD based test from this paper outperforms the KSD test of [10] for this example. Moreover, the power of the tests decreases with increasing dimension, which is due to the fact that a difference in only one component of a d-dimensional random vector is harder to uncover with increasing dimension.

Altogether, the findings of the simulation study suggest that the KSD-based test proposed in this paper may have superior finite sample performance over the KSD-based test of [10]. Moreover, it seems that

both KSD-based tests have superior power compared to the MMD-based tests of [5] and [10]. This is not surprising, as the MMD-based tests do not incorporate any information about the underlying density of the fitted family of distributions, in contrast to the KSD. Generalizing these simulation results to other scenarios would requirea more comprehensive analysis, which, however, lies beyond the scope of this paper.

Acknowledgments

The authors would like to express their deepest gratitude to Aleksey Min. Large parts of this project were conducted under his master thesis supervision of Veronika Reimoser at TUM. Moreover, we would like to thank the whole Chair of Mathematical Finance at TUM for allowing us to use their computational resources.

Funding

This work was supported by the Swiss National Science Foundation under Grant 186858.

A Proofs

A.1 Technical results

Lemma 2. *Assumption 3 and 4 imply*

(i) $\mathbb{E}\left[\sup_{\theta\in\Theta}\|h_\theta(X,X')\|_2^2\right]<\infty$ *and* $\mathbb{E}\left[\sup_{\theta\in\Theta}\|h_\theta(X,X)\|_2^2\right]<\infty$,

(ii) $\mathbb{E}\left[\sup_{\theta\in\Theta}\|\nabla_\theta h_\theta(X,X')\|_2^2\right]<\infty$ *and* $\mathbb{E}\left[\sup_{\theta\in\Theta}\|\nabla_\theta h_\theta(X,X)\|_2^2\right]<\infty$,

(iii) $\mathbb{E}\left[\sup_{\theta\in\Theta}\|\nabla_\theta^m h_\theta(X,X')\|_1\right]<\infty$ *and* $\mathbb{E}\left[\sup_{\theta\in\Theta}\|\nabla_\theta^m h_\theta(X,X)\|_1\right]<\infty$ *for all* $2\le m\le 3$.

Further, for all $\theta\in\Theta$ *we have* $\mathbb{E}\left[\nabla_\theta^m h_\theta(X,X')\right]=\nabla_\theta^m\mathbb{E}\left[h_\theta(X,X')\right]$ *for all* $1\le m\le 3$.

Proof. Note that the last statement of the lemma follows from $(i)-(iii)$ by an application of the Leibniz rule with the majorant $\sup_{\theta\in\Theta}\nabla_\theta^m h_\theta(X,X')$. Thus, it remains to show $(i)-(iii)$. Since k and its partial derivatives are bounded we can forget about their influence when showing finiteness of expectations.

From Assumption 4, we can deduce that

$$\mathbb{E}\left[\sup_{\theta\in\Theta}|s_{p_\theta,i}(X)|^a\sup_{\theta\in\Theta}|s_{p_\theta,j}(X)|^v\right]\le\mathbb{E}\left[\sup_{\theta\in\Theta}|s_{p_\theta,i}(X)|^{2a}\right]^{1/2}\mathbb{E}\left[\sup_{\theta\in\Theta}|s_{p_\theta,j}(X)|^{2b}\right]^{1/2}<\infty$$

for all $a,b\in\{0,1,2\}$, and similarly, for $\mathbb{E}\left[\sup_{\theta\in\Theta}|s_{p_\theta,j}(X)|\sup_{\theta\in\Theta}|s_{p_\theta,i}(X)|\right]$. Therefore, since $\mathbb{E}\left[h(X,X')\right]$ and $\mathbb{E}\left[h(X,X)\right]$ can be estimated by such terms, (i) is satisfied.

