Journal of Machine Learning Research 26 (2025) 1-52 Submitted 9/23; Revised 12/24; Published 5/25

Distribution Free Tests for Model Selection Based on Maximum Mean Discrepancy with Estimated Parameters

Florian Brück FLORIAN.BRUECK@UNIGE.CH
Research Institute for Statistics and Information Science
University of Geneva, Switzerland

Jean-David Fermanian JEAN-DAVID.FERMANIAN@ENSAE.FR
Ensae-CREST & IP-Paris, France

Aleksey Min MIN@TUM.DE
Department of Mathematics
Technical University of Munich, Germany

Editor: Pradeep Ravikumar

Abstract

There exist several testing procedures based on the maximum mean discrepancy (MMD) to address the challenge of model specification. However, these testing procedures ignore the presence of estimated parameters in the case of composite null hypotheses. In this paper, we first illustrate the effect of parameter estimation in model specification tests based on the MMD. Second, we propose simple model specification and model selection tests in the case of models with estimated parameters. All our tests are asymptotically standard normal under the null, even when the true underlying distribution belongs to the competing parametric families. A simulation study and a real data analysis illustrate the performance of our tests in terms of power and level.

Keywords: distribution free test statistics, maximum mean discrepancy, model specification, model comparison.

1. Introduction

Model selection is an "umbrella term" that refers to different important statistical problems. One common occurrence of model selection is in the context of parametric or semiparametric model estimation. In such cases, the estimated model must be statistically validated, leading to the field of "specification testing". In this work, we will consider models that specify the law of a Data Generating Process (DGP), i.e., we consider models of probability measures. Thus, a model, say $\mathcal{M}$, is associated with a family of probability measures $\mathcal{M} := \{P_\alpha, \alpha \in \Theta_1\}$, where Θ_1 denotes some set of parameters. The true underlying law of the DGP is denoted by P. The empirical law of a sample of n i.i.d. observations from P is denoted $\mathbb{P}_n$. To obtain consistent specification tests, it is necessary to measure the distance between a suggested model and the true underlying DGP. This is commonly done in terms of a semi-metric π between probability measures. The distance between the model $\mathcal{M}$ and the

©2025 Florian Brück, Jean-David Fermanian, Aleksey Min.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v26/23-1199.html.

law P of the DGP is then defined as

$$\pi(\mathcal{M}, P) := \inf_{\alpha\in\Theta_1} \pi(P_\alpha, P).$$

This definition allows to conduct a hypothesis test of

$$\mathcal{H}^{(\pi)}_{0,\mathcal{M}} : \pi(\mathcal{M}, P) = 0,$$

which is a natural way of testing the hypothesis that the model for the DGP is correctly specified. Typically, the best-fitting probability measure(s) in the family $\mathcal{M}$ are defined by so-called pseudo-true parameter(s) (also called pseudo-true value(s)) $\alpha_\star$ that minimize the distance between the model $\mathcal{M}$ and the DGP, i.e., $\alpha_\star \in \arg\min_{\alpha\in\Theta_1} \pi(P_\alpha, P)$. In general, pseudo-true values are unknown and not unique. Assume that one of them, still denoted $\alpha_\star$, can be consistently estimated by a random sequence $(\alpha_n)_{n\geq 1}$. Thus, P_{α_n} should be close to an optimal model in $\mathcal{M}$. Further, P is often also unknown and needs to be estimated via $\mathbb{P}_n$. In practical applications, a consistent test of $\mathcal{H}^{(\pi)}_{0,\mathcal{M}}$ therefore needs to be based on the asymptotic distribution of (an estimator of) $\pi(P_{\alpha_n}, \mathbb{P}_n)$ under the null hypothesis $\mathcal{H}^{(\pi)}_{0,\mathcal{M}}$.

Another common occurrence of the term model selection is when there exist several competing models which must be compared statistically. Here, model selection rather means validating one of the competing model as "the better one". The aim is then to identify the model that is the closest to the DGP, a task that is also called "model comparison". The seminal paper by Vuong (1989) was the first to propose a general framework for this task. The author proposed to use the Kullback-Leibler (KL) divergence for model comparison: a model $\mathcal{M}_1$ is preferred over a model $\mathcal{M}_2$ when its distance to the true model is smaller, i.e., when $\pi(\mathcal{M}_1, P) < \pi(\mathcal{M}_2, P)$ and π is chosen as the KL divergence. In general, model selection between competing models is based on tests of the null hypothesis

$$\mathcal{H}^{(\pi)}_{0,\mathcal{M}_1,\mathcal{M}_2} : \pi(\mathcal{M}_1, P) = \pi(\mathcal{M}_2, P).$$

In most circumstances, the competing models are parametric and may be misspecified, i.e., none of the models will satisfy $\pi(\mathcal{M}_i, P) = 0$, $i \in \{1, 2\}$. For instance, when $\mathcal{M}_1 = \{P_\alpha, \alpha \in \Theta_1\}$ and $\mathcal{M}_2 = \{Q_\beta, \beta \in \Theta_2\}$, their pseudo-true values are defined as above by

$$\alpha_\star \in \operatorname{argmin}_{\alpha\in\Theta_1} \pi(P_\alpha, P), \text{ and } \beta_\star \in \operatorname{argmin}_{\beta\in\Theta_2} \pi(Q_\beta, P).$$

Then, for a given tuple $(\alpha_\star, \beta_\star)$ of pseudo-true values, $\mathcal{H}^{(\pi)}_{0,\mathcal{M}_1,\mathcal{M}_2}$ may be rewritten as

$$\mathcal{H}^{(\pi)}_{0,\mathcal{M}_1,\mathcal{M}_2} : \pi(P_{\alpha_\star}, P) = \pi(Q_{\beta_\star}, P).$$

Once some estimated pseudo-true values α_n and β_n and a sample from P are available, a model selection test is typically based on the random quantity

$$\mathcal{T}_n := \hat{\pi}(P_{\alpha_n}, \mathbb{P}_n) - \hat{\pi}(Q_{\beta_n}, \mathbb{P}_n),$$

where $\hat{\pi}(P_1, P_2)$ denotes an estimator of $\pi(P_1, P_2)$ for any tuple of probability measures (P_1, P_2). Unfortunately, the asymptotic distribution of $\mathcal{T}_n$ is generally complex. Indeed, the randomness of the estimated parameters of both models matters to state the limiting law of $\mathcal{T}_n$ under $\mathcal{H}^{(\pi)}_{0,\mathcal{M}_1,\mathcal{M}_2}$. This is particularly the case for overlapping models. These are models $\mathcal{M}_1$ and $\mathcal{M}_2$ for which $P_{\alpha_\star} = Q_{\beta_\star}$, a situation that often cannot be excluded. For example, this is the case for the KL divergence, where $\mathcal{T}_n$ is no longer distribution free (Vuong, 1989, Theorem 3.3).

Since the seminal works of Durbin (1973a,b), numerous consistent specification tests have been proposed in the statistical literature, depending on the chosen distance π. However, only a few of them manage composite assumptions. This is the case with the Cramer-von Mises and Kolmogorov distances (Pollard, 1980), the uniform distance between multivariate cdfs' (Beran and Millar, 1989), the total variation distance or, equivalently, the L_1 distance between the underlying densities (Cao and Lugosi, 2005) and a L_2-type distance between characteristic functions (Fan, 1997). Recently, the topic has regained attention, since testing composite assumptions based on distances that are popular in generative machine learning/statistics has become an important problem in applied machine learning/statistics. This is notably the case for Wasserstein's distance (Hallin et al., 2021), Kernel Stein Discrepancies (Liu et al., 2016; Chwialkowski et al., 2016) and the Maximum-Mean Discrepancy (Key et al., 2025).

The aim of this paper is to provide a solution for "specification testing" and "model comparison" when π is the Maximum Mean Discrepancy (MMD). The MMD has become a very popular distance in machine learning and statistics since its introduction in Smola et al. (2007). The MMD can be easily estimated even in high-dimensions and it has a nice interpretation in terms of embeddings of probability measures in Reproducing Kernel Hilbert Spaces (RKHS). See Muandet et al. (2017) for a more recent review of the topic. Using the MMD as a discrepancy measure has been proven useful in many statistical applications including robust inference (Chérief-Abdellatif and Alquier, 2022; Alquier and Gerber, 2024), change-point detection (Arlot et al., 2019), goodness-of-fit tests (Gretton et al., 2012; Bounliphone et al., 2015) or copula estimation (Alquier et al., 2023). Further, is has been applied in generative machine learning (Dziugaite et al., 2015; Li et al., 2015, 2017; Sutherland et al., 2017; Zhou et al., 2020) and in a variety of other domains including transfer learning (Long et al., 2017), the computation of functions of random variables (Schölkopf et al., 2015), Bayesian statistics (Fukumizu et al., 2013; Park et al., 2016; Cherief-Abdellatif and Alquier, 2020), clustering (Jegelka et al., 2009), conditional independence testing (Zhang et al., 2011), adaptive MCMC methods (Sejdinovic et al., 2014), causal inference (Lopez-Paz et al., 2015), dynamical systems (Song et al., 2009) and in the construction of sampling algorithms (Hofert et al., 2021; Brück, 2025), among many others. In general, the computational complexity of the MMD is quadratic in the number of data points, which can be restrictive with large data sets. Nonetheless, several recent research papers have provided computationally efficient procedures: see, e.g., the linear statistic in Gretton et al. (2012) or Chatalic et al. (2022).

Let us recall the basics about the MMD, following the approach of Smola et al. (2007). Consider some probability measures defined on a topological space $\mathcal{S}$ equipped with its corresponding Borel sigma-algebra $\mathcal{A}$. Instead of comparing these probability measures directly in the space of probability measures, they may be mapped into an RKHS of real-valued functions $\mathcal{H}$ defined on $\mathcal{S}$. The latter space is associated with a symmetric and positive definite function $k : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$, called kernel, which may be chosen by practitioners. For many kernels k (called "characteristic"), these mappings are injective and the MMD distance between probability measures is then defined as the distance between their respective embeddings in the space of functions $\mathcal{H}$. More specifically, consider some random element X in a topological space $(\mathcal{S},\mathcal{A})$, whose law is P. The embedding of the probability measure P into $\mathcal{H}$ is given by the map $P \mapsto \mathbb{E}_P\left[k(\cdot,X)\right] =: \mu_P$. The latter map implicitly depends on the kernel k, but this dependence is suppressed to lighten notations. When k is characteristic, $P \mapsto \mu_P$ is injective and we have $P_1 \neq P_2$ iff $\mu_{P_1} \neq \mu_{P_2}$. Thus, the MMD defines a distance on the space of probability measures on $\mathcal{S}$ via $\mathrm{MMD}(P_1,P_2) := \|\mu_{P_1} - \mu_{P_2}\|_{\mathcal{H}}$. One can deduce that

$$\mathrm{MMD}^2(P_1,P_2) = \mathbb{E}_{X,X'\sim P_1}\Big[\mathbb{E}_{Y,Y'\sim P_2}\big[k(X,X') - 2k(X,Y) + k(Y,Y')\big]\Big].$$

Therefore, the computation of $\mathrm{MMD}(P_1,P_2)$ does not involve any operations in the Hilbert space $\mathcal{H}$ but is solely dependent on expectations of known functionals w.r.t. P_1 and P_2. Moreover, given an i.i.d. sample $(X_1,X_2,\dots,X_n)$ from P_1 and $(Y_1,Y_2,\dots,Y_n)$ from P_2, $\mathrm{MMD}^2(P_1,P_2)$ can be empirically estimated by

$$\widehat{\mathrm{MMD}}^2(P_1,P_2) := \frac{1}{n(n-1)} \sum_{\substack{i,j=1\\ i\neq j}}^{n} \big\{k(X_i,X_j) - 2k(X_i,Y_j) + k(Y_i,Y_j)\big\}. \tag{1}$$

The latter unbiased estimator of $\mathrm{MMD}^2(P_1,P_2)$ is a U-statistic of degree two associated with the symmetric map

$$h\big((x_1,y_1),(x_2,y_2)\big) := k(x_1,x_2) - k(x_1,y_2) - k(x_2,y_1) + k(y_1,y_2), \tag{2}$$

where $(x_j,y_j) \in \mathcal{S}^2, j \in \{1,2\}$. Hereafter, such maps will be called "U-statistic kernel" (or "U-kernel") to distinguish them from the kernel k. Clearly, $\widehat{\mathrm{MMD}}^2(P_1,P_2)$ may be used to test whether or not $(X_1,X_2,\dots,X_n)$ and $(Y_1,Y_2,\dots,Y_n)$ follow the same underlying probability measure, i.e., to test the null hypothesis $\mathcal{H}_0 : P_1 = P_2$. A consistent test may be deduced from standard U-statistics theory (Serfling, 1980, Section 5) by observing that, under $\mathcal{H}_0$,

$$n\widehat{\mathrm{MMD}}^2(P_1,P_2) \xrightarrow{\text{law}} \sum_{i=1}^{\infty} \lambda_i \left(\mathcal{X}_i^2 - 2\right),$$

where $\mathcal{X}_i \sim \mathcal{N}(0,2)$ and the λ_i denote the (possibly infinitely many) eigenvalues associated with the functional equation $\mathbb{E}\Big[\big\{k(X,y) - \mu_{P_1}(X) - \mu_{P_1}(y) + \mathbb{E}_Y[\mu_{P_1}(Y)]\big\}\psi(X)\Big] = \lambda\psi(y)$ for every $y \in \mathcal{S}$ (Gretton et al., 2012). However, the computation of the eigenvalues λ_i is challenging, which is a limitation of a test for $\mathcal{H}_0$ based on $n\widehat{\mathrm{MMD}}^2(P_1,P_2)$. On the other hand, even if closed-form expressions for these eigenvalues are rarely available, estimating them is feasible through the computation of the eigenvalues of a particular Gram matrix: see Gretton et al. (2009).

Our corresponding null hypothesis for model specification will be defined hereafter as

$$\mathcal{H}_{0,\mathcal{M}} :\ \mathrm{MMD}(P_{\alpha_\star},P) = 0,$$

whereas the null hypothesis for model selection will be defined as

$$\mathcal{H}_{0,\mathcal{M}_1,\mathcal{M}_2} :\ \mathrm{MMD}(P_{\alpha_\star},P) = \mathrm{MMD}(Q_{\beta_\star},P).$$

When some of our results are valid only under $\mathcal{H}_{0,\mathcal{M}}$ or $\mathcal{H}_{0,\mathcal{M}_1,\mathcal{M}_2}$, this will be explicitly specified. For notational convenience, the dependence of $\mathcal{H}_{0,\mathcal{M}}$ and $\mathcal{H}_{0,\mathcal{M}_1,\mathcal{M}_2}$ on $(\alpha_\star, \beta_\star)$ will remain implicit. Moreover, from now on, and unless explicitly stated otherwise, we allow that $\alpha_\star$ and $\beta_\star$ are some general unknown "optimal" parameters, not necessarily pseudo-true values, to keep the mathematical framework as general as possible. For example, $P_{\alpha_\star}$ or $Q_{\beta_\star}$ may be an "optimal model" because of some particular properties such as sparsity, ease of calibration, fairness, etc. In such circumstances, the unknown $\alpha_\star$ (resp. $\beta_\star$) are minimizing some loss function that may not be the MMD distance between $\mathcal{M}_1$ (resp. $\mathcal{M}_2$) and P.

The null hypothesis $\mathcal{H}_{0,\mathcal{M}}$ may be seen as a standard zero assumption for two sample testing. The main difficulty comes from the fact that $\alpha_\star$ is unknown and has to be estimated. In principle, when a sequence $(\alpha_n)_{n\geq 1}$ weakly tends to $\alpha_\star$, resorting to the asymptotic distribution of $\widehat{\text{MMD}}^2(P_{\alpha_n}, P)$ allows to conduct specification testing, i.e., testing the null hypothesis $\mathcal{H}_{0,\mathcal{M}} : \text{MMD}(P_{\alpha_\star}, P) = 0$. Unfortunately, the limiting law of $\widehat{\text{MMD}}^2(P_{\alpha_n}, P)$ might be more complex than that of $\widehat{\text{MMD}}^2(P_{\alpha_\star}, P)$. Moreover, it is usually impossible to generate independent samples from P_{α_n} without resorting to the sample from P, since α_n is usually estimated on the sample from P, adding another layer of complexity. Nonetheless, Key et al. (2025) recently showed that some consistent test of $\mathcal{H}_{0,\mathcal{M}}$ can be obtained by bootstrapping $\text{MMD}^2(P_{\alpha_n}, \mathbb{P}_n)$.