Next we show (ii). First, observe that $\partial_{\theta_j}h_\theta(X,X)$ is given by

$$k(X,X)\sum_{1\le i\le p}2s_{p_\theta,i}(X)\partial_{\theta_j}s_{p_\theta,i}(X)+\nabla_x k(X,X)^\intercal\partial_{\theta_j}s_{p_\theta}(X)+\nabla_{x'}k(X,X)^\intercal\partial_{\theta_j}s_{p_\theta}(X).\tag{4}$$

Now, let us show (ii). When considering $\|\nabla_\theta h_\theta(X,X)\|_2^2$, we only need to show the finiteness of expectations of squares of terms appearing in (4). For this purpose it is enough to observe that

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\left|s_{p_\theta,i}(X)^a\partial_{\theta_j}s_{p_\theta,i}(X)^b\right|\right]\le\mathbb{E}\left[\sup_{\theta\in\Theta}\left|s_{p_\theta,i}(X)^{2a}\right|\right]^{1/2}\mathbb{E}\left[\sup_{\theta\in\Theta}\left|\partial_{\theta_j}s_{p_\theta,i}(X)^{2b}\right|\right]^{/1/2}<\infty$$

for $a,b\in\{0,1,2\}$. Thus, (ii) follows, as the argument for $\|\nabla_\theta h_\theta(X,X')\|_2^2$ is similar, but fewer moments are needed due to the independence of X,X'.

It remains to show (iii). Observe that each element in $\nabla_\theta^m h_\theta(X,X')$ is of the form

$$k(X,X')\sum_{1\le i\le p}\sum_{0\le j\le m}\partial_\theta^j s_{p_\theta,i}(X)\partial_\theta^{m-j}s_{p_\theta,i}(X')+\nabla_x k(X,X')^\intercal\partial_\theta^m s_{p_\theta}(X)+\nabla_{x'}k(X,X')^\intercal\partial_\theta^m s_{p_\theta}(X')\tag{5}$$

and each element in $\nabla_\theta^m h_\theta(X,X)$ of the form

$$k(X,X)\sum_{1\le i\le p}\sum_{1\le j\le m}2\partial_\theta^{m-j}s_{p_\theta,i}(X)\partial_\theta^j s_{p_\theta,i}(X)+\nabla_x k(X,X)^\intercal\partial_\theta^m s_{p_\theta}(X)+\nabla_{x'}k(X,X)^\intercal\partial_\theta^m s_{p_\theta}(X),\tag{6}$$

where $0\le j\le m$, ∂_θ^j is an abstract notation for $\frac{\partial^j}{\partial\theta_{i_1}\dots\partial\theta_{i_j}}$ for some $(i_l)_{1\le l\le j}\in[p]$, and with the convention $\partial^0 f(\theta)=f(\theta)$. Therefore, we get that $\mathbb{E}\left[\sup_{\theta\in\Theta}\|\nabla_\theta^m h_\theta(X,X')\|_1\right]<\infty$ and $\mathbb{E}\left[\sup_{\theta\in\Theta}\|\nabla_\theta^m h_\theta(X,X)\|_1\right]<\infty$ whenever

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\left|\partial_\theta^j s_{p_\theta,i}(X)\partial_\theta^{m-j}s_{p_\theta,i}(X')\right|\right]\le\mathbb{E}\left[\sup_{\theta\in\Theta}\left|\partial_\theta^j s_{p_\theta,i}(X)\right|\right]\mathbb{E}\left[\sup_{\theta\in\Theta}\left|\partial_\theta^{m-j}s_{p_\theta,i}(X')\right|\right]<\infty$$

and

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\left|\partial_\theta^j s_{p_\theta,i}(X)\partial_\theta^{m-j} s_{p_\theta,i}(X)\right|\right] \leq \mathbb{E}\left[\sup_{\theta\in\Theta}\left|\partial_\theta^j s_{p_\theta,i}(X)\right|^2\right]^{1/2} \mathbb{E}\left[\sup_{\theta\in\Theta}\left|\partial_\theta^{m-j} s_{p_\theta,i}(X')\right|^2\right]^{1/2} < \infty,$$

which is satisfied by Assumption 4. Thus, $\mathbb{E}\left[\sup_{\theta\in\Theta}\|\nabla_\theta^m h_\theta(X,X')\|_1\right] < \infty$. □

Let $H := (H_{h,l})_{1\leq h,l\leq p}$ be defined via the maps $H_{h,l} : \Theta \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as

$$H_{h,l}(\theta, x, x') = \frac{\partial^2}{\partial\theta_h \partial\theta_l} h_\theta(x, x'). \tag{7}$$