For model comparison, a similar phenomenon as for the test of Vuong (1989) occurs, even if $\alpha_\star$ and $\beta_\star$ were known. When $P_{\alpha_\star} = P = Q_{\beta_\star}$, the rate of convergence of $\widehat{\text{MMD}}^2(P_{\alpha_\star}, P) - \widehat{\text{MMD}}^2(Q_{\beta_\star}, P)$ is n^{-1}; but, otherwise, the rate of convergence is $n^{-1/2}$, as noticed in Bounliphone et al. (2015). Thus, to built a consistent test of $\mathcal{H}_{0,\mathcal{M}_1,\mathcal{M}_2}$, one needs to deal with different rates of convergences. These rates depend on whether or not $P_{\alpha_\star} = P = Q_{\beta_\star}$, a situation that is clearly unknown a priori. A two-step procedure with a pre-test of $P_{\alpha_\star} = P = Q_{\beta_\star}$ seems natural, but it is undesirable due to the difficulty of determining the critical values of $n\big\{\widehat{\text{MMD}}^2(P_{\alpha_n}, P) - \widehat{\text{MMD}}^2(Q_{\beta_n}, P)\big\}$ and multiple testing issues.

In this paper, we provide a solution to the latter problem of model specification and model selection tests based on the MMD. The contributions of our paper can be summarized as follows. First, we investigate the influence of parameter estimation on the asymptotic distribution of the estimator $\widehat{\text{MMD}}^2(P_{\alpha_n}, P)$. We show that it is necessary to account for the influence of parameter estimation to test $\mathcal{H}_{0,\mathcal{M}}$: see Section 2.1 for an illustrative example. Second, as we find that the asymptotic distribution of $n\widehat{\text{MMD}}^2(P_{\alpha_n}, P)$ under the null is rather complex to handle in practical applications, we provide new asymptotically distribution free approaches for specification testing and model comparison. In particular, we show that our novel test statistics converge to a standard normal distribution under the null, allowing straightforward calculations of critical values. Our ideas stem from a generalization of the sample splitting approach introduced in Schennach and Wilhelm (2017), which was used to obtain a standard normal distributed test statistic for Vuong's likelihood ratio test with estimated parameters (Vuong, 1989). Adapting their core ideas, this allows to propose test statistics that are relatively simple to compute and whose asymptotic standard normal law is not influenced by parameter estimation.

The paper is organized as follows. Our test statistics for model selection are introduced in Section 2. In Section 3, we formalize the mathematical framework and state the asymptotic distribution of our test statistic for model specification. A similar structure is followed in Section 4 to manage model comparison tests. Section 5 contains a short simulation study to illustrate the empirical performance of the proposed tests. We apply our methodology to stock returns data in Section 6. Section 7 summarizes the results and sketches further extensions of the framework. Most technical details, particularly the proofs of the main theorems, have been postponed to appendices.

2. Non-Degenerate MMD-based Tests for Model Selection

At this stage, a reader may wonder why our model specification test will not be based solely on the test statistic $\widehat{\text{MMD}}^2(P_{\alpha_n}, P)$, which is an estimate of $\widehat{\text{MMD}}^2(P_{\alpha_\star}, P)$ whose asymptotic law has been investigated in Gretton et al. (2006). To illustrate the technical difficulties in working with $\widehat{\text{MMD}}^2(P_{\alpha_n}, P)$, assume that the estimator α_n of a value $\alpha_\star$ is obtained from the sample drawn under P, which is the natural setup when P_{α_n} is a model for P. Then, any sample from P_{α_n} is inherently dependent due to the common α_n and it is not independent of the initial sample from P. Note that Gretton et al. (2006) requires i.i.d. samples from P_{α_n} to compute $\widehat{\text{MMD}}^2(P_{\alpha_n}, P)$, which is problematic. This issue has first been noticed, but not further investigated, by Lloyd and Ghahramani (2015). However, since the dependence between the samples from P_{α_n} and P deteriorates with increasing n, one might hope that $n\widehat{\text{MMD}}^2(P_{\alpha_n}, P)$ has the same asymptotic law as $n\widehat{\text{MMD}}^2(P_{\alpha_\star}, P)$. Unfortunately, the following example illustrates that this is not the case. Furthermore, it shows that the asymptotic distribution of $n\widehat{\text{MMD}}^2(P_{\alpha_n}, P)$ is significantly more complicated than the asymptotic distribution of $n\widehat{\text{MMD}}^2(P_{\alpha_\star}, P)$. Therefore, we later introduce novel distribution free test statistics when conducting model specification or comparison tests based on the MMD.

2.1 An Illustrative Example

On a probability space $(\Omega, \mathcal{B}, \mathbb{P})$, consider the random variable $X \sim P = \mathcal{N}(0,1)$. Let $(X_i)_{i=1,\dots,n}$ be an i.i.d. sample drawn from P, the law of the DGP. Define a parametric model for the law of X by $\mathcal{M} := \{\mathcal{N}(\alpha, 1),\ \alpha \in \mathbb{R}\}$. For a given parameter α, set

$$Y_i(\alpha) := Y_i + \alpha \sim P_\alpha, \ \text{ where } Y_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \ \ i \in \{1, \dots, n\}.$$

Select $\alpha_\star = 0$ and the "optimal model" $\mathcal{N}(\alpha_\star, 1)$ is identical to the law of the DGP. In practice, $\alpha_\star$ is unknown and will be estimated by $\alpha_n := n^{-1}\sum_{i=1}^n X_i$, for example. Assume we want to test whether $P_{\alpha_\star} = P$, or equivalently $\text{MMD}(P_{\alpha_\star}, P) = 0$, using the test statistic $n\widehat{\text{MMD}}^2(P_{\alpha_n}, P)$. To this purpose, we choose the Gaussian kernel $k(x_1, x_2) = \exp\big(-(x_1 - x_2)^2\big)$, which is characteristic. This is a convenient choice since any map $\alpha \mapsto h\big((X_i, Y_i(\alpha)), (X_j, Y_j(\alpha))\big)$, as defined in (2), is differentiable. Therefore, a second order Taylor expansions around $\alpha_\star = 0$ yields

$$\begin{aligned} n\widehat{\mathrm{MMD}}^2(P_{\alpha_n}, P) &= n\widehat{\mathrm{MMD}}^2(P_{\alpha_\star}, P) \\ &+ \sqrt{n}\,(\alpha_n - 0)\frac{\sqrt{n}}{n(n-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{n}\frac{\partial}{\partial\alpha}h\big((X_i, Y_i+\alpha),(X_j,Y_j+\alpha)\big)|_{\alpha=0} \\ &+ \frac{n(\alpha_n-0)^2}{2}\frac{1}{n(n-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{n}\frac{\partial^2}{\partial\alpha^2}h\big((X_i, Y_i+\alpha),(X_j,Y_j+\alpha)\big)|_{\alpha=0} + o_{\mathbb{P}}(n\alpha_n^2) \\ &= n\widehat{\mathrm{MMD}}^2(P_{\alpha_\star}, P) + \sqrt{n}\alpha_n\sqrt{n}\tilde{U}_{1,n} + \frac{n\alpha_n^2}{2}\tilde{U}_{2,n} + o_{\mathbb{P}}(1), \end{aligned}$$

where $\tilde{U}_{1,n}$ and $\tilde{U}_{2,n}$ denote the U-statistics of degree two corresponding to the first and second order derivatives of $\alpha \mapsto h\big((X_i, Y_i+\alpha),(X_j,Y_j+\alpha)\big)$ at $\alpha_\star$. Thus, $n\widehat{\mathrm{MMD}}^2(P_{\alpha_n}, P)$ can be decomposed into the sum of the "usual" test statistic for a known and fixed $\alpha_\star$ plus the random term $E_n := \sqrt{n}\alpha_n\sqrt{n}\widetilde{U}_{1,n} + n\alpha_n^2\widetilde{U}_{2,n}/2$ that can be attributed to the noise created by the estimation of $\alpha_\star$. Obviously, $n\widehat{\mathrm{MMD}}^2(P_{\alpha_n}, P)$ has the same limiting law as $n\widehat{\mathrm{MMD}}^2(P_{\alpha_\star}, P)$ if and only if $E_n = o_{\mathbb{P}}(1)$. Since $\sqrt{n}\alpha_n \xrightarrow{\text{law}} \mathcal{N}(0,1)$, $E_n = o_{\mathbb{P}}(1)$ if and only if $\sqrt{n}\tilde{U}_{1,n} + \sqrt{n}\alpha_n\tilde{U}_{2,n}/2 = o_{\mathbb{P}}(1)$. By standard results of U-statistics theory and simple calculations, $\sqrt{n}\widetilde{U}_{1,n}$ weakly tends to a $\mathcal{N}(0,\sigma_1^2)$ random variable with $\sigma_1^2 = 8\sqrt{3}/(63\sqrt{7})$ and $\tilde{U}_{2,n} \to 16/(5\sqrt{5})$ a.s. This implies $\sqrt{n}\tilde{U}_{1,n} + \sqrt{n}\alpha_n\tilde{U}_{2,n}/2 \neq o_{\mathbb{P}}(1)$. Therefore, the limiting law of $n\widehat{\mathrm{MMD}}^2(P_{\alpha_n}, P)$ is not equal to the limiting law of $n\widehat{\mathrm{MMD}}^2(P_{\alpha\star}, P)$, due to the influence of parameter estimation.

2.2 New Asymptotically Distribution Free Test Statistics

Let us first introduce some notation. For sequences of random elements $(X_1,\ldots,X_n)$ and $(Y_1,\ldots,Y_n)$, we denote $[\boldsymbol{X}]_{i:j} := \big(X_i, X_{i+1},\ldots,X_j\big)$, $[\boldsymbol{Y}]_{i:j} := \big(Y_i, Y_{i+1},\ldots,Y_j\big)$ and $[\boldsymbol{X},\boldsymbol{Y}]_{i:j} := \big((X_i,Y_i),(X_{i+1},Y_{i+1}),\ \ldots,\ (X_j,Y_j)\big)$ for $1\le i<j\le n$. For $x_1,\ldots,x_n$ and $y_1,\ldots,y_n$, we similarly denote $[\boldsymbol{x}]_{i:j} := \big(x_i, x_{i+1},\ldots,x_j\big)$, $[\boldsymbol{y}]_{i:j} := \big(y_i, y_{i+1},\ldots,y_j\big)$ and $[\boldsymbol{x},\boldsymbol{y}]_{i:j} := \big((x_i,y_i),(x_{i+1},y_{i+1}),\ \ldots,\ (x_j,y_j)\big)$ for $1\le i<j\le n$. In the same manner, $\boldsymbol{X}_{i_1,i_2,\ldots,i_k}$, $\boldsymbol{Y}_{i_1,i_2,\ldots,i_k}$ and $[\boldsymbol{X},\boldsymbol{Y}]_{i_1,i_2,\ldots,i_k}$ are defined, where the index set indicates the components of concatenated variables.

Our new test statistics for model specification and model comparison will be based on a weighted combination of two test statistics, both being a potential candidate for this task. For the sake of simplicity, let us start by comparing two fixed probability measures P_1 and P_2. The first ingredient of our new test statistics is $\widehat{\mathrm{MMD}}^2(P_1,P_2)$ as introduced in (1). Besides $\widehat{\mathrm{MMD}}^2(P_1,P_2)$, a test of $\mathcal{H}_0 : P_1 = P_2$ may also be based on the U-statistic

$$\begin{aligned} \widehat{\mathrm{MMD}}_q^2(P_1,P_2) &:= \frac{1}{n/2(n/2-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{n/2}\big\{k(X_{2i-1},X_{2j-1}) - k(Y_{2j},X_{2i}) \\ &- k(Y_{2i},X_{2j}) + k(Y_{2i-1},Y_{2j-1})\big\} = \frac{1}{n/2(n/2-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{n/2} q\big([\boldsymbol{X},\boldsymbol{Y}]_{2i-1,2i}\,,[\boldsymbol{X},\boldsymbol{Y}]_{2j-1,2j}\big), \end{aligned}$$

introducing the U-statistic kernel

$$q\big([\boldsymbol{x},\boldsymbol{y}]_{1:2}\,,[\boldsymbol{x},\boldsymbol{y}]_{3:4}\big) := k(x_1,x_3) - k(x_4,y_2) - k(x_2,y_4) + k(y_1,y_3).$$

Hereafter, we will assume that n is even for simplicity. Note that $\widehat{\mathrm{MMD}}^2_q$ is a U-statistic based on the sample $\big([\boldsymbol{X},\boldsymbol{Y}]_{1:2},[\boldsymbol{X},\boldsymbol{Y}]_{3:4},\dots,[\boldsymbol{X},\boldsymbol{Y}]_{(n-1):n}\big)$ from $P_1\otimes P_2\otimes P_1\otimes P_2$ of size $n/2$. It is easy to check that $\widehat{\mathrm{MMD}}^2_q(P_1,P_2)$ is an unbiased estimator of $\mathrm{MMD}^2(P_1,P_2)$. By standard U-statistic arguments, we have

$$\sqrt{n}\big\{\widehat{\mathrm{MMD}}^2_q(P_1,P_2) - \mathrm{MMD}^2(P_1,P_2)\big\} \xrightarrow{\text{law}} \mathcal{N}(0,\sigma_q^2),$$

with $\sigma_q^2 > 0$, essentially if and only if P_1 or P_2 is not a Dirac (see (12) below). Therefore, a test of $\mathcal{H}_0 : P_1 = P_2$ may always be based on $\widehat{\mathrm{MMD}}^2_q(P_1,P_2)$. Unfortunately, it may result in a power loss compared to a test based on $\widehat{\mathrm{MMD}}^2(P_1,P_2)$ since $\widehat{\mathrm{MMD}}^2_q(P_1,P_2)$ does not use all pairs of observations which are available from a sample of size n drawn from $P_1\otimes P_2$. It should be noted that a test of $\mathcal{H}_0$ based on $\widehat{\mathrm{MMD}}^2_q(P_1,P_2)$ is similar to the idea of Shekhar et al. (2022), who have recently proposed a MMD-based test statistic of $\mathcal{H}_0$ that is essentially a non-degenerate two-sample U-statistic. Similarly to the expected behavior for $\widehat{\mathrm{MMD}}^2_q(P_1,P_2)$, Shekhar et al. (2022) observe a power loss of their test statistic in comparison to a test based on $\widehat{\mathrm{MMD}}^2(P_1,P_2)$. Therefore, it is desirable to approximately keep the power of a test based on $\widehat{\mathrm{MMD}}^2(P_1,P_2)$, while resorting to the critical values of a normal distribution under the null hypothesis.

To this purpose, we will consider a weighted sum of $\widehat{\mathrm{MMD}}^2(P_1,P_2)$ and $\widehat{\mathrm{MMD}}^2_q(P_1,P_2)$ in the spirit of Schennach and Wilhelm (2017): introduce some (possibly random and data dependent) weights $\epsilon_n > 0$ and define the test statistic

$$\widehat{\mathrm{MMD}}^2_{\epsilon_n}(P_1,P_2) := \widehat{\mathrm{MMD}}^2(P_1,P_2) + \epsilon_n \widehat{\mathrm{MMD}}^2_q(P_1,P_2). \tag{3}$$

If $\epsilon_n := \epsilon > 0$ is a constant, it is obvious that $\sqrt{n}\,\widehat{\mathrm{MMD}}^2_{\epsilon_n}(P_1,P_2) \xrightarrow{\text{law}} \epsilon\mathcal{N}(0,\sigma_q^2)$ under $\mathcal{H}_0$, since $\sqrt{n}\,\widehat{\mathrm{MMD}}^2(P_1,P_2)$ tends to zero in probability when $P_1 = P_2$. However, the choice $\epsilon_n = \epsilon > 0$ may lead to a power loss, similar to a test based on $\sqrt{n}\,\widehat{\mathrm{MMD}}^2_q(P_1,P_2)$. Therefore, we impose that ϵ_n tends to zero in probability hereafter and we will prove that a test of $\mathcal{H}_0$ can be conducted via the test statistic

$$\mathcal{T}_n(P_1,P_2) := \frac{\sqrt{n}\,\widehat{\mathrm{MMD}}^2_{\epsilon_n}(P_1,P_2)}{\hat{\sigma}_n}, \tag{4}$$

where $\hat{\sigma}_n^2 := \hat{\sigma}_n^2(\epsilon_n,P_1,P_2)$ denotes an estimator of the asymptotic variance of the numerator of $\mathcal{T}_n(P_1,P_2)$. With our choice of $\hat{\sigma}_n$, which is described in Section 3.2 and 3.3, and under $\mathcal{H}_0$, the test statistic $\mathcal{T}_n(P_1,P_2)$ converges weakly to $\mathcal{N}(0,1)$. Imposing $\epsilon_n \to 0$ allows to approximately keep the power of $n\widehat{\mathrm{MMD}}^2(P_1,P_2)$ under the alternative and also to avoid the problem of computing the critical values of the asymptotic law of $n\widehat{\mathrm{MMD}}^2(P_1,P_2)$.

Instead of considering a tuple of fixed probability measures (P_1,P_2) we now consider an underlying parametric model $\mathcal{M} := \{P_\alpha,\ \alpha\in\Theta_1\}$ for the DGP, as in Section 1. With the same notations as above, a "specification test" of $\mathcal{H}_{0,\mathcal{M}} : \mathrm{MMD}(P_{\alpha_\star}, P) = 0$ can be based on the statistic

$$\mathcal{T}_n(\mathcal{M}, P) := \sqrt{n}\, \frac{\widehat{\mathrm{MMD}}^2_{\epsilon_n}(P_{\alpha_n}, P)}{\hat{\sigma}_n}, \tag{5}$$

for some sequence of parameters $(\alpha_n)_{n\geq 1}$ that weakly converges to $\alpha_\star$ at rate $n^{-1/2}$. Typically, the estimated parameters α_n are obtained from an i.i.d. sample $(X_1, X_2, \dots, X_n)$ from P. To mimic the previous situation and to calculate $\mathcal{T}_n(\mathcal{M}, P)$, we have to draw a sample $\big(Y_1(\alpha_n), Y_2(\alpha_n), \dots, Y_n(\alpha_n)\big)$ from P_{α_n}. However, such a sample cannot be i.i.d. due to the common dependence on α_n. Moreover, it cannot be independent of $(X_1, X_2, \dots, X_n)$ when α_n is deduced from the latter sample. Additionally, the denominator $\hat{\sigma}_n$ in (5) depends on ϵ_n and needs to be calculated from the sample $\big(X_i, Y_i(\alpha_n)\big)_{i=1,\dots,n}$. Nevertheless, it is shown in Section 3 that $\mathcal{T}_n(\mathcal{M}, P)$ still converges to a standard normal distribution under $\mathcal{H}_{0,\mathcal{M}}$, regardless of the technical problems induced by parameter estimation.

In the case of two competing models $\mathcal{M}_1$ and $\mathcal{M}_2$ for P, one may conduct a model comparison based on the difference $\mathrm{MMD}^2(P_{\alpha_\star}, P) - \mathrm{MMD}^2(Q_{\beta_\star}, P)$ to judge which family of probability measures is closest to the true underlying model P. The idea of testing the null hypothesis $\mathrm{MMD}(P_{\alpha_\star}, P) = \mathrm{MMD}(Q_{\beta_\star}, P)$ based on $\sqrt{n}\big\{\widehat{\mathrm{MMD}}^2(P_{\alpha_\star}, P) - \widehat{\mathrm{MMD}}^2(Q_{\beta_\star}, P)\big\}$ was first proposed by Bounliphone et al. (2015). However, the latter authors only considered fixed competing probability measures and excluded the "degenerate" situation $P_{\alpha_\star} = P = Q_{\beta_\star}$. In the case of parametric models for P, the estimates α_n and β_n of $\alpha_\star$ and $\beta_\star$ are usually obtained from $(X_1, X_2, \dots, X_n)$. The framework of Bounliphone et al. (2015) requires access to independent i.i.d. samples from P_{α_n} and Q_{β_n}, which also need to be independent of the sample from P. Obviously, this is impossible when α_n and β_n are evaluated on the sample $(X_i)_{i=1,\dots,n}$. Note that estimation of α_n and β_n on separate hold-out sets from another P-sample does not resolve the issue, since the samples from P_{α_n} and Q_{β_n} are inherently dependent due to the common dependence w.r.t. the estimated parameters. Therefore, a direct application of the test statistic proposed in Bounliphone et al. (2015) to model comparison problems is not mathematically justified. Moreover, since it has to be excluded that $P = P_{\alpha_\star} = Q_{\beta_\star}$, their test is not applicable to every modeling problem as the assumption $P = P_{\alpha_\star} = Q_{\beta_\star}$ may be reasonable for some competing generative models, e.g. such as GANs.

In our paper, we rectify these shortcomings by introducing a test for the null hypothesis $\mathcal{H}_{0,\mathcal{M}_1,\mathcal{M}_2} : \mathrm{MMD}(P_{\alpha_\star}, P) = \mathrm{MMD}(Q_{\beta_\star}, P)$, based on

$$\mathcal{T}_n(\mathcal{M}_1, \mathcal{M}_2, P) := \sqrt{n}\frac{\widehat{\mathrm{MMD}}^2_{\epsilon_n}(P_{\alpha_n}, P) - \widehat{\mathrm{MMD}}^2_{\epsilon_n}(Q_{\beta_n}, P)}{\hat{\tau}_n}, \tag{6}$$

where $\hat{\tau}_n^2 = \hat{\tau}_n^2(\epsilon_n, P_{\alpha_n}, Q_{\beta_n}, P)$ denotes a natural estimator of the asymptotic variance of $\sqrt{n}\big\{\widehat{\mathrm{MMD}}^2_{\epsilon_n}(P_{\alpha_n}, P) - \widehat{\mathrm{MMD}}^2_{\epsilon_n}(Q_{\beta_n}, P)\big\}$, which is specified in Section 4.1. Moreover, $(\alpha_n)_{n\geq 1}$ (resp. $(\beta_n)_{n\geq 1}$) denotes a sequences of parameters which weakly converges to $\alpha_\star$ (resp. $\beta_\star$) at rate $n^{-1/2}$. We prove that $\mathcal{T}_n(\mathcal{M}_1, \mathcal{M}_2, P) \xrightarrow{\text{law}} \mathcal{N}(0,1)$ under $\mathcal{H}_{0,\mathcal{M}_1,\mathcal{M}_2}$, even if $P = P_{\alpha_\star} = Q_{\beta_\star}$ and regardless of the dependence induced by α_n and β_n.

3. Asymptotic Behavior of MMD-based Specification Tests

This section establishes the asymptotic normality of our test statistic for model specification $\mathcal{T}_n(\mathcal{M}, P)$.

3.1 Fundamental Regularity Conditions on $\mathcal{M}$, k and α_n

Let us formalize our mathematical framework. Hereafter, we will assume that the sample space $\mathcal{S}$ is some topological space equipped with its Borel sigma-algebra. Due to the Moore-Aronszajn Theorem (Aronszajn, 1950), there exists a unique RKHS $\mathcal{H}$ of real-valued functions on $\mathcal{S}$ that is associated with our kernel $k : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$. It can be proven that $\mathcal{H} = \overline{\text{span}\{k(x,\cdot) \,|\, x \in \mathcal{S}\}}$ where the closure is taken w.r.t. the RKHS norm. For instance, when $\mathcal{S} = \mathbb{R}$, the RKHS associated with the popular Gaussian kernel $k(x,y) = \exp(-(x-y)^2/\sigma^2)$ is the space of functions $f : x \mapsto \exp(-x^2/\sigma^2)\sum_{j=0}^{+\infty} v_j x^j$, for some coefficients $(v_j)_{j\geq 0}$ that satisfy $\sum_{j\geq 0} j!\sigma^{2j} v_j^2/2^j < \infty$. The scalar product of two elements $f(x) = \exp(-x^2/\sigma^2)\sum_{j=0}^{+\infty} v_j x^j$ and $g(x) = \exp(-x^2/\sigma^2)\sum_{j=0}^{+\infty} w_j x^j$ in $\mathcal{H}$ is then $< f, g >= \sum_{j\geq 0} j!\sigma^{2j} v_j w_j/2^j$ (Minh, 2010, Theorem 1). See other examples of RKHS $\mathcal{H}$ in Berlinet and Thomas-Agnan (2011, Chapter 7). Let $(X_1, X_2, \ldots)$ denote an i.i.d. sequence drawn from the law P of the DGP. To state our results, we will need several conditions of regularity.

Assumption 1 *The kernel k is a measurable and bounded map from $\mathcal{S} \times \mathcal{S}$ to $\mathbb{R}$. Moreover, it is characteristic: the map $P \mapsto \int_{\mathcal{S}} k(\cdot, x)\, P(\mathrm{d}x)$ from the space of Borel probability measures on $\mathcal{S}$ to $\mathcal{H}$ is injective.*

Note that the integral $\int_{\mathcal{S}} k(\cdot, x)\, P(\mathrm{d}x)$ has to be interpreted as a Bochner integral; see, e.g., Dinculeanu (2000, Chapter 1). Moreover, the boundedness of k implies that the mean embedding is well-defined for any probability measure: see Section 3.1 in Muandet et al. (2020), for instance. Thus, Assumption 1 is sufficient to ensure that $\text{MMD}(P_1, P_2)$ is a valid distance between two probability measures. For example, the Gaussian and Laplace kernels on $\mathbb{R}^d$ satisfy Assumption 1: see Fukumizu et al. (2007) and Sriperumbudur et al. (2010) for a thorough account on kernels satisfying Assumption 1. The characteristic property is key in many applications of the MMD and has been studied in depth in the literature. It is closely related to, but different from, the concept of "universality" (Sriperumbudur et al., 2011; Simon-Gabriel and Schölkopf, 2018). It is required for some kernel measures of conditional dependence (Fukumizu et al., 2007, 2009). Notably, Nishiyama and Fukumizu (2016) stated the characteristicity of kernels defined by pdfs' of symmetric infinitely divisible distributions. Recently, Szabó and Sriperumbudur (2018) studied the characteristic and universal properties of product kernels. See Muandet et al. (2017, Section 3.3.1) and the references therein too.

Next, let us consider $\mathcal{M} := \left\{P_\alpha; \alpha \in \Theta_1\right\}$, a parametric family of probability measures on $\mathcal{S}$.

Assumption 2 *The space Θ_1 is a compact subset of $\mathbb{R}^{p_\alpha}$ and its interior is non-empty. There exists a topological space $\mathcal{U}$ equipped with its Borel sigma-algebra, a random element $U \sim P_U$ in $\mathcal{U}$ and a measurable map $F : \mathcal{U} \times \Theta_1 \to \mathcal{S}$ such that the law of $F(U; \alpha)$ is P_α for any $\alpha \in \Theta_1$. For a given parameter $\alpha_\star$ that belongs to the interior of Θ_1, the map*

$\alpha \mapsto \mathrm{MMD}^2(P_\alpha, P)$ *is twice-continuously differentiable in a neighborhood of* $\alpha_\star$. *Further, the random variable* $\int_{\mathcal{U}} k\big(X, F(u;\alpha_\star)\big)\,\mathrm{d}P_U(u)$ *is not constant a.e. and* $supp(P) \cap supp(P_{\alpha_\star}) \neq \varnothing$.

Essentially, this assumption ensures that the parametrization of $\mathcal{M}$ is sufficiently regular. First, P_α smoothly varies in α w.r.t. the MMD. Second, there exists a convenient way of simulating from the model P_α for every $\alpha \in \Theta_1$. In particular, Assumption 2 ensures that a single i.i.d. sequence of random elements $(U_1, U_2, \dots)$, $U_i \sim P_U$, is sufficient to obtain an i.i.d. sequence $\big(F(U_1;\alpha), F(U_2;\alpha), \dots\big)$ from P_α, for every $\alpha \in \Theta_1$. Thus, from now on, we will assume that we have an i.i.d. sequence $(U_1, U_2, \dots)$ from P_U at hand, which is also assumed to be independent of $(X_1, X_2, \dots)$. Furthermore, as is standard in mathematical statistics, we assume that the two independent i.i.d. sequences are induced by a common abstract probability space $(\Omega, \mathcal{B}, \mathbb{P})$. If not indicated otherwise, expectations $\mathbb{E}[\cdot]$ are always taken w.r.t. $\mathbb{P}$. Moreover, $W_n = o_{\mathbb{P}}(1)$ (resp. $W_n = O_{\mathbb{P}}(1)$) means that a sequence of random elements $(W_n)_{n\geq 1}$ tends to zero in probability (resp. is bounded in probability) in the probability space $(\Omega, \mathcal{B}, \mathbb{P})$. The last condition of Assumption 2 is very mild and of purely technical nature. It ensures that we are not working with constant random variables later.

Every model $\mathcal{M}$ can be defined in terms of the tuple (F, U) by setting $\mathcal{M} := \{P_\alpha := \mathrm{Law}\big(F(U;\alpha)\big), \alpha \in \Theta_1\}$. Defining a model by a class of probability measures in this way is common when working with generative (also called simulation-based) models, which are very popular in machine learning and which often also implicitly appear in classical statistics: see, e.g., Bond-Taylor et al. (2021) for a review of the topic and Dziugaite et al. (2015); Li et al. (2015, 2017); Sutherland et al. (2017); Zhou et al. (2020) for generative models based on the MMD. The map F will therefore be called a generating function of the model $\mathcal{M}$. To illustrate the role of U and F, assume that P_α is a parametric family of probability measures on the real line. Then U could be chosen as a uniformly distributed random variable on $(0,1)$ and $F(\cdot;\alpha)$ could denote the inverse-quantile function of P_α. Nevertheless, there might exist several tuples (F, U) to describe the same model $\mathcal{M}$ and not every representation of $\mathcal{M}$ might satisfy Assumption 2.

At this stage, we have not specified what we mean by "optimal" concerning $\alpha_\star$. Formally, $\alpha_\star$ could be arbitrarily chosen, even if, in practice, this value is most often the minimizer of some distance between the family $(P_\alpha)_{\alpha\in\Theta_1}$ and the true probability measure P. In the latter case, we called $\alpha_\star$ a pseudo-true value (see Section 1). For the moment, we keep the discussion as general as possible: we do not impose that $\alpha_\star$ is a pseudo-true value, even if this is implicitly implied by $\mathcal{H}_{0,\mathcal{M}}$. Note that we neither require to specify the statistical method of inference for $\alpha_\star$ by α_n nor which data is used to calculate α_n (full sample, sample splitting or overlapping).