Corollary 2. *Assume that* $(X_i)_{i\in\mathbb{N}} \overset{i.i.d.}{\sim} Q$. *Under* H_0^C *and the conditions of Lemma 2 we have* $n\,\widehat{\mathrm{KSD}}_q^2(p_{\theta_0}) \xrightarrow{d} Z_1$ *and* $\sqrt{n}\frac{1}{n(n-1)}\sum_{i,j\in[n],i\neq j}\nabla_\theta h_{\theta_0}(X_i,X_j) \xrightarrow{d} Z_2$, *where* $Z_1 \sim \sum_{j=1}^\infty \lambda_j(T_j^2-1)$ *with* $(T_j)_{j\in\mathbb{N}} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ *and the* λ_j *are the eigenvalues of the operator* $\psi \mapsto \mathbb{E}_{X\sim Q}[h_{\theta_0}(x,X)\psi(X)]$ *as well as* $Z_2 \sim \mathcal{N}(0, 4\,Var\left(\mathbb{E}_X\left[\nabla_\theta h_{\theta_0}(X,X')\right]\right)$. *Moreover* $\frac{1}{n(n-1)}\sum_{i,j\in[n],i\neq j} H_{h,l}(\theta_0,X_i,X_j) \to H^\star_{h,l}$ *for all* $h,l\in[p]$.

Proof. Follows from standard U-statistics theory since the respective moment conditions are satisfied by Lemma 2. □

A.2 Proof of Theorem 1

By Assumption 4, for all $x \in \mathcal{X}$ we have that $p_\theta(X) > 0$ almost surely for $X \sim Q$ and that $p_\theta(x) \in \mathcal{C}^4(\theta)$, so $h_\theta(x,x') \in \mathcal{C}^3(\theta)$ for any $x, x' \in \mathcal{X}$. We are therefore able to perform a second order Taylor expansion around θ_0, that yields

$$\begin{aligned} n\,\widehat{\mathrm{KSD}}_q^2(p_{\hat\theta_n}) = n\,\widehat{\mathrm{KSD}}_q^2(p_{\theta_0}) &+ \sqrt{n}(\hat\theta_n-\theta_0)^T\cdot\sqrt{n}\,\frac{1}{n(n-1)}\sum_{i,j\in[n],i\neq j}\nabla_\theta h_{\theta_0}(X_i,X_j) \\ &+ \sqrt{n}(\hat\theta_n-\theta_0)^T\cdot\left(\frac{1}{n(n-1)}\sum_{i,j\in[n],i\neq j} H(\theta_0,X_i,X_j)\right)\cdot\sqrt{n}(\hat\theta_n-\theta_0) \\ &+ R(\hat\theta_n), \end{aligned}$$

where $R(\hat\theta_n)$ denotes the random remainder term and H is defined in (7). Note that by Lemma 2 we have $\mathbb{E}\left[\nabla_\theta h_{\theta_0}(X,X')\right] = \nabla_\theta \mathbb{E}\left[h_{\theta_0}(X,X')\right] = \nabla_\theta\,\mathrm{KSD}_q(p_{\theta_0}) = 0$ as θ_0 is the unique minimizer of $\theta \mapsto \mathrm{KSD}_q^2(p_\theta)$ by Assumption 4. Moreover, $\mathbb{E}\left[H(\theta_0,X_i,X_j)\right] = \nabla_\theta^2\mathbb{E}\left[h_{\theta_0}(X,X')\right] = H^*$. Thus, the result immediately follows from Assumption 5 if we can show that $R(\hat\theta_n) = o_{\mathbb{P}}(1)$. By Taylor's theorem, the remainder term is of the form

$$R(\hat\theta_n) = n\sum_{i,j,h\in[p]} R_{i,j,h}(\hat\theta_n)\prod_{l\in\{i,j,h\}}(\hat\theta_{n,l}-\theta_{o,l}) \quad \text{with } \|R_{i,j,h}(\hat\theta_n)\| \leq \frac{1}{3!}\sup_{\theta\in\Theta}\left\|\nabla_\theta^3\,\widehat{\mathrm{KSD}}_q^2(p_\theta)\right\|_\infty.$$

For $n\in\mathbb{N}$, we can upper bound

$$\sup_{\theta\in\Theta}\left\|\nabla_\theta^3\,\widehat{\mathrm{KSD}}_q^2(p_\theta)\right\|_\infty \leq \frac{1}{n(n-1)}\sum_{i,j\in[n],i\neq j}\sup_{\theta\in\Theta}\left\|\nabla_\theta^3 h_\theta(X_i,X_j)\right\|_\infty.$$

The r.h.s. of this is a U-statistic. Since $\mathbb{E}_{X,X'\sim Q}\left[\sup_{\theta\in\Theta}\left\|\nabla_\theta^3 h_\theta(X,X')\right\|_1\right] < \infty$, standard U-statistic results yield that the U-statistic converges in distribution to its expectation, which is finite. Thus, $\sup_{\theta\in\Theta}\|\nabla_\theta^3 \widehat{\mathrm{KSD}}_q^2(p_\theta)\|_\infty$ is bounded in probability, yielding

$$R(\hat{\theta}_n) = O_{\mathbb{P}}(n\|\hat{\theta}_n - \theta_0\|^3) = O_{\mathbb{P}}(\|\hat{\theta}_n - \theta_0\|) = o_{\mathbb{P}}(1)$$

and the result follows.