To simplify the notations in the following, define the functions

$$\begin{aligned} h\big([\boldsymbol{x}, \boldsymbol{u}]_{1:2};\alpha\big) &:= h\big((x_1, F(u_1;\alpha)), (x_2, F(u_2;\alpha))\big) \\ &= k(x_1, x_2) - k\big(x_1, F(u_2;\alpha)\big) - k\big(x_2, F(u_1;\alpha)\big) \\ &\quad + k\big(F(u_1;\alpha), F(u_2;\alpha)\big), \text{ and} \end{aligned} \tag{7}$$

$$\begin{aligned} q\big([\boldsymbol{x}, \boldsymbol{u}]_{1:4};\alpha\big) &:= k\big(x_1, x_3\big) - k\big(x_4, F(u_2;\alpha)\big) - k\big(x_2, F(u_4;\alpha)\big) \\ &\quad + k\big(F(u_1;\alpha), F(u_3;\alpha)\big), \end{aligned} \tag{8}$$

where the arguments x_j and u_k belong to $\mathcal{S}$ and $\mathcal{U}$ respectively. Furthermore, define the family of maps

$$\tilde{h}(x,y;\alpha):\ \alpha \mapsto \mathbb{E}\Big[h\big((x,y),(X,F(U;\alpha))\big)\Big], \tag{9}$$

which is indexed by $(x,y) \in \mathcal{S}\times\mathcal{S}$. The gradient of any map $\alpha \mapsto H(\alpha)$ at some $\bar{\alpha} \in \Theta_1$ will be denoted by $\nabla_\alpha H(\bar{\alpha})$, and $\nabla_{\alpha^\intercal} H(\bar{\alpha})$ denotes its transpose (row) vector. Concerning second order derivatives, $\nabla_{\alpha,\alpha^\intercal} H(\bar{\alpha})$ denotes the Hessian matrix of H evaluated at $\alpha = \bar{\alpha}$.

3.2 Asymptotic Variance Estimation

A key ingredient to obtain the asymptotic normality of $\mathcal{T}_n(\mathcal{M},P)$ will be the choice of a suitable estimator of the asymptotic variance of $\sqrt{n}\widehat{\mathrm{MMD}}^2_{\epsilon_n}(P_{\alpha_n},P)$. In this section, we propose some intuitive estimators of the asymptotic variances of $\sqrt{n}\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P)$ and $\sqrt{n}\widehat{\mathrm{MMD}}^2_q(P_{\alpha_n},P)$, which we later combine to obtain a suitable estimator of the asymptotic variance of $\sqrt{n}\widehat{\mathrm{MMD}}^2_{\epsilon_n}(P_{\alpha_n},P)$. To this purpose, recall that, for any fixed parameter α, it is well-known (Serfling, 1980, p. 192) that the asymptotic variance of $\sqrt{n}\big\{\widehat{\mathrm{MMD}}^2(P_\alpha,P) - \mathrm{MMD}^2(P_\alpha,P)\big\}$ is

$$\sigma^2_\alpha := \mathrm{Var}\Big(2\tilde{h}\big(X,F(U;\alpha);\alpha\big)\Big).$$

The corresponding empirical counterpart of σ^2_α is

$$\tilde{\sigma}^2_\alpha := \frac{4}{n}\sum_{i=1}^n \left\{\frac{1}{n-1}\sum_{\substack{j=1\\ i\neq j}}^n h\Big(\big(X_i,F(U_i;\alpha)\big),\big(X_j,F(U_j;\alpha)\big)\Big) - \widehat{\mathrm{MMD}}^2(P_\alpha,P)\right\}^2.$$

Since our goal is to estimate $\sigma^2_{\alpha_\star}$ for unknown $\alpha_\star$, we replace $\alpha_\star$ with α_n and define an estimator of $\sigma^2_{\alpha_\star}$ by

$$\tilde{\sigma}^2_{\alpha_n} := \frac{4}{n}\sum_{i=1}^n \left\{\frac{1}{n-1}\sum_{\substack{j=1\\ i\neq j}}^n h\Big(\big(X_i,F(U_i;\alpha_n)\big),\big(X_j,F(U_j;\alpha_n)\big)\Big) - \widehat{\mathrm{MMD}}^2(P_{\alpha_n},P)\right\}^2. \tag{10}$$

Similarly, the asymptotic variance of $\sqrt{n}\big\{\widehat{\mathrm{MMD}}^2_q(P_\alpha,P) - \mathrm{MMD}(P_\alpha,P)\big\}$ for any $\alpha \in \Theta_1$ is

$$\sigma^2_{q,\alpha} := \mathrm{Var}\Big(2\sqrt{2}\mathbb{E}_{[\boldsymbol{X},\boldsymbol{U}]_{3:4}}\big[q\big([\boldsymbol{X},\boldsymbol{U}]_{1:4};\alpha\big)\big]\Big).$$

Analogously, this allows to define an estimator of $\sigma^2_{q,\alpha_\star}$ via

$$\tilde{\sigma}^2_{q,\alpha_n} := \frac{16}{n}\sum_{i=1}^{n/2} \left\{\frac{1}{n/2-1}\sum_{\substack{j=1\\ i\neq j}}^{n/2} q\Big([\boldsymbol{X},\boldsymbol{U}]_{2i-1,2i,2j-1,2j};\alpha_n\Big) - \widehat{\mathrm{MMD}}^2(P_{\alpha_n},P)\right\}^2. \tag{11}$$

One should observe that both estimators are always non-negative and of computational complexity $O(n^2)$. Further, $\sigma^2_{\alpha_\star} = 0$ when $P_{\alpha_\star} = P$, but $\sigma^2_{q,\alpha_\star}$ is always strictly positive under Assumption 2, which will later be important in our theoretical derivations. To see this, note that, if $\sigma_{q,\alpha_\star} = 0$ then the random element

$$\mathbb{E}_{[\boldsymbol{X},\boldsymbol{U}]_{3:4}}\big[q\big([\boldsymbol{X},\boldsymbol{U}]_{1:4};\alpha_\star\big)\big] = \mathbb{E}\big[k(X_1,X_3)|X_1\big] - \mathbb{E}\Big[k\big(X_4,F(U_2;\alpha_\star)\big)|U_2\Big] \\ -\mathbb{E}\Big[k\big(X_2,F(U_4;\alpha_\star)\big)|X_2\Big] + \mathbb{E}\Big[k\big(F(U_1;\alpha_\star),F(U_3;\alpha_\star)\big)|U_1\Big] \tag{12}$$

is constant almost surely. Therefore, the four random variables on the r.h.s. of (12) are constant almost surely. In particular, this would imply that $\mathbb{E}\big[k\big(X,F(U;\alpha_\star)\big)|X\big]$ is constant, a situation that has been excluded by Assumption 2.

3.3 Differentiable Generating Functions (Model Specification)

In this section, we treat the case of $\alpha \mapsto F(u;\alpha)$ being twice differentiable for every $u \in \mathrm{supp}(U)$. Even if this assumption may appear relatively demanding (see the discussion in the beginning of Section 3.4 below), the proofs of our results are significantly simpler and intuitive in this case. This is why we choose to first provide our results under the assumption of a smooth generating function, once it is composed with the kernel. Later, we generalize our results to the case of possibly non-differentiable maps.

Assumption 3 *The maps $\alpha \mapsto k\big(x,F(u;\alpha)\big)$ and $\alpha \mapsto k\big(F(u;\alpha),F(\tilde{u};\alpha)\big)$ are twice differentiable for every $x \in \mathcal{S}$ and $u,\tilde{u} \in \mathcal{U}$.*

As a consequence, the maps h and q, as defined by (7) and (8) respectively, are twice differentiable w.r.t. α. Note that Assumption 3 is mainly an assumption on the smoothness of $F(u;\alpha)$, since most of the commonly used kernels k are smooth functions. Further, we require some usual conditions of regularity, expressed in terms of moments. These conditions are not only imposed on h and q but also on the auxiliary function

$$g\big([\boldsymbol{x},\boldsymbol{u}]_{1:3};\alpha\big) := \frac{4}{3}\big\{h([\boldsymbol{x},\boldsymbol{u}]_{1,2};\alpha)h([\boldsymbol{x},\boldsymbol{u}]_{1,3};\alpha) + h([\boldsymbol{x},\boldsymbol{u}]_{2,1};\alpha)h([\boldsymbol{x},\boldsymbol{u}]_{2,3};\alpha) \\ + h([\boldsymbol{x},\boldsymbol{u}]_{3,2};\alpha)h([\boldsymbol{x},\boldsymbol{u}]_{3,1};\alpha)\big\}, \tag{13}$$

which frequently appears in the proofs. For any $\delta > 0$, let $B_\delta(\alpha_\star)$ be an open ball in Θ_1 that is centered at $\alpha_\star$ and whose radius is δ. We require the following conditions of regularity.

Assumption 4 *There exists a $\delta > 0$ s.t. $\mathbb{E}\Big[\sup_{\alpha_1 \in B_\delta(\alpha_\star)} \|\nabla^2_{\alpha,\alpha^\intercal} h\big([\boldsymbol{X},\boldsymbol{U}]_{1:2};\alpha_1\big)\|\Big] < \infty$, and*

$$\mathbb{E}\Big[\sup_{\alpha_1 \in B_\delta(\alpha_\star)} \|\nabla^2_{\alpha,\alpha^\intercal} q\big([\boldsymbol{X},\boldsymbol{U}]_{1:4};\alpha_1\big)\|\Big] + \mathbb{E}\Big[\sup_{\alpha_1 \in B_\delta(\alpha_\star)} \|\nabla^2_{\alpha,\alpha^\intercal} g\big([\boldsymbol{X},\boldsymbol{U}]_{1:3};\alpha_1\big)\|\Big] < \infty.$$

Moreover, $\mathbb{E}\big[\nabla_\alpha h\big([\boldsymbol{X},\boldsymbol{U}]_{1:2};\alpha_\star\big)\big] = \nabla_\alpha \mathbb{E}\big[h\big([\boldsymbol{X},\boldsymbol{U}]_{1:2};\alpha_\star\big)\big]$, $\mathbb{E}\big[\nabla_\alpha g\big([\boldsymbol{X},\boldsymbol{U}]_{1:3};\alpha_\star\big)\big] = \nabla_\alpha \mathbb{E}\big[g\big([\boldsymbol{X},\boldsymbol{U}]_{1:3};\alpha_\star\big)\big]$ and $\mathbb{E}\big[\nabla_\alpha q\big([\boldsymbol{X},\boldsymbol{U}]_{1:4};\alpha_\star\big)\big] = \nabla_\alpha \mathbb{E}\big[q\big([\boldsymbol{X},\boldsymbol{U}]_{1:4};\alpha_\star\big)\big]$.

Under the latter assumptions, Lemma 2 in the appendix shows that $\tilde{\sigma}^2_{\alpha_n} = O_{\mathbb{P}}(n^{-1})$ when $P_{\alpha_\star} = P$, whereas $\tilde{\sigma}^2_{q,\alpha_n} \to \tilde{\sigma}^2_{q,\alpha_\star} > 0$ remains positive. This justifies a model specification test based on a weighted combination of $\sqrt{n}\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P)$ and $\sqrt{n}\widehat{\mathrm{MMD}}^2_q(P_{\alpha_n},P)$, which automatically "switches" between the two statistics, depending on whether or not $P_{\alpha_\star} = P$.

Theorem 1 *Assume $\epsilon_n \to 0$, $\epsilon_n\sqrt{n} \to \infty$ in probability, $\sqrt{n}(\alpha_n - \alpha_\star) = O_{\mathbb{P}}(1)$ and that Assumptions 1-4 are satisfied.*

1. *If $P = P_{\alpha_\star}$, i.e., under $\mathcal{H}_{0,\mathcal{M}}$, we have*

$$\mathcal{T}_n(\mathcal{M}, P) = \sqrt{n}\frac{\widehat{\mathrm{MMD}}^2_{\epsilon_n}(P_{\alpha_n}, P)}{\hat{\sigma}_n} \xrightarrow{law} \mathcal{N}(0,1),$$

   where $\hat{\sigma}_n = \tilde{\sigma}_{\alpha_n} + \epsilon_n \tilde{\sigma}_{q,\alpha_n}$.

2. *If $P \neq P_{\alpha_\star}$, i.e., if $\mathcal{H}_{0,\mathcal{M}}$ is not true, then $\mathcal{T}_n(\mathcal{M}, P)$ tends to infinity in probability.*

The proof is postponed to Section A.2. As a consequence of Theorem 1, a consistent distribution free test of $\mathcal{H}_{0,\mathcal{M}}$ can be conducted with the test statistic $\mathcal{T}_n(\mathcal{M}, P)$ (see Algorithm 1 below).

Remark 1 *Note that Theorem 1 obviously covers the case of simple zero assumptions, i.e., testing $\mathcal{H}_0 : P_1 = P$ for some given probability P_1 and with the test statistic (4). Our theory directly applies by defining Θ_1 as the singleton $\{\alpha_\star\}$, and setting $P_{\alpha_\star} = P_1$. The technical assumptions 2-4 are no longer required in this case, since the sequence (α_n) becomes constant and it is no longer necessary to differentiate the kernels h and q.*

Remark 2 *In Theorem 1, it is possible to replace the denominator $\hat{\sigma}_n = \tilde{\sigma}_{\alpha_n} + \epsilon_n \tilde{\sigma}_{q,\alpha_n}$ by $\left(\tilde{\sigma}^2_{\alpha_n} + \epsilon_n^2 \tilde{\sigma}^2_{q,\alpha_n}\right)^{1/2}$ since both quantities are asymptotically equivalent under our assumptions. Moreover, to lighten our theoretical developments, we have neglected the covariance between $\widehat{\mathrm{MMD}}^2(P_{\alpha_n}, P)$ and $\widehat{\mathrm{MMD}}^2_q(P_{\alpha_n}, P)$ as this covariance multiplied by ϵ_n is always asymptotically negligible compared to $\tilde{\sigma}^2_{\alpha_n} + \epsilon_n^2 \tilde{\sigma}^2_{q,\alpha_n}$ (invoke the Cauchy-Schwarz inequality in the degenerate case, as $\tilde{\sigma}_{\alpha_n} = o_{\mathbb{P}}(\epsilon_n)$).*

3.4 Non-Differentiable Generating Function (Model Specification)

The assumption of differentiability of $\alpha \mapsto F(U;\alpha)$ may be considered as relatively strong in many practical applications in statistics and machine learning. For example, let $F(\cdot;\alpha)$ denote a deep neural network with the ReLu activation function and a parameter vector α and let U be uniformly distributed on $[0,1]$. Then, defining $P_\alpha := \mathrm{law}(F(U;\alpha))$ yields a universal approximator of any probability distribution P in terms of the MMD: see Yang et al. (2022, Theorem 2.8). Obviously, $\alpha \mapsto F(u;\alpha)$ is not differentiable for any u, and the results from Section 3.3 cannot be applied. In this section, we show that our test may still be applied even if the generating function is not differentiable, imposing some regularity conditions on $\mathbb{E}\big[k\big(F(U;\alpha),\cdot\big)\big]$ which are detailed in Appendix A.3.

Since we cannot apply a Taylor expansion w.r.t. α to $\mathcal{T}_n(\mathcal{M}, P)$ when $\alpha \mapsto F(U;\alpha)$ is not differentiable, we need more sophisticated tools than in Section 3.3 to derive the asymptotic normality of $\mathcal{T}_n(\mathcal{M}, P)$. Here, we rely on the framework of empirical U-processes introduced by Arcones and Giné (1993, 1994). See Appendix A.3 for technical details. Under our proposed assumptions, we obtain the limiting law of the test statistic defined in (5) when dealing with non-differentiable generating functions, which can be used to test the null hypothesis $\mathcal{H}_{0,\mathcal{M}}$. To be short, we recover the results of Theorem 1.

Theorem 2 *Let Assumptions 1-2 and 6-8 in Appendix A.3 hold. If $\epsilon_n \to 0$, $\epsilon_n\sqrt{n} \to \infty$ in probability and $\sqrt{n}(\alpha_n - \alpha_\star) = O_{\mathbb{P}}(1)$, then the conclusions of Theorem 1 apply.*

The proof can be found in Section A.3 in the appendix. Since Assumptions 6-8 are quite abstract, we further illustrate the practical relevance of the proposed framework by verifying that a ReLu-type generative neural network satisfies Assumptions 1-2 and 6-8 in Appendix C. We summarize our proposed model specification test in Algorithm 1.

Algorithm 1: MMD-based test of $\mathcal{H}_{0,\mathcal{M}} : \mathrm{MMD}(P_{\alpha_\star}, P) = 0$

Requirements: I.i.d. sample $(X_i)_{1\le i\le n}$ from P, generative model $F(U;\alpha) \sim P_\alpha$, estimator α_n of $\alpha_\star$, tuning parameter ϵ_n and confidence level γ.