A.3 Proof of Theorem 2

First, recall that by our assumptions $\theta \mapsto \mathrm{KSD}_q(p_\theta)$ is continuous. Due to the compactness of Θ we get that H_1^C implies $\mathrm{KSD}_q(p_{\theta_0}) > 0$. We prove the theorem by contradiction and assume there exists an event A with $\mathbb{P}(A) > 0$ such that $\liminf_{n\to\infty} \widehat{\mathrm{KSD}}_q^2(p_{\hat{\theta}_n}) = 0$ on A. This means that there exists a collection of indices $(a_n)_{n\geq 1}$ such that the subsequence $(\theta_{a_n})_{n\geq 1}$ satisfies

$$\lim_{n\to\infty} \widehat{\mathrm{KSD}}_q^2(p_{\hat{\theta}_{a_n}}) = 0 \quad \text{on } A.$$

Additionally, since Θ is compact, the Bolzano-Weierstrass theorem implies that the sequence $(\hat{\theta}_{a_n})_{n\in\mathbb{N}}$ has a subsequence $(\hat{\theta}_{b_n})_{n\in\mathbb{N}}$ that converges towards a $\theta^* \in \Theta$. By the mean value theorem,

$$\left|\widehat{\mathrm{KSD}}_q^2(p_{\hat{\theta}_{b_n}}) - \widehat{\mathrm{KSD}}_q^2(p_{\theta^*})\right| \leq \|\hat{\theta}_{b_n} - \theta^*\|_1 \frac{1}{n(n-1)} \sum_{i,j\in[n], i\neq j} \sup_{\theta\in\Theta} \|\nabla_\theta h(X_i, X_j)\|_1.$$

Since $\mathbb{E}\left[\sup_{\theta\in\Theta}\|\nabla_\theta h(X_i,X_j)\|_1\right] < \infty$ by Lemma, 2 we get $\left|\widehat{\mathrm{KSD}}_q^2(p_{\hat{\theta}_{b_n}}) - \widehat{\mathrm{KSD}}_q^2(p_{\theta^*})\right| \to 0$, implying that $\mathrm{KSD}_q(p_{\theta^*}) = \lim_{n\to\infty} \widehat{\mathrm{KSD}}_q^2(p_{\theta^*}) = \lim_{n\to\infty} \widehat{\mathrm{KSD}}_q^2(p_{\hat{\theta}_{b_n}}) = 0$, a contradiction to H_1^C.

A.4 Proof of Lemma 1

Lemma 2 immediately implies that the conditions for an application of [2, Theorem 2.4 and Corollary 2.6] are satisfied. As a consequence, we obtain that $nU_n^\star h_{\theta_0,n} \to Z_1$ and $\sqrt{n}\,(U_n^\star \nabla h_{\theta_0} - U_n \nabla h_{\theta_0}) \to Z_2$ (see [2, Remark 2.7]). Joint convergence immediately follows from [2, Remark 2.10 ii].

A.5 Proof of Proposition 1

Apply a third-order Taylor expansion to $U_n^\star h_{\hat{\theta}_n^\star,n}$ to get

$$U_n^\star h_{\theta_n^\star,n} = U_n^\star h_{\theta_0,n} + (\theta_n^\star - \theta_0)^\intercal U_n^\star \nabla_\theta h_{\theta_0,n} + (\theta_n^\star - \theta_0)^\intercal \left(U_n^\star \nabla_\theta^2 h_{\theta_0,n}\right)(\theta_n^\star - \theta_0) + R_n(\theta_n^\star).$$