1 Sample $\big(F(U_i,\alpha_n)\big)_{1\le i\le n}$, where $(U_i)_{1\le i\le n} \overset{i.i.d.}{\sim} P_U$;

2 Compute $\mathcal{T}_n(\mathcal{M},P) = \sqrt{n}\widehat{\mathrm{MMD}}^2_{\epsilon_n}(P_{\alpha_n},P)/(\tilde\sigma_{\alpha_n} + \epsilon_n\tilde\sigma_{q,\alpha_n})$ according to (3), (10) and (11) ;

3 Reject $\mathrm{MMD}(P_{\alpha_\star}, P) = 0$ when $\big|\mathcal{T}_n(\mathcal{M},P)\big| > \Phi^{-1}(1-\gamma/2)$; otherwise, accept.

---

4. Asymptotic Behavior of MMD-based Tests for Model Comparison

Let us specify the mathematical framework that is required to prove the asymptotic normality of the test statistic for model comparison introduced in (6). First we introduce additional notation. For some sequences of random elements $(X_1, X_2, \ldots)$, $(U_1, U_2, \ldots)$ and $(V_1, V_2, \ldots)$, we denote $[\boldsymbol{X},\boldsymbol{U},\boldsymbol{V}]_{i:j} := \big((X_i,U_i,V_i),(X_{i+1},U_{i+1},V_{i+1}),\ldots,(X_j,U_j,V_j)\big)$ for $1 \le i < j \le n$. Similarly, we denote $[\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{i:j} := \big((x_i,u_i,v_i),(x_{i+1},u_{i+1},v_{i+1}),\ldots,(x_j,u_j,v_j)\big)$. In the same manner, $[\boldsymbol{X},\boldsymbol{U},\boldsymbol{V}]_{i_1,i_2,\ldots,i_k}$ and $[\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{i_1,i_2,\ldots,i_k}$ are defined, where the index set indicates the components of the concatenated variables.

4.1 Regularity Assumptions and Asymptotic Variance Estimation

Consider two competing parametric models $\mathcal{M}_1 = \{P_\alpha; \alpha \in \Theta_1\}$ and $\mathcal{M}_2 = \{Q_\beta; \beta \in \Theta_2\}$ for the DGP. The goal is to evaluate whether or not one of the models is closer to the true law P of the data than the other in terms of the MMD. Our null hypothesis is then written $\mathcal{H}_{0,\mathcal{M}_1,\mathcal{M}_2} : \mathrm{MMD}(P_{\alpha_\star}, P) = \mathrm{MMD}(Q_{\beta_\star}, P)$.

For convenience, the asymptotic behavior of our test statistic $\mathcal{T}_n(\mathcal{M}_1,\mathcal{M}_2,P)$ will be stated under the assumption that the optimal parameters are pseudo-true values for the MMD, i.e., $\alpha_\star := \mathrm{argmin}_{\alpha\in\Theta_1} \mathrm{MMD}(P_\alpha,P)$ and $\beta_\star := \mathrm{argmin}_{\beta\in\Theta_2} \mathrm{MMD}(Q_\beta,P)$. This constraint is due to the fact $\mathcal{H}_{0,\mathcal{M}_1,\mathcal{M}_2}$ can be satisfied even though $P_{\alpha_\star} \ne P \ne Q_{\beta_\star}$, which would then introduce additional randomness in our test statistic. In this case, some terms of the form $\nabla_\alpha \mathrm{MMD}(P_\alpha,P)|_{\alpha=\alpha_\star} \ne 0 \ne \nabla_\beta \mathrm{MMD}(Q_\beta,P)|_{\beta=\beta_\star}$ would appear in the asymptotic variance, adding significant complications in the estimation procedure, which is why we have refrained from investigating this case. To estimate $\alpha_\star$ (resp. $\beta_\star$), we could set $\alpha_n \in \mathrm{argmin}_{\alpha\in\Theta_1} \widehat{\mathrm{MMD}}^2(P_\alpha,P)$ (resp. $\beta_n \in \mathrm{argmin}_{\beta\in\Theta_2} \widehat{\mathrm{MMD}}^2(Q_\beta,P)$), which would yield consistent and asymptotically normal estimators under some regularity conditions (Briol et al., 2019). Nonetheless, this is not mandatory. Thus, the choice of α_n and β_n will remain unspecified hereafter.

Similarly to the existence of a random variable U and a generating function $F(\cdot;\cdot)$ for the model $\mathcal{M}_1$, we assume the existence of a random variable V in some topological space $\mathcal{V}$ and a generating function $G(\cdot;\cdot) : \mathcal{V} \times \Theta_2 \to \mathcal{S}$ of $\mathcal{M}_2$ such that $G(V;\beta) \sim Q_\beta$ for every $\beta \in \Theta_2$. As in the previous section, we assume that we have access to an i.i.d. sequence $(V_1, V_2, \ldots)$ from P_V, which is also independent of $(X_1, X_2, \ldots)$. We also assume the random vectors (X_i, U_i, V_i), $i \in \{1, \ldots, n\}$, are independently drawn and that they are defined on the same abstract probability space $(\Omega, \mathcal{B}, \mathbb{P})$.

In terms of notations and to distinguish quantities that are related to either model $\mathcal{M}_1$ or $\mathcal{M}_2$, we will use the same notation as in Section 3 but an upper index $\cdot^{(1)}$ (resp. $\cdot^{(2)}$) to refer to a quantity related to $\mathcal{M}_1$ (resp. $\mathcal{M}_2$). For instance, the map h introduced in (7) will be denoted as $h^{(1)}\big([\boldsymbol{x},\boldsymbol{u}]_{1:2};\alpha\big) := h\big((x_1, F(u_1;\alpha)), (x_2, F(u_2;\alpha))\big)$ when referring to $\mathcal{M}_1$, whereas we will denote $h^{(2)}\big([\boldsymbol{x},\boldsymbol{v}]_{1:2};\beta\big) := h\big((x_1, G(v_1;\beta)), (x_2, G(v_2;\beta))\big)$ when referring to $\mathcal{M}_2$. To distinguish between the two parametric models, the letter u (resp. v) will be reserved for the first model (resp. second model).

In the following, we will mimic the ideas of Section 3. Recall that, in the definition of $\mathcal{T}_n(\mathcal{M}_1, \mathcal{M}_2, P)$ in (6), we have not yet specified the estimator $\hat{\tau}_n^2$ of the asymptotic variance of $\sqrt{n}\big\{\widehat{\mathrm{MMD}}^2_{\epsilon_n}(P_{\alpha_n}, P) - \widehat{\mathrm{MMD}}^2_{\epsilon_n}(Q_{\beta_n}, P)\big\}$. Again, we will use an estimator of the form $\hat{\tau}_n = \hat{\tau}_1 + \epsilon_n \hat{\tau}_2$, where $\hat{\tau}_1^2$ (resp. $\hat{\tau}_2^2$) denotes an estimator of the asymptotic variance of $\sqrt{n}\big\{\widehat{\mathrm{MMD}}^2(P_{\alpha_n}, P) - \widehat{\mathrm{MMD}}^2(Q_{\beta_n}, P)\big\}$ (resp. of $\sqrt{n}\big\{\widehat{\mathrm{MMD}}^2_q(P_{\alpha_n}, P) - \widehat{\mathrm{MMD}}^2_q(Q_{\beta_n}, P)\big\}$). To this aim and in accordance with the definitions of σ^2_α and $\sigma^2_{q,\alpha}$ in Section 3.1 define

$$\sigma^2_{\alpha,\beta} := \mathrm{Var}\Big(2\,\mathbb{E}_{X_2,U_2,V_2}\big[h^{(1)}\big([\boldsymbol{X},\boldsymbol{U}]_{1:2};\alpha\big) - h^{(2)}\big([\boldsymbol{X},\boldsymbol{V}]_{1:2};\beta\big)\big]\Big), \text{ and} \tag{14}$$

$$\sigma^2_{q,\alpha,\beta} := \mathrm{Var}\Big(2\sqrt{2}\,\mathbb{E}_{[\boldsymbol{X},\boldsymbol{U},\boldsymbol{V}]_{3:4}}\big[q^{(1)}\big([\boldsymbol{X},\boldsymbol{U}]_{1:4};\alpha\big) - q^{(2)}\big([\boldsymbol{X},\boldsymbol{V}]_{1:4};\beta\big)\big]\Big).$$

Defining

$$\begin{aligned} h\big([\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{1:2};\alpha,\beta\big) &:= h^{(1)}\big([\boldsymbol{x},\boldsymbol{u}]_{1:2};\alpha\big) - h^{(2)}\big([\boldsymbol{x},\boldsymbol{v}]_{1:2};\beta\big), \text{ and} \\ q\big([\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{1:4};\alpha,\beta\big) &:= q^{(1)}\big([\boldsymbol{x},\boldsymbol{u}]_{1:4};\alpha\big) - q^{(2)}\big([\boldsymbol{x},\boldsymbol{v}]_{1:4};\beta\big) \end{aligned} \tag{15}$$

we can introduce their corresponding estimators, for a given tuple (α,β), via

$$\tilde{\sigma}^2_{\alpha,\beta} := \frac{4}{n}\sum_{i=1}^n \left\{ \frac{1}{n-1}\sum_{\substack{j=1\\ i\neq j}}^n h\Big([\boldsymbol{X},\boldsymbol{U},\boldsymbol{V}]_{i:j};\alpha,\beta\Big) - \Big(\widehat{\mathrm{MMD}}^2(P_\alpha,P) - \widehat{\mathrm{MMD}}^2(Q_\beta,P)\Big)\right\}^2, \text{ and} \tag{16}$$

$$\tilde{\sigma}^2_{q,\alpha,\beta} := \frac{16}{n}\sum_{i=1}^{n/2}\left\{\frac{1}{n/2-1}\sum_{\substack{j=1\\ i\neq j}}^{n/2} q\Big([\boldsymbol{X},\boldsymbol{U},\boldsymbol{V}]_{2i-1,2i,2j-1,2j};\alpha,\beta\Big) - \Big(\widehat{\mathrm{MMD}}^2(P_\alpha,P) - \widehat{\mathrm{MMD}}^2(Q_\beta,P)\Big)\right\}^2. \tag{17}$$

Again, $\tilde{\sigma}^2_{\alpha_n,\beta_n}$ and $\tilde{\sigma}^2_{q,\alpha_n,\beta_n}$ are used as estimators of $\sigma^2_{\alpha_\star,\beta_\star}$ and $\sigma^2_{q,\alpha_\star,\beta_\star}$ respectively.

4.2 Differentiable Generating Functions (Model Comparison)

We are able to derive a test for model comparison that is never degenerate when $(U_i)_{i\geq 1}$ and $(V_i)_{i\geq 1}$ are independent. We need a slight extension of Assumption 4 to prove the results.

Assumption 5 *We have supp$(P_{\alpha_\star}) \cap$ supp$(P) \cap$ supp$(Q_{\beta_\star}) \neq \varnothing$ and there exists $\delta > 0$ s.t.*

$$\mathbb{E}\Bigg[\sup_{(\alpha_1,\beta_1)\in B_\delta((\alpha_\star,\beta_\star))} \|\nabla^2_{(\alpha,\beta),(\alpha,\beta)^\top} g\big([\boldsymbol{X},\boldsymbol{U},\boldsymbol{V}]_{1:3};(\alpha_1,\beta_1)\big)\|\Bigg] < \infty$$

$$\text{and } \nabla_{(\alpha,\beta)}\mathbb{E}\big[g\big([\boldsymbol{X},\boldsymbol{U},\boldsymbol{V}]_{1:3};(\alpha_\star,\beta_\star)\big)\big] = \mathbb{E}\big[\nabla_{(\alpha,\beta)} g\big([\boldsymbol{X},\boldsymbol{U},\boldsymbol{V}]_{1:3};(\alpha_\star,\beta_\star)\big)\big].$$

Theorem 3 *Assume that $\epsilon_n \to 0$, $\epsilon_n\sqrt{n} \to \infty$ in probability, $\alpha_\star \in \operatorname{argmin}_{\alpha\in\Theta_1} \mathrm{MMD}^2(P_\alpha, P)$ and $\beta_\star \in \operatorname{argmin}_{\beta\in\Theta_2} \mathrm{MMD}^2(Q_\beta, P)$, $\sqrt{n}(\alpha_n - \alpha_\star) = O_{\mathbb{P}}(1)$ and $\sqrt{n}(\beta_n - \beta_\star) = O_{\mathbb{P}}(1)$, the samples $(U_i)_{i\geq 1}$ and $(V_i)_{i\geq 1}$ are independent and that Assumptions 1-5 are satisfied by the competing models $\mathcal{M}_1$ and $\mathcal{M}_2$.*

1. *Under $\mathcal{H}_{0,\mathcal{M}_1,\mathcal{M}_2} : \mathrm{MMD}(P_{\alpha_\star}, P) = \mathrm{MMD}(Q_{\beta_\star}, P)$, we have*

$$\mathcal{T}_n(\mathcal{M}_1,\mathcal{M}_2,P) = \sqrt{n}\frac{\widehat{\mathrm{MMD}}^2_{\epsilon_n}(P_{\alpha_n},P) - \widehat{\mathrm{MMD}}^2_{\epsilon_n}(Q_{\beta_n},P)}{\hat{\tau}_n} \xrightarrow{law} \mathcal{N}(0,1),$$

   where $\hat{\tau}_n = \tilde{\sigma}_{\alpha_n,\beta_n} + \epsilon_n\tilde{\sigma}_{q,\alpha_n,\beta_n}$.

2. *If $\mathrm{MMD}(P_{\alpha_\star}, P) > \mathrm{MMD}(Q_{\beta_\star}, P)$, then*

$$\sqrt{n}\frac{\widehat{\mathrm{MMD}}^2_{\epsilon_n}(P_{\alpha_n},P) - \widehat{\mathrm{MMD}}^2_{\epsilon_n}(Q_{\beta_n},P)}{\hat{\tau}_n} \to +\infty \text{ in probability.}$$

3. *If $\mathrm{MMD}(P_{\alpha_\star}, P) < \mathrm{MMD}(Q_{\beta_\star}, P)$, then*

$$\sqrt{n}\frac{\widehat{\mathrm{MMD}}^2_{\epsilon_n}(P_{\alpha_n},P) - \widehat{\mathrm{MMD}}^2_{\epsilon_n}(Q_{\beta_n},P)}{\hat{\tau}_n} \to -\infty \text{ in probability.}$$

The proof can be found in Appendix A.4. As in Remark 1, the latter theorem also covers the case of known parameters $\alpha_\star$ and $\beta_\star$, i.e., the case when Θ_1 and Θ_2 are singletons, solely requiring Assumption 1.

4.3 Non-Differentiable Generating Functions (Model Comparison)

When $\alpha \mapsto F(u;\alpha)$ or $\beta \mapsto G(v;\beta)$ is not twice differentiable, we rely on similar techniques as in Section 3.4 to deduce the asymptotic normality of $\mathcal{T}_n(\mathcal{M}_1,\mathcal{M}_2,P)$. Some technical conditions related to empirical U-processes are required: see Section A.5 in the appendix.

Theorem 4 *Assume $\epsilon_n \to 0$, $\epsilon_n\sqrt{n} \to \infty$ in probability, $(U_i)_{i\geq 1}$ and $(V_i)_{i\geq 1}$ are independent, $\alpha_\star \in \operatorname{argmin}_{\alpha\in\Theta_1} \mathrm{MMD}(P_\alpha, P)$ and $\beta_\star \in \operatorname{argmin}_{\beta\in\Theta_2} \mathrm{MMD}(Q_\beta, P)$, $\sqrt{n}(\alpha_n - \alpha_\star) = O_{\mathbb{P}}(1)$ and $\sqrt{n}(\beta_n - \beta_\star) = O_{\mathbb{P}}(1)$, and that Assumptions 1-2 and 10-11 in Appendix A.5 are satisfied for the two competing models $\mathcal{M}_1$ and $\mathcal{M}_2$. Then the conclusions of Theorem 3 apply.*

We summarize our proposed model selection test in Algorithm 2.