Note that by [2, Theorem 2.4 a) and c)], $U_n^\star \nabla_\theta h_{\theta_0,n}$ is the bootstrapped version of a U-statistic with degenerate core

$$\nabla_\theta h_{\theta_0} - \mathbb{E}_{X\sim Q}\left[\nabla_\theta h_{\theta_0}(X,\cdot)\right] - \mathbb{E}_{X\sim Q}\left[\nabla_\theta h_{\theta_0}(\cdot,X)\right] + \mathbb{E}_{X,X'\sim Q}\left[\nabla_\theta h_{\theta_0}(X,X')\right] \tag{8}$$

and the same is true for $U_n^\star \nabla_\theta^2 h_{\theta_0,n}$. Thus, they are $O^\star(n^{-1})$. Therefore, $nU_n^\star h_{\theta_n^\star,n} = nU_n^\star h_{\theta_0,n} + O^\star(n^{-1/2}) + nR_n(\theta_n^\star) \overset{\star}{\to} Z_1$ if $R_n(\theta_n^\star) = o^\star(n^{-1})$. To see that $R_n(\theta_n^\star) = o^\star(n^{-1})$, observe that $R_n(\theta_n^\star)$ is a sum of terms of the form

$$\tau := (\theta_n^\star - \theta_0)_i (\theta_n^\star - \theta_0)_j (\theta_n^\star - \theta_0)_k U_n^\star \left(\frac{\partial^3}{\partial\theta_i \partial\theta_j \partial\theta_k} h_{\tilde{\theta},n}\right),$$

where $\tilde{\theta}$ is in the line segment joining $\theta_n^\star$ and θ_0. Since one can check that Assumption 6 implies that

$$\mathbb{E}\left[\left|\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k}h_{\tilde{\theta},n}(X,X')\right|\right] \leq \mathbb{E}\left[\sup_{\theta\in\Theta}\left|\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k}h_{\tilde{\theta},n}(X,X')\right|\right] < \infty$$

and one similarly obtains $\mathbb{E}\left[\left|\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k}h_{\tilde{\theta},n}(X,X)\right|\right] < \infty$ the bootstrap law of large numbers [2, Theorem 4.1] implies $\tau = O^\star(n^{-3/2})$, proving the claim.

A.6 Proof of Theorem 3

From the proof of Proposition 1, we already know that $nU_n^\star h_{\theta_n^\star,n} = nU_n^\star h_{\theta_0} + o^\star(1)$. Thus, it remains to investigate the remaining term

$$\begin{aligned}
&U_n^\star\nabla_\theta h_{\theta_n^\star} - U_n\nabla_\theta h_{\hat{\theta}_n} = U_n^\star\nabla_\theta h_{\theta_0} - U_n\nabla_\theta h_{\theta_0} + U_n^\star\nabla_\theta^2 h_{\theta_0}(\theta_n^\star-\theta_0) - U_n\nabla_\theta^2 h_{\theta_0}(\hat{\theta}_n-\theta_0) + R_n(\theta_n^\star,\hat{\theta}_n)\\
&= U_n^\star\nabla_\theta h_{\theta_0} - U_n\nabla_\theta h_{\theta_0} + U_n^\star\nabla_\theta^2 h_{\theta_0}(\theta_n^\star-\hat{\theta}_n) + \left(U_n^\star\nabla_\theta^2 h_{\theta_0} - U_n\nabla_\theta^2 h_{\theta_0}\right)(\hat{\theta}_n-\theta_0) + R_n(\theta_n^\star,\hat{\theta}_n).
\end{aligned}$$

Note that each term in $R_n(\theta_n^\star,\hat{\theta}_n)$ is of the form

$$(\theta_n^\star-\theta_0)_i(\theta_n^\star-\theta_0)_j\left(U_n^\star\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k}h_{\tilde{\theta}_1}\right) \text{ or } -(\hat{\theta}_n-\theta_0)_i(\hat{\theta}_n-\theta_0)_j U_n\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k}h_{\tilde{\theta}_2},$$

where $\tilde{\theta}_1$ is on the line segment joining θ_0 and $\theta_n^\star$, and $\tilde{\theta}_2$ is on the line segment joining θ_0 and $\hat{\theta}_n$. Therefore, this term is of order $O^\star(n^{-1})$, as all moment conditions required in [2, Theorem 4.1] are satisfied for $\sup_{\theta\in\Theta}\nabla_\theta^3 h_\theta$. By similar arguments, $\left(U_n^\star\nabla_\theta^2 h_{\theta_0} - U_n\nabla_\theta^2 h_{\theta_0}\right)(\hat{\theta}_n-\theta_0) = O^\star(n^{-1})$. This implies that

$$\begin{aligned}
&(\theta_n^\star-\hat{\theta}_n)^\intercal\left(U_n^\star\nabla_\theta h_{\theta_n^\star} - U_n\nabla_\theta h_{\hat{\theta}_n}\right)\\
&= (\theta_n^\star-\hat{\theta}_n)^\intercal\left(U_n^\star\nabla_\theta h_{\theta_0} - U_n\nabla_\theta h_{\theta_0}\right) + (\theta_n^\star-\hat{\theta}_n)^\intercal U_n^\star\nabla_\theta^2 h_{\theta_0}(\theta_n^\star-\hat{\theta}_n) + O^\star(n^{-3/2}).
\end{aligned}$$

Now, [2, Theorem 4.1] yields $U_n^\star\nabla_\theta^2 h_{\theta_0} \xrightarrow{\star} \mathbb{E}_{X,X'\sim Q}\left[\nabla_\theta^2 h_{\theta_0}(X,X')\right]$ and we have derived in Lemma 1 that $\sqrt{n}\left(U_n^\star\nabla_\theta h_{\theta_0} - U_n\nabla_\theta h_{\theta_0}\right) \to Z_2$. Therefore, by the joint convergence, Assumption 7 and Slutsky's Lemma, we get

$$\begin{aligned}
&nU_n^\star h_{\theta_n^\star,n} + n(\theta_n^\star-\hat{\theta}_n)^\intercal\left(U_n^\star\nabla_\theta h_{\theta_n^\star} - U_n\nabla_\theta h_{\hat{\theta}_n}\right)\\
&= nU_n^\star h_{\theta_0} + (\theta_n^\star-\hat{\theta}_n)^\intercal\left(U_n^\star\nabla_\theta h_{\theta_0} - U_n\nabla_\theta h_{\theta_0}\right) + (\theta_n^\star-\hat{\theta}_n)^\intercal\left(U_n^\star\nabla_\theta^2 h_{\theta_0}\right)(\theta_n^\star-\hat{\theta}_n) + o^\star(1)\\
&\xrightarrow{\star} Z_1 + Z_2^\intercal Z_3 + Z_3^\intercal\mathbb{E}_{X,X'\sim Q}\left[\nabla_\theta^2 h_{\theta_0}(X,X')\right]Z_3,
\end{aligned}$$

which proves the claim.

A.7 Proof of Corollary 1

Under H_0^C, it immediately follows from Theorem 3 that $nU_n^\star h_{\theta_n^\star,n} + n(\theta_n^\star-\hat{\theta}_n)\left(U_n^\star\nabla_\theta h_{\theta_n^\star} - U_n\nabla_\theta h_{\hat{\theta}_n}\right) \xrightarrow{\star} Z$, noting that $\mathbb{E}_{X,X'\sim Q}\left[\nabla_\theta^2 h_{\theta_0}(X,X')\right] = H^\star$. Under H_1^C, Remark 1 implies that $n\widetilde{\mathrm{KSD}}_q^2(\theta_n^\star) \xrightarrow{\star} Y := Z_4 + Z_2^\intercal Z_3 + Z_3\mathbb{E}\left[\nabla_\theta^2 h_{\theta_0}(X,X')\right]Z_3$, where $nU_n^\star h_{\theta_0,n} \xrightarrow{\star} Z_4$ with $Z_4 \sim \lim_{n\to\infty} U_n\bar{h}_{\theta_0}$, where $\bar{h}_{\theta_0}$ is the degenerate U-statistic kernel obtained from centering h_{θ_0} as in (8). Thus, $Y = O^\star(1)$.

B Inconsistency of the naive wild bootstrap for U-statistics with estimated parameters