Algorithm 2: MMD based test of $\mathcal{H}_{0,\mathcal{M}_1,\mathcal{M}_2} : \mathrm{MMD}(P_{\alpha_\star}, P) = \mathrm{MMD}(Q_{\beta_\star}, P)$

Requirements: I.i.d. sample $(X_i)_{1\le i\le n}$ from P, generative models $F(U;\alpha) \sim P_\alpha$ and $G(V,\beta) \sim Q_\beta$, estimator α_n of $\mathrm{argmin}_{\alpha\in\Theta_1} \mathrm{MMD}(P_\alpha, P)$, estimator β_n of $\mathrm{argmin}_{\beta\in\Theta_2} \mathrm{MMD}(Q_\beta, P)$, tuning parameter ϵ_n and confidence level γ.

1 Sample $\big(F(U_i,\alpha_n)\big)_{1\le i\le n}$ and $\big(G(V_i,\beta_n)\big)_{1\le i\le n}$, where $(U_i)_{1\le i\le n} \overset{i.i.d.}{\sim} P_U$ and $(V_i)_{1\le i\le n} \overset{i.i.d.}{\sim} P_V$ are independent;

2 Compute $\mathcal{T}_n(\mathcal{M}_1,\mathcal{M}_2,P) =$ $\sqrt{n}\big\{\widehat{\mathrm{MMD}}^2_{\epsilon_n}(P_{\alpha_n},P) - \widehat{\mathrm{MMD}}^2_{\epsilon_n}(Q_{\beta_n},P)\big\}/(\tilde\sigma_{\alpha_n,\beta_n} + \epsilon_n \tilde\sigma_{q,\alpha_n,\beta_n})$according to (3), (16) and (17);

3 Reject $\mathrm{MMD}(P_{\alpha_\star},P) = \mathrm{MMD}(Q_{\beta_\star},P)$ when $\big|\mathcal{T}_n(\mathcal{M}_1,\mathcal{M}_2,P)\big| > \Phi^{-1}(1-\gamma/2)$; otherwise, accept.

5. Simulation Study

The simulation study investigates the finite sample performance of our tests for model specification and model selection.

5.1 Monte Carlo Study for Model Specification

To investigate the performance of the model specification test based on the test statistic $\mathcal{T}_n(\mathcal{M},P)$, let us generalize the example from Section 2.1 to an arbitrary dimension p, inspired by the toy example from Gretton et al. (2012). Consider a p−dimensional random vector $X \sim P = \mathcal{N}(0, I_p)$, where P still denotes the law of the DGP and I_p is the p−dimensional identity matrix.

As a first example, the model $\mathcal{M}$ for the law of X is defined by

$$Y(\alpha) = Y + \alpha \sim P_\alpha, \text{ with } Y \sim \mathcal{N}\big(0, \sigma^2 I_p\big)$$

for some known variance σ^2, and $\alpha := (\alpha_1,\dots,\alpha_p)$ is a p-dimensional vector to be estimated. For every σ^2, the "optimal" parameter is $\alpha_\star = 0$. Moreover, $P = P_{\alpha_\star}$ if $\sigma^2 = 1$. We will vary the standard deviation σ of the competing models by setting $\sigma \in \{1.0, 1.1, 1.2, 1.3, 1.4\}$. Furthermore, consider a p−dimensional Gaussian kernel $k(X_1,X_2) = e^{-\|X_1-X_2\|_2^2/p}$. In the literature (e.g., Gretton et al., 2012), the exponent of the Gaussian kernel is often normalized by an expression containing the empirical median. Since an influence of this estimation step on the asymptotic distribution should also be investigated in the future, we do not consider this type of normalization.

As in Section 2.1, we estimate α by the empirical mean of $(X_1,\dots,X_n)$, an i.i.d. sample from P, i.e., $\alpha_n = n^{-1}\sum_{i=1}^n X_i$. We generate by simulation an i.i.d. sample $(Y_1,\dots,Y_n)$ from $\mathcal{N}(0,\sigma^2 I_p)$ to build the sample $\big(Y_1(\alpha_n),\dots,Y_n(\alpha_n)\big)$ from P_{α_n}. Note that, given α_n, the quantities $Y_k(\alpha_n)$ are mutually independent, but not unconditionally. In this simulation study, we set the number of Monte Carlo replications to 1000, i.e., we generate 1000 independent replications of the test statistics $\mathcal{T}_n(\mathcal{M},P)$ and report the empirical level/power of a test of $\mathcal{H}_{0,\mathcal{M}}$. As a comparison, we also provide the empirical level/power of a test of $\mathcal{H}_{0,\mathcal{M}}$ which is solely based on $\sqrt{n}\widehat{\mathrm{MMD}}^2_q(P_{\alpha_n},P)$.

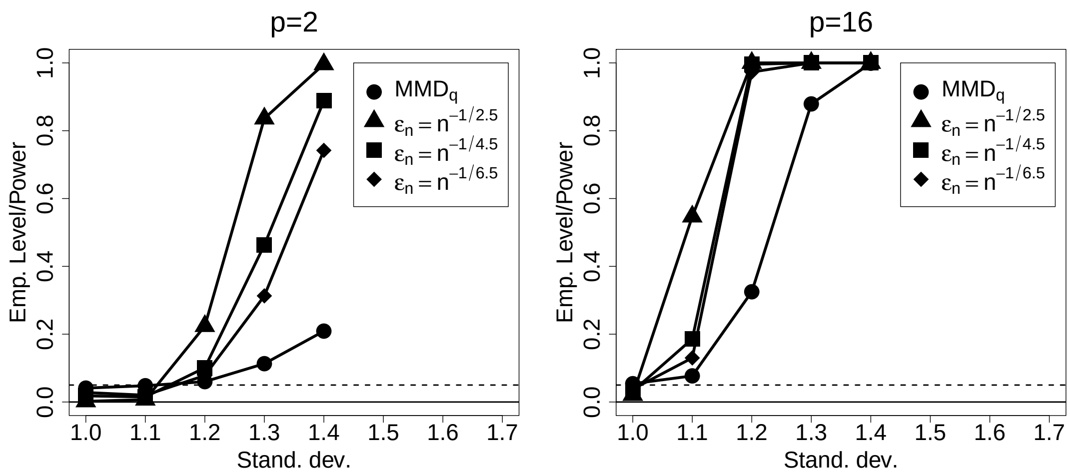

Figure 1: Empirical level and power of the tests based on $\mathcal{T}_n(\mathcal{M}, P)$ and $\sqrt{n}\widehat{\mathrm{MMD}}_q^2(P_{\alpha_n}, P)$ for dimensions $p = 2$ (left) and $p = 16$ (right), a sample size $n = 500$, as well as for different choices of ϵ_n (see the legend) and varying standard deviation. The rejection probabilities are estimated using 1000 replications of the tests based on samples of size n. The black dashed line indicates the significance level 0.05.

For the two tests based on the statistics $\mathcal{T}_n(\mathcal{M}, P)$ and $\sqrt{n}\widehat{\mathrm{MMD}}_q^2(P_{\alpha_n}, P)$ with a level 5%, Figure 1 shows the empirical proportion of rejections of the null hypothesis $\mathcal{H}_{0,\mathcal{M}}$ for dimensions $p \in \{2, 16\}$, sample size $n = 500$ as well as for different choices of $\epsilon_n \in \{n^{-1/2.5}, n^{-1/4.5}, n^{-1/6.5}\}$. In Figure 1, we restrict ourselves to this moderate sample size in order to empirically illustrate an influence of a choice of the weights ϵ_n on the performance of the proposed tests. Note that the empirical proportions of rejections of $\mathcal{H}_{0,\mathcal{M}}$ when $\sigma = 1.0$ are the empirical levels of these tests. When $\sigma > 1.0$, they are their empirical powers.

First of all, we observe in Figure 1 that all tests keep their empirical level sufficiently well for all considered ϵ_n. Furthermore, if the convergence rate of ϵ_n to zero is decreasing, then the empirical power of the test based on $\mathcal{T}_n(\mathcal{M}, P)$ is also decreasing, confirming our intuition that ϵ_n has to tend to zero with n. However, the power is always higher than the empirical power of the test solely based on $\sqrt{n}\widehat{\mathrm{MMD}}_q^2(P_{\alpha_n}, P)$. For any considered tuning parameter ϵ_n, the test based on $\mathcal{T}_n(\mathcal{M}, P)$ always outperforms the test based on $\sqrt{n}\widehat{\mathrm{MMD}}_q^2(P_{\alpha_n}, P)$ for the considered sample size $n = 500$ and all dimensions, confirming the relevance of $\mathcal{T}_n(\mathcal{M}, P)$. In all our experiments, we have observed that the empirical power of all tests is increasing with increasing sample sizes and also for increasing dimension. In the sequel, we fix $\epsilon_n = n^{-1/2.5}$ since this choice empirically yields the highest powers of the proposed MMD specification test.

As a second example, define the family of competing models for the law of X as

$$Y(\sigma) = \alpha_0 \mathbf{1} + diag(\sigma_1, \dots, \sigma_p) Y \sim P_\sigma, ;\ Y \sim \mathcal{N}(0, I_p)$$

for some pre-specified marginal mean $\alpha_0 \in \mathbb{R}$, where the marginal standard deviations are $\sigma_1, \dots, \sigma_p$. We set $\sigma := (\sigma_1, \dots, \sigma_p)$ and $\mathbf{1} = (1, \dots, 1)$. If we fix $\alpha_0 = 0$, then the "optimal" parameters are $\sigma_1^* = \dots = \sigma_p^* = 1$ and $P = P_{\sigma^*}$, where $\sigma^* = (\sigma_1^*, \dots, \sigma_p^*)$. Now, we vary the

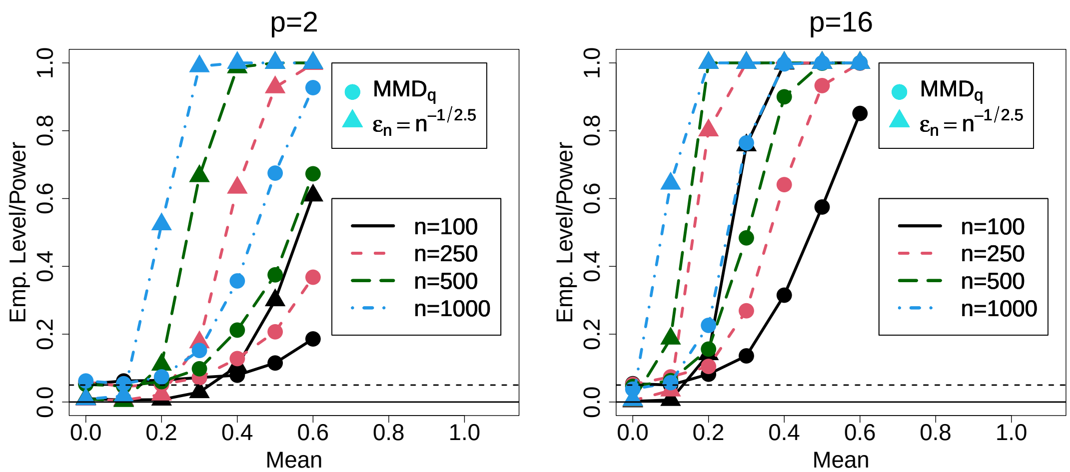

Figure 2: Empirical level and power of the tests based on $\mathcal{T}_n(\mathcal{M}, P)$ and $\sqrt{n}\widehat{\mathrm{MMD}}_q^2(P_{\sigma_n}, P)$ for dimensions $p = 2$ (left) and $p = 16$ (right) as well as for sample sizes $n = 100, 250, 500, 1000$ (see the legend), $\epsilon_n = n^{-1/2.5}$ and varying mean. The rejection probabilities are estimated using 1000 samples of size n. The black dashed line indicates the significance level 0.05.

mean α_0 of the competing model $Y(\sigma)$ by setting $\alpha_0 \in \{0, 0.1, 0.2, \dots, 0.6\}$. As in the previous example, we consider the p–dimensional Gaussian kernel $k(X_1, X_2) = e^{-\|X_1 - X_2\|_2^2/p}$ and set the significance level at 0.05. Furthermore, we estimate σ_j by the empirical standard deviation of the j–th marginal i.i.d. sample from P, namely $\sigma_{j,n}^2 = n^{-1}\sum_{i=1}^n (X_{ij} - \overline{X}_{ij})^2$, and set $\sigma_n := (\sigma_{1,n}, \dots, \sigma_{p,n})$. Thus, we use the two samples $(X_1, \dots, X_n)$ and $\big(Y_1(\sigma_n), \dots, Y_n(\sigma_n)\big)$ from P and P_{σ_n} to test the null hypothesis $\mathcal{T}_n(\mathcal{M}, P)$. In the simulation study, the number of Monte Carlo replications is again 1000 and we report the empirical level/power of a test of $\mathcal{H}_{0,\mathcal{M}}$.

For the two tests based on $\mathcal{T}_n(\mathcal{M}, P)$ and $\sqrt{n}\widehat{\mathrm{MMD}}_q^2(P_{\sigma_n}, P)$, Figure 2 shows the empirical proportion of rejections of the null hypothesis $\mathcal{H}_{0,\mathcal{M}}$ for dimensions $p \in \{2, 16\}$, sample sizes $n \in \{100, 250, 500, 1000\}$ and $\epsilon_n = n^{-1/2.5}$. Note that the empirical proportions of rejections of $\mathcal{H}_{0,\mathcal{M}}$ for the case $\alpha_0 = 0$ are the empirical levels of the tests. When $\alpha_0 > 0$, they are their empirical powers. All considered tests keep their empirical level reasonably well and their power increases with an increasing sample size. Again, the tests based on $\mathcal{T}_n(\mathcal{M}, P)$ are always more powerful than the test based on $\sqrt{n}\widehat{\mathrm{MMD}}_q^2(P_{\sigma_n}, P)$. In this framework, note the poor power of our MMD specification tests for a small sample size $(n = 100)$ and a small dimension $(d = 2)$.