We implicitly assume in this section that all moment conditions for the convergence of the wild bootstrap are satisfied, which allows us to focus on the relevant issue of inconsistency of the wild bootstrap for degenerate U-statistics under parameter estimation. In [10] the authors propose to bootstrap the V-statistic $n\,\widehat{\mathrm{KSD}}_q^2(p_{\hat{\theta}_n}) + n^{-1}\sum_{i\in[n]} h_{\hat{\theta}_n}(X_i, X_i)$ via

$$n^{-1}\sum_{i\neq j\in[n]} W_iW_jh_{\hat{\theta}_n}(X_i,X_j) + n^{-1}\sum_{i\in[n]} W_iW_ih_{\hat{\theta}_n}(X_i,X_i),$$

where $(W_i)_{i\in\mathbb{N}}$ are i.i.d. Rademacher random variables independent of $(X_i)_{i\in\mathbb{N}}$. It is easy to see that by the law of large numbers $n^{-1}\sum_{i\in[n]} W_iW_ih_{\hat{\theta}_n}(X_i,X_i) = n^{-1}\sum_{i\in[n]} h_{\hat{\theta}_n}(X_i,X_i) \to \mathbb{E}\left[h_{\theta_0}(X,X)\right]$. Moreover, $g_{\hat{\theta}_n}\left((W,X),(W',X')\right) := WW'h_{\hat{\theta}_n}(X,X')$ is a symmetric U-statistic core for the sample $((W_i,X_i))_{i\in\mathbb{N}}$. Additionally, for every symmetric function f of two arguments, we have that $g_f((W,X),(W',X')) := WW'f(X,X')$ is a symmetric degenerate U-statistic core for the sample $((W_i,X_i))_{i\in\mathbb{N}}$, as

$$\mathbb{E}_{W_1,X_1}\left[W_1W_2f(X_1,X_2)\right] = 0$$

since W_2 is independent of X_2. Thus, we apply a first-order Taylor expansion to obtain

$$\begin{aligned} n^{-1}\sum_{i\neq j\in[n]} g_{\hat{\theta}_n}\left((W,X),(W',X')\right) &= n^{-1}\sum_{i\neq j\in[n]} g_{\theta_0}\left((W,X),(W',X')\right) \\ &\quad + (\hat{\theta}_n-\theta_0)^\intercal n^{-1}\sum_{i\neq j\in[n]} \nabla_\theta g_{\theta_0}\left((W,X),(W',X')\right) + o^\star(1). \end{aligned}$$

Now, by the previous arguments, we have

$$n^{-1}\sum_{i\neq j\in[n]} \nabla_\theta g_{\theta_0}\left((W,X),(W',X')\right) = O_{\mathbb{P}}(1) \text{ and } n^{-1}\sum_{i\neq j\in[n]} \nabla^2_\theta g_{\theta_0}\left((W,X),(W',X')\right) = O_{\mathbb{P}}(1).$$

Applying that $Z_n = O_{\mathbb{P}}(1) \Rightarrow Z_n = O^\star(1)$, which can derived similarly as [6, Lemma 3], we obtain that

$$n^{-1}\sum_{i\neq j\in[n]} g_{\hat{\theta}_n}\left((W,X),(W',X')\right) = n^{-1}\sum_{i\neq j\in[n]} g_{\theta_0}\left((W,X),(W',X')\right) + o^\star(1) \xrightarrow{\star} Z_1,$$

where $Z_1 \sim \lim_{n\to\infty} n\,\widehat{\mathrm{KSD}}_q^2(p_{\theta_0})$ by the standard wild bootstrap for degenerate U-statistics. Therefore, the wild bootstrap procedure proposed by [10] does not yield an asymptotically correct confidence interval as its limiting distribution is given by $Z_1 + \mathbb{E}\left[h_{\theta_0}(X,X)\right]$, which disregards the influence of parameter estimation.

We remark that the calculations above in no means are dependent on the framework of the KSD. All claims hold for general parameter-dependent U-statistics, which shows that the naive wild bootstrap for degenerate U-statistics with estimated parameters does generally not provide the correct limiting law.

C Additional plots for the simulation study

References

[1] Andreas Anastasiou et al. "Stein's Method Meets Computational Statistics: A Review of Some Recent Developments". In: *Statistical Science* 38.1 (2023), pp. 120 –139. DOI: 10.1214/22-STS863.

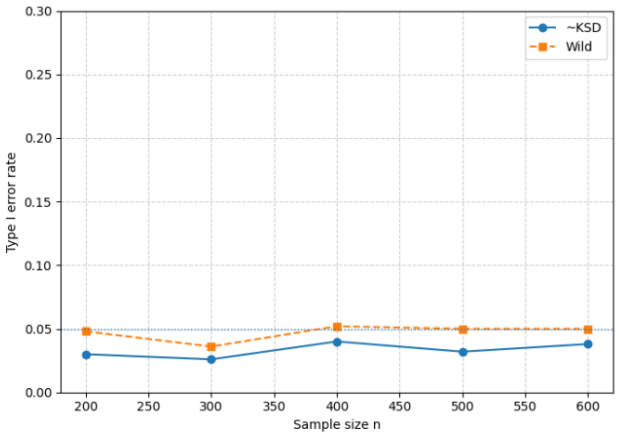

Figure 3: Simulation under the null hypothesis for dimension $d = 2$.