5.2 Monte Carlo Study for Model Comparison

In the third example, we focus on model comparison by considering two competing parametric models $\mathcal{M}_1$ and $\mathcal{M}_2$ for the law of X. Assume that the dimension p is even. The first model $\mathcal{M}_1$ is defined by

$$Y(\alpha) = Y + \alpha \sim P_\alpha, ; \; Y \sim \mathcal{N}\big(0, diag(1, \dots, 1, \sigma^2, \dots, \sigma^2)\big),$$

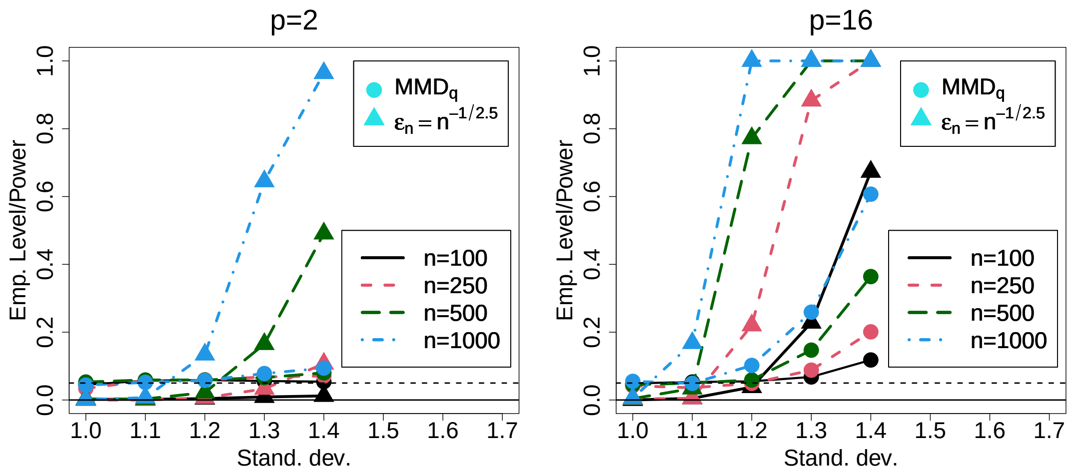

Figure 3: Degenerate case for comparison of two models: Empirical level and power of the tests based on $\mathcal{T}_n(\mathcal{M}_1, \mathcal{M}_2, P)$ and $\sqrt{n}\big(\widehat{\mathrm{MMD}}_q^2(P_{\alpha_n}, P) - \widehat{\mathrm{MMD}}_q^2(Q_{\beta_n}, P)\big)$ for dimensions $p = 2$ (left) and $p = 16$ (right) as well as for sample sizes $n = 100, 250, 500, 1000$ (see the legend), $\epsilon_n = n^{-1/2.5}$ and varying standard deviation σ in Model $\mathcal{M}_1$. Model $\mathcal{M}_2$ coincides with the true model ($\beta = 0$). The rejection probabilities are estimated using 1000 replications of the tests based on samples of size n. The black dashed line indicates the significance level 0.05.

for some pre-specified variance σ^2, where $\alpha = (\alpha_1, \dots, \alpha_p)$. Thus, the first $p/2$ margins of $Y(\alpha)$ have variance 1 and the remaining $p/2$ margins have variance σ^2. If $\sigma^2 = 1$, the model $\mathcal{M}_1$ coincides with the true model when α equals the "optimal" parameter $\alpha_\star = 0$. The second model $\mathcal{M}_2$ is defined by

$$Z(\beta) = Z + \beta \sim Q_\beta \, ; \; Z \sim \mathcal{N}(0, I_p) \, ,$$

where $\beta = (\beta_1, \dots, \beta_p)$. If $\beta = 0$, the model $\mathcal{M}_2$ also coincides with the law of the DGP. Therefore, we may be in the degenerate situation, when the two competing models with optimal parameters coincide with the law of the DGP. As in the first example, we vary the standard deviation σ by setting $\sigma \in \{1.0, 1.1, 1.2, 1.3, 1.4\}$. Further, we estimate α and β by the empirical mean of the i.i.d. sample from P, $\alpha_n = \beta_n = n^{-1} \sum_{i=1}^n X_i$. Then, we independently generate the two samples $\big(Y_1(\alpha_n), \dots, Y_n(\alpha_n)\big)$ and $\big(Z_1(\beta_n), \dots, Z_n(\beta_n)\big)$ to test the null hypothesis $\mathcal{H}_{0,\mathcal{M}_1,\mathcal{M}_2} : \mathrm{MMD}(P_{\alpha_\star}, P) = \mathrm{MMD}(Q_{\beta_\star}, P)$. In the simulation study, we set the number of Monte Carlo replications to 1000, i.e., we generate 1000 independent replications of the test statistic $\mathcal{T}_n(\mathcal{M}_1, \mathcal{M}_2, P)$ and report the empirical level/power of a test of $\mathcal{H}_{0,\mathcal{M}_1,\mathcal{M}_2}$. As a comparison, we also provide the level/power of a test of $\mathcal{H}_{0,\mathcal{M}_1,\mathcal{M}_2}$ which is solely based on $\sqrt{n}\big(\widehat{\mathrm{MMD}}_q^2(P_{\alpha_n}, P) - \widehat{\mathrm{MMD}}_q^2(Q_{\beta_n}, P)\big)$. For the two considered tests, Figure 3 shows the empirical proportion of rejections of the null hypothesis $\mathcal{H}_{0,\mathcal{M}_1,\mathcal{M}_2}$ for dimensions $p \in \{2, 16\}$, sample sizes $n \in \{100, 250, 500, 1000\}$ and $\epsilon_n = n^{-1/2.5}$. The empirical proportions of rejection of $\mathcal{H}_{0,\mathcal{M}_1,\mathcal{M}_2}$ for the case $\sigma = 1.0$ are the empirical levels of our tests. The cases $\sigma > 1.0$ correspond to their empirical powers.

First of all, we observe in Figure 3 that all tests keep their empirical level fairly well. As previously, their power is relatively small for small a dimension ($d = 2$) and a small sample

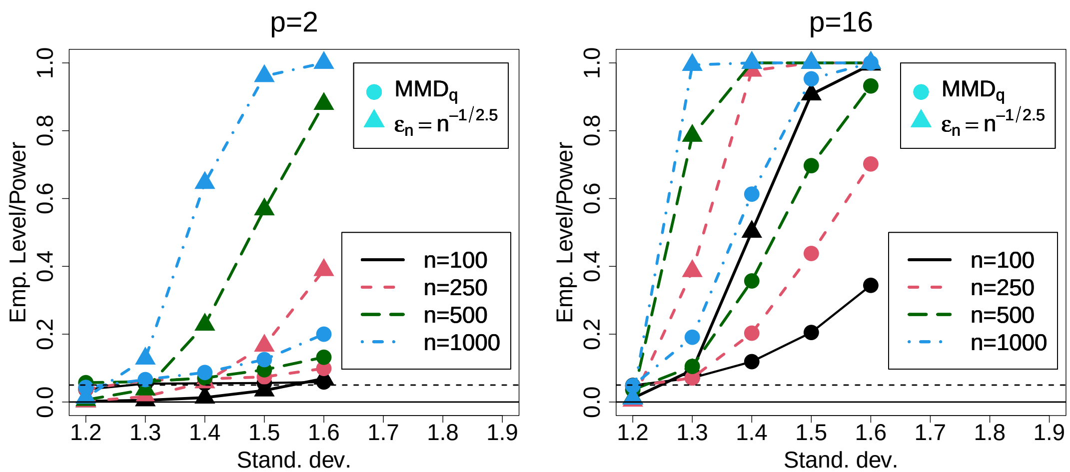

Figure 4: Non-degenerate case for comparison of two models: Empirical level and power of the tests based on $\mathcal{T}_n(\mathcal{M}_1, \mathcal{M}_2, P)$ and $\sqrt{n}\big(\widehat{\text{MMD}}^2_q(P_{\alpha_n}, P) - \widehat{\text{MMD}}^2_q(Q_{\beta_n}, P)\big)$ for dimensions $p = 2$ (left), $p = 16$ (right) as well as for sample sizes $n = 100, 250, 500, 1000$ (see the legend), $\epsilon_n = n^{-1/2.5}$ and varying standard deviation in Model $\mathcal{M}_1$. Both models do not coincide with the true model. The rejection probabilities are estimated using 1000 replications of the tests based on samples of size n. The black dashed line indicates the significance level 0.05.

size ($n = 100, 250$). Note that the empirical power of our MMD test is always higher than that provided by the test solely based on $\sqrt{n}\big(\widehat{\text{MMD}}^2_q(P_{\alpha_n}, P) - \widehat{\text{MMD}}^2_q(Q_{\beta_n}, P)\big)$. For the considered tuning parameter $\epsilon_n = n^{-1/2.5}$, the test based on $\mathcal{T}_n(\mathcal{M}_1, \mathcal{M}_2, P)$ always outperforms the test based on the competitor test statistic $\sqrt{n}\big(\widehat{\text{MMD}}^2_q(P_{\alpha_n}, P) - \widehat{\text{MMD}}^2_q(Q_{\beta_n}, P)\big)$ for the two considered dimensions and $n \in \{500, 1000\}$. As expected, the empirical power of all tests is increasing with increasing sample sizes. As we have already observed, it is also increasing for increasing dimensions.

For the fourth example, we modify the third example to avoid the degenerate case. Now, the models $\mathcal{M}_1$ and $\mathcal{M}_2$ are given by

$$Y(\alpha) = Y + \alpha \sim P_{\alpha},;\ Y \sim \mathcal{N}\left(0, diag(1.2^2, \ldots, 1.2^2, \sigma^2, \ldots, \sigma^2)\right), \text{ and}$$

$$Z(\beta) = Z + \beta \sim Q_{\beta};\ Z \sim \mathcal{N}\left(0, 1.2^2 I_p\right),$$

respectively. Thus, both models cannot coincide with the DGP, reflecting the non-degenerate case. However, for $\sigma = 1.2$, they coincide and are therefore equally far away from the DGP. We vary the standard deviation σ by setting $\sigma \in \{1.2,\ 1.3,\ 1.4,\ 1.5,\ 1.6\}$.

For the two tests based on the statistics $\mathcal{T}_n(\mathcal{M}_1, \mathcal{M}_2, P)$ and $\sqrt{n}\big(\widehat{\text{MMD}}^2_q(P_{\alpha_n}, P) - \widehat{\text{MMD}}^2_q(Q_{\beta_n}, P)\big)$, Figure 4 shows the empirical proportion of rejections of the null hypothesis $\mathcal{H}_{0,\mathcal{M}_1,\mathcal{M}_2}$ for dimensions $p \in \{2, 16\}$, sample sizes $n \in \{100, 250, 500, 1000\}$ and $\epsilon_n = n^{-1/2.5}$. The empirical proportions of rejections of $\mathcal{H}_{0,\mathcal{M}_1,\mathcal{M}_2}$ for the case $\sigma = 1.2$ (resp. $\sigma > 1.2$) are the empirical levels (resp. powers) of the tests. In Figure 4, we clearly observe that the test

Country	Specification test for the normal mixture	Specification test for the scaled t-distribution	Model comparison test
Austria	0.3872	0.8288	0.6019
Germany	0.7143	0.4689	0.6981
Ireland	0.8301	0.6783	0.9001
Italy	0.9638	0.5020	0.5140
Netherlands	0.6206	0.5335	0.9051
Singapore	0.6548	0.7121	0.9582
Sweden	0.7638	0.8206	0.9426

Table 1: P-values for the model specification and model comparison tests with $\epsilon_n = n^{-1/2.5}$ for the standardized residuals of Austria, Germany, Ireland, Italy, Netherlands, Singapore and Sweden (univariate models).

based on $\mathcal{T}_n(\mathcal{M}_1, \mathcal{M}_2, P)$ is again more powerful than a test based on $\sqrt{n}\big(\widehat{\text{MMD}}_q^2(P_{\alpha_n}, P) - \widehat{\text{MMD}}_q^2(Q_{\beta_n}, P)\big)$. Further, similar conclusions as in the third example can be drawn.

The `R` code for the implemented tests that are used in the simulation study is accessible at `https://github.com/Flo771994/MMD_tests_for_model_selection`. Fabian Baier recently translated the `R` code into Python, and the corresponding implementation is now available at `https://github.com/fabianbaiertum/BFM-test`.

6. Empirical Analysis

We illustrate the proposed MMD specification and model selection tests on historical stock returns, using the same data set as Brück et al. (2023). This data set consists of the daily log-returns of the MSCI indices for Austria, Germany, Ireland, Italy, the Netherlands, Singapore and Sweden from December 31, 1998 till March 12, 2018. The 5007 daily log-returns of seven countries constitute a multivariate time series, for which the considered MMD tests are not applicable. Therefore, we follow a standard econometrical approach by filtering every univariate time series of log-returns through ARMA(p,q)-GARCH(1,1) techniques (see McNeil et al., 2005, Chapter 4). Using BIC, we find that the model ARMA(0,0)-GARCH(1,1) is the most suitable one for the considered returns. Thus, we get so-called "standardized" residuals, that can reasonably be considered as i.i.d., with expectation zero and variance 1. We then focus on the law of such residuals.

Using the MMD specification and model selection tests, we first answer the question of whether a scaled t-distribution or a mixture of two normal distributions are appropriate for fitting the univariate standardized residuals, two standard choices in financial econometrics. Obviously, all distributions in the two considered models are restricted to have expectation 0 and variance 1. Furthermore, as proposed in the previous section, we use the nuisance parameter $\epsilon_n = n^{-1/2.5}$ for $n = 5007$. Table 1 reports the p-values for the proposed MMD-based tests. Our conclusions are consistent with those obtained by Brück et al. (2023) with the corrected Clarke test and the one-step Vuong test. Namely, none of the two considered univariate distributions for standardized residuals can be preferred with respect

Type of the test / Copula	Normal	Student t	Clayton	Gumbel
Model specification	0.8444	0.7177	$1.9 \cdot 10^{-29}$	$9.79 \cdot 10^{-25}$
Model comparison	0.8555		0.0006	

Table 2: P-values for the model specification and model comparison tests with $\epsilon_n = n^{-1/2.5}$ for seven-dimensional standardized residuals (multivariate models).

to another. This is not surprising because both parametric families are commonly used in applied financial econometrics.

In the next step, we would like to find an appropriate multivariate model for the seven dimensional vector of standardized residuals. Since this problem is not an easy task, we split it into two sub-tasks: first, the choice of marginal models, and, second, the choice of a copula model, as in McNeil et al. (2005, Chapter 5) and many others. According to Table 1, the scaled t-distribution is appropriate for modeling the marginal standardized residuals. Thus, we select it as the univariate marginal model for each of the seven standardized residuals. In order to estimate our parametric copula models under consideration, we transform all univariate standardized residuals by their marginal empirical distributions to get pseudo-observations. The copula parameters are then estimated by semiparametric pseudo-maximum likelihood (Genest et al., 1995). This standard technique yields asymptotically normal estimators.

Now, we can compare the normal copula and the t-copula (McNeil et al., 2005, Chapter 5), and perform also specification tests for them. The full multivariate distribution of the seven dimensional standardized residuals is finally specified by the marginal scaled t-distributions coupled with either the normal copula or the Student t copula. The first three p-values in the left part of Table 2 indicate that the specification test as well as the model selection test cannot reject the corresponding null hypotheses at the 5% significance level. It appears that the degree of freedom of the Student-t copula cannot fully capture all bivariate tail dependencies and, therefore, cannot be favored in this context, despite its common preference in applied financial econometrics.

Alternatively, we also consider the Clayton copula and the Gumbel copula. They are governed only by a single parameter, which is probably not enough to reasonably model a random vector of dimension seven. When they are combined with marginal scaled t-distributions, the p-values of the specifications tests for the Clayton copula and the Gumbel copula in the right part of Table 2 are not surprisingly very small. Therefore, the null hypotheses of model specification can be rejected for both models, at any standard nominal levels. Moreover, the null hypothesis that these two models are equivalently well suited is rejected at the 5% level. The test statistics is equal to 3.4271, indicating that the multivariate model with the Gumbel copula is preferred to the one with the Clayton copula.

7. Conclusion and Outlook

We have provided novel MMD-based model specification and model selection tests when the model parameters are estimated in a first-stage. In comparison with an approach solely based on $\widehat{\mathrm{MMD}}^2(P_{\alpha_n}, P)$, these tests circumvent the major difficulty of computing the critical values of a complicated asymptotic distribution, but instead simply require to resort to the critical values of the standard normal distribution. Moreover, since our distribution free testing procedures are also valid when no parameter estimation is conducted, they yield valuable alternatives to the two sample test proposed by Gretton et al. (2006). The testing procedures are summarized in Algorithm 1 and Algorithm 2, respectively.

Both proposed test statistics depend on a tuning parameter, whose "optimal" choice is still an open problem that may be investigated in the future, for example by a local power analysis in the same spirit as Schennach and Wilhelm (2017). Moreover, it remains future work to derive certain properties of our testing procedures, such as the local power and uniformity of convergence over sets of DGPs. Moreover, Chérief-Abdellatif and Alquier (2022) recently investigated parameter estimation based on the MMD for dependent input data. Since our test statistics are basically composed of U-statistics, future research might be concerned with relaxing the assumption of i.i.d. observations in the application of our MMD-based model specification and model selection tests by resorting to the vast literature on U-statistics of dependent data.

Acknowledgments and Disclosure of Funding

Parts of this project were conducted during a research stay of Florian Brück at CREST, which was funded by the FMJH scholarship "Junior Scientific Visibility". During his time at the University of Geneva, Florian Brück was funded by the Grant 186858 of the Swiss National Science Foundation and he additionally acknowledges that parts of this project were conducted during his time as a PhD student at Technical University Munich. Jean-David Fermanian was supported by the labex Ecodec (reference project ANR-11-LABEX-0047).