[2] Miguel A. Arcones and Evarist Giné. "On the Bootstrap of U and V Statistics". In: *The Annals of Statistics* 20.2 (1992), pp. 655–674. URL: http://www.jstor.org/stable/2241977.

[3] Alessandro Barp et al. "Minimum Stein Discrepancy Estimators". In: *Advances in Neural Information Processing Systems* 32 (2019).

[4] J. Bretagnolle. "Lois limites du Bootstrap de certaines fonctionnelles". In: *Annales de l'I.H.P. Probabilités et statistiques* 19.3 (1983), pp. 281–296. URL: https://www.numdam.org/item/AIHPB_1983__19_3_281_0/.

[5] Florian Brück, Jean-David Fermanian, and Aleksey Min. "Distribution free MMD tests for model selection with estimated parameters". In: *Journal of Machine Learning Research* (to appear).

[6] Guang Cheng and Jianhua Z. Huang. "Bootstrap consistency for general semiparametric M-estimation". In: *The Annals of Statistics* 38.5 (2010), pp. 2884 –2915. DOI: 10.1214/10-AOS809.

[7] Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. "A Kernel Test of Goodness of Fit". In: *Proceedings of the 33rd International Conference on Machine Learning* 48 (2016), pp. 2606–2615.

[8] Kacper P Chwialkowski, Dino Sejdinovic, and Arthur Gretton. "A wild bootstrap for degenerate kernel tests". In: *Advances in Neural Nnformation Processing Systems* 27 (2014).

[9] Jackson Gorham and Lester Mackey. "Measuring Sample Quality with Stein's Method". In: *Advances in Neural Information Processing Systems* 28 (2015), pp. 226–234.

[10] Oscar Key et al. "Composite Goodness-of-fit Tests with Kernels". In: *Journal of Machine Learning Research* 26.51 (2025), pp. 1–60. URL: http://jmlr.org/papers/v26/24-0276.html.

[11] Anne Leucht and Michael H. Neumann. "Consistency of general bootstrap methods for degenerate U-type and V-type statistics". In: *Journal of Multivariate Analysis* 100.8 (2009), pp. 1622–1633. ISSN: 0047-259X. DOI: https://doi.org/10.1016/j.jmva.2009.01.008.

[12] Anne Leucht and Michael H. Neumann. "Dependent wild bootstrap for degenerate U- and V-statistics". In: *Journal of Multivariate Analysis* 117 (2013), pp. 257–280. DOI: https://doi.org/10.1016/j.jmva.2013.03.003.

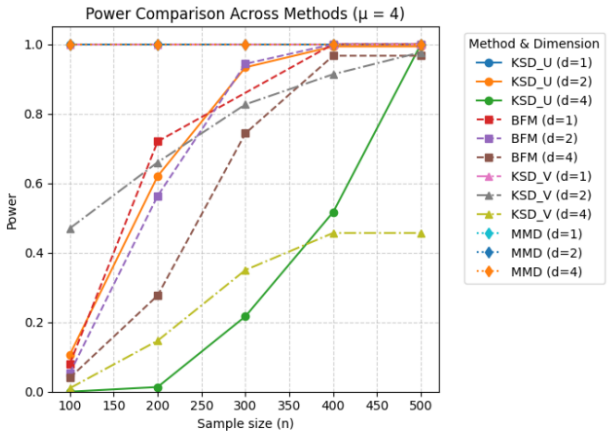

Figure 4: Simulation under the alternative for dimensions $d \in \{1, 2, 4\}$ and $\mu = 4$.

[13] Qiang Liu, Jason Lee, and Michael Jordan. "A Kernelized Stein Discrepancy for Goodness-of-fit Tests and Model Evaluation". In: *Proceedings of the 33rd International Conference on Machine Learning* 48 (2016), pp. 276–284.

[14] Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 1980. DOI: 10.1002/9780470316481.

[15] Tertius de Wet and Ronald H. Randles. "On the Effect of Substituting Parameter Estimators in Limiting χ^2 U and V Statistics". In: *The Annals of Statistics* 15.1 (1987), pp. 398–412.