Appendix A. Proofs

Hereafter, for any symmetric map ψ and any i.i.d. sample $(Z_1, \dots, Z_n)$, the associated U-statistic of degree r is denoted as $\mathcal{U}_n \psi := \big(r!\binom{n}{r}\big)^{-1} \sum_{(i_1, i_2, \dots, i_r) \in I_{r,n}} \psi(Z_{i_1}, \dots, Z_{i_r})$, where

$$I_{r,n} := \{(i_1, \dots, i_r) : i_j \in \mathbb{N}, 1 \leq i_j \leq n; i_j \neq i_k \text{ if } j \neq k\}.$$

In our case, Z_i will be (X_i, U_i) or the concatenation of several similar random vectors. More specifically, consider a class $\mathcal{L}$ of symmetric real-valued functions on $\otimes_{i=1}^m (\mathcal{S} \otimes \mathcal{U})$. With a slight abuse of notation, denote as $(\mathcal{U}_n^{(m)} \ell)_{\ell \in \mathcal{L}}$ the empirical U-process which acts on the sample

$$\big([\boldsymbol{X}, \boldsymbol{U}]_{1:m}, [\boldsymbol{X}, \boldsymbol{U}]_{(m+1):2m} \dots, [\boldsymbol{X}, \boldsymbol{U}]_{(n-m+1):n}\big),$$

and whose degree is determined by ℓ. The latter sample is drawn from $\otimes_{i=1}^m (P \otimes P_U)$ and is of size n/m (which is implicitly assumed to be an integer). Note that $\mathcal{U}_n^{(1)} \ell$ is just $\mathcal{U}_n \ell$ and we drop the upper index in this case.

Let us illustrate our notation with the two classes of symmetric functions

$$\mathcal{F} := \Big\{ [\boldsymbol{x}, \boldsymbol{u}]_{1:2} \mapsto h\big([\boldsymbol{x}, \boldsymbol{u}]_{1:2}; \alpha\big) \Big| \, \alpha \in B_\delta(\alpha_\star) \Big\}, \text{ and} \tag{18}$$

$$\mathcal{F}_q := \Big\{ [\boldsymbol{x}, \boldsymbol{u}]_{1:4} \mapsto q\big([\boldsymbol{x}, \boldsymbol{u}]_{1:4}; \alpha\big) \Big|\ \alpha \in B_\delta(\alpha_\star) \Big\}, \tag{19}$$

for some $\delta > 0$, where h and q are defined in (7) and (8), respectively. The empirical U-process of degree one indexed by the set of functions $\mathcal{F}$ is defined as the stochastic process

$$\mathcal{U}_n : \ \mathcal{F} \mapsto \mathbb{R};\ f \mapsto \mathcal{U}_n f := \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} f\big([\boldsymbol{X}, \boldsymbol{U}]_{i,j}\big).$$

Moreover, the empirical U-process of degree 2 indexed by $\mathcal{F}_q$ is the stochastic process

$$\mathcal{U}_n^{(2)} : \ \mathcal{F}_q \mapsto \mathbb{R};\ f \mapsto \mathcal{U}_n^{(2)} f := \frac{1}{n/2(n/2-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n/2} f\big([\boldsymbol{X}, \boldsymbol{U}]_{2i-1,2i,2j-1,2j}\big).$$

This notation allows to rewrite the estimators of $\mathrm{MMD}^2(P_{\alpha_\star}, P)$ in terms of empirical U-processes as

$$\widehat{\mathrm{MMD}}^2(P_{\alpha_n}, P) = \mathcal{U}_n h\left(\cdot; \alpha_n\right) \ \text{ and } \ \widehat{\mathrm{MMD}}_q^2(P_{\alpha_n}, P) = \mathcal{U}_n^{(2)} q\left(\cdot; \alpha_n\right).$$

Moreover, let the operator $\mathbb{P}\ell$ denote $\mathbb{E}\left[\ell(Z)\right]$, where the random vector Z is in accordance with the arguments of ℓ. Thus, $\big((\mathcal{U}_n^{(m)} - \mathbb{P})\ell\big)_{\ell \in \mathcal{L}}$ denotes the centered empirical U-processes whose asymptotic behavior was investigated in Arcones and Giné (1993, 1994), among others.

A.1 Theoretically Convenient Variance Estimation

Let us begin this section by proposing alternative estimators of $\sigma^2_{\alpha_\star}$, $\sigma^2_{q,\alpha_\star}$,$\sigma^2_{\alpha_\star,\beta_\star}$ and $\sigma^2_{q,\alpha_\star,\beta_\star}$, the asymptotic variances of $\widehat{\mathrm{MMD}}^2(P_{\alpha_\star}, P)$, $\widehat{\mathrm{MMD}}_q^2(P_{\alpha_\star}, P)$, $\widehat{\mathrm{MMD}}^2(P_{\alpha_\star}, P) - \widehat{\mathrm{MMD}}^2(Q_{\beta_\star}, P)$ and $\widehat{\mathrm{MMD}}_q^2(P_{\alpha_\star}, P) - \widehat{\mathrm{MMD}}_q^2(Q_{\beta_\star}, P)$ respectively. In Sections 3.2 and 4.1, we have defined empirical estimators $\tilde{\sigma}^2_{\alpha_n}$ and $\tilde{\sigma}^2_{q,\alpha_n}$ of $\sigma^2_{\alpha_\star}$ and $\sigma^2_{q,\alpha_\star}$. Unfortunately, these estimators are not U-statistics. Thus, an analysis of their asymptotic properties is not very convenient. In the following, we will introduce alternative U-statistics estimators and show that they are asymptotically equivalent - up to a term which tends to 0 sufficiently fast - to the estimators defined in Sections 3.2 and 4.1. This will allow us to develop asymptotic theory for the simpler to analyze U-statistic estimators and to derive the corresponding asymptotic properties of the estimators introduced in Sections 3.2 and 4.1 straightforwardly.

First, for any fixed α, we symmetrize $\tilde{\sigma}^2_\alpha$ and discard the terms for which the indices j and k are equal, since their contribution is negligible. Doing so, we obtain a proper U-statistic, which is much more convenient to work with. This leads to the following estimator of σ^2_α:

$$\begin{aligned}\hat{\sigma}^2_\alpha := \frac{4}{3n(n-1)(n-2)} & \sum_{\substack{i,j,k=1 \\ i \neq k, i \neq j, k \neq j}}^{n} \big\{ h\big([\boldsymbol{X}, \boldsymbol{U}]_{i,j}; \alpha\big) h\big([\boldsymbol{X}, \boldsymbol{U}]_{i,k}; \alpha\big) \\ &+ h\big([\boldsymbol{X}, \boldsymbol{U}]_{j,i}; \alpha\big) h\big([\boldsymbol{X}, \boldsymbol{U}]_{j,k}; \alpha\big) + h\big([\boldsymbol{X}, \boldsymbol{U}]_{k,j}; \alpha\big) h\big([\boldsymbol{X}, \boldsymbol{U}]_{k,i}; \alpha\big) \big\} \\ &- 4\big\{\widehat{\mathrm{MMD}}^2(P_\alpha, P)\big\}^2 = \mathcal{U}_n g(\cdot; \alpha) - 4\big\{\widehat{\mathrm{MMD}}^2(P_\alpha, P)\big\}^2.\end{aligned}$$

Thus, $\hat{\sigma}_\alpha^2$ can be decomposed into the weighted sum of the squared $\widehat{\mathrm{MMD}}^2(P_\alpha, P)$ and a U-statistic of degree three with the symmetric U-kernel $g(\cdot;\alpha)$, as defined in (13). Obviously, an estimator of $\sigma_{\alpha_\star}^2$ can now be defined as

$$\hat{\sigma}_{\alpha_n}^2 := \mathcal{U}_n g(\cdot;\alpha_n) - 4\left\{\widehat{\mathrm{MMD}}^2(P_{\alpha_n}, P)\right\}^2. \tag{20}$$

Similarly, we can define a U-statistic estimator of $\sigma_{q,\alpha}^2$ given by

$$\begin{aligned}\hat{\sigma}_{q,\alpha}^2 \quad := \quad & \frac{2}{n(n/2-1)(n/2-2)} \sum_{\substack{i,j,k=1 \\ i\neq k, i\neq j, k\neq j}}^{n/2} \xi\big([\boldsymbol{X},\boldsymbol{U}]_{2i-1,2i,2j-1,2j,2k-1,2k};\alpha\big) \\ & -8\left\{\widehat{\mathrm{MMD}}^2(P_\alpha,P)\right\}^2 = \mathcal{U}_n^{(2)}\xi(\cdot;\alpha) - 8\left\{\widehat{\mathrm{MMD}}^2(P_\alpha,P)\right\}^2.\end{aligned}$$

where we define a symmetric (in $[\boldsymbol{x},\boldsymbol{u}]_{1:2}$, $[\boldsymbol{x},\boldsymbol{u}]_{3:4}$ and $[\boldsymbol{x},\boldsymbol{u}]_{5:6}$) U-kernel as

$$\begin{aligned}\xi(\cdot;\alpha) : [\boldsymbol{x},\boldsymbol{u}]_{1:6} \mapsto & \frac{8}{3}\Big\{q\big([\boldsymbol{x},\boldsymbol{u}]_{1,2,3,4};\alpha\big)q\big([\boldsymbol{x},\boldsymbol{u}]_{1,2,5,6};\alpha\big) \\ & + q\big([\boldsymbol{x},\boldsymbol{u}]_{3,4,1,2};\alpha\big)q\big([\boldsymbol{x},\boldsymbol{u}]_{3,4,5,6};\alpha\big) + q\big([\boldsymbol{x},\boldsymbol{u}]_{5,6,3,4};\alpha\big)q\big([\boldsymbol{x},\boldsymbol{u}]_{5,6,1,2};\alpha\big)\Big\}.\end{aligned} \tag{21}$$

Note that the operator $\mathcal{U}_n^{(2)}\xi(\cdot,\alpha)$ is a usual U-statistic of degree three on the sample $[\boldsymbol{X},\boldsymbol{U}]_{(2i-1):(2i)}$ for $i = 1,\dots,n/2$. Now, an estimator of $\sigma_{q,\alpha_\star}^2$ can be defined as $\hat{\sigma}_{q,\alpha_n}^2$.

In the model selection framework we can analogously define U-statistic estimators for the asymptotic variances $\sigma_{\alpha,\beta}^2$ and $\sigma_{q,\alpha,\beta}^2$. For a given tuple (α,β), we define

$$\begin{aligned}\hat{\sigma}_{\alpha,\beta}^2 := & \frac{1}{n(n-1)(n-2)} \sum_{\substack{i,j,k=1 \\ i\neq j\neq k}}^{n} g\big([\boldsymbol{X},\boldsymbol{U},\boldsymbol{V}]_{i,j,k};\alpha,\beta\big) - 4\left\{\widehat{\mathrm{MMD}}^2(P_\alpha,P) - \widehat{\mathrm{MMD}}^2(Q_\beta,P)\right\}^2 \\ = & \ \mathcal{U}_n g(\cdot;\alpha,\beta) - 4\left\{\widehat{\mathrm{MMD}}^2(P_\alpha,P) - \widehat{\mathrm{MMD}}^2(Q_\beta,P)\right\}^2,\end{aligned} \tag{22}$$

as an estimator of $\sigma_{\alpha,\beta}^2$ where we define the symmetric (w.r.t. triplets $(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v})$ or concatenated pairs of triplets) U-statistic kernel

$$\begin{aligned}g\big([\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{1:3};\alpha,\beta\big) := & \frac{4}{3}\Big\{h\big([\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{1,2};\alpha,\beta\big)h\big([\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{1,3};\alpha,\beta\big) \\ & + h\big([\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{2,1};\alpha,\beta\big)h\big([\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{2,3};\alpha,\beta\big) + h\big([\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{3,2};\alpha,\beta\big)h\big([\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{3,1};\alpha,\beta\big)\Big\}.\end{aligned} \tag{23}$$

Moreover, we can define an estimator of $\sigma_{q,\alpha,\beta}^2$ based on $\xi\big([\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{1:6};\alpha,\beta\big)$ by

$$\begin{aligned}\hat{\sigma}_{q,\alpha,\beta}^2 \quad := \quad & \frac{2}{n(n/2-1)(n/2-2)} \sum_{\substack{i,j,k=1 \\ i\neq j\neq k}}^{n/2} \xi\big([\boldsymbol{X},\boldsymbol{U},\boldsymbol{V}]_{2i-1,2i,2j-1,2j,2k-1,2k};\alpha,\beta\big) \\ & -8\left\{\widehat{\mathrm{MMD}}^2(P_\alpha,P) - \widehat{\mathrm{MMD}}^2(Q_\beta,P)\right\}^2 \\ = \quad & \mathcal{U}_n^{(2)}\xi(\cdot;\alpha,\beta) - 8\left\{\widehat{\mathrm{MMD}}^2(P_\alpha,P) - \widehat{\mathrm{MMD}}^2(Q_\beta,P)\right\}^2,\end{aligned}$$

where we define

$$\xi\big([\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{1:6};\alpha,\beta\big) \coloneqq \frac{8}{3}\big\{q\big([\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{1,2,3,4};\alpha,\beta\big)q\big([\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{1,2,5,6};\alpha,\beta\big)+ \tag{24}$$
$$q\big([\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{3,4,1,2};\alpha,\beta\big)q\big([\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{3,4,5,6};\alpha,\beta\big)+q\big([\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{5,6,3,4};\alpha,\beta\big)q\big([\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{5,6,1,2};\alpha,\beta\big)\big\}.$$

Therefore, U-statistic estimators for $\sigma^2_{\alpha_\star,\beta_\star}$ and $\sigma^2_{q,\alpha_\star,\beta_\star}$ are given by

$$\hat{\sigma}^2_{\alpha_n,\beta_n} \coloneqq \mathcal{U}_n g(\cdot;\alpha_n,\beta_n) - 4\big\{\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P) - \widehat{\mathrm{MMD}}^2(Q_{\beta_n},P)\big\}^2, \text{ and}$$
$$\hat{\sigma}^2_{q,\alpha_n,\beta_n} \coloneqq \mathcal{U}_n^{(2)}\xi(\cdot;\alpha_n,\beta_n) - 8\big\{\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P) - \widehat{\mathrm{MMD}}^2(Q_{\beta_n},P)\big\}^2,$$

which are sums of parameter dependent U-statistics. The following Lemma shows that $\tilde{\sigma}^2_{\alpha_n}$, $\tilde{\sigma}^2_{q,\alpha_n}$, $\tilde{\sigma}^2_{\alpha_n,\beta_n}$ and $\tilde{\sigma}^2_{q,\alpha_n,\beta_n}$ are asymptotically equivalent to $\hat{\sigma}^2_{\alpha_n}$, $\hat{\sigma}^2_{q,\alpha_n}$, $\hat{\sigma}^2_{\alpha_n,\beta_n}$ and $\hat{\sigma}^2_{q,\alpha_n,\beta_n}$.

Lemma 1 *Under Assumptions 1-2,* $\tilde{\sigma}^2_{\alpha_n} = \hat{\sigma}^2_{\alpha_n} + O_{\mathbb{P}}(n^{-1})$*,* $\tilde{\sigma}^2_{q,\alpha_n} = \hat{\sigma}^2_{q,\alpha_n} + O_{\mathbb{P}}(n^{-1})$*,* $\tilde{\sigma}^2_{\alpha_n,\beta_n} = \hat{\sigma}^2_{\alpha_n,\beta_n} + O_{\mathbb{P}}(n^{-1})$ *and* $\tilde{\sigma}^2_{q,\alpha_n,\beta_n} = \hat{\sigma}^2_{q,\alpha_n,\beta_n} + O_{\mathbb{P}}(n^{-1})$*. Additionally,* $\sigma_{\alpha_\star} > 0$ *whenever* $P_{\alpha_\star} \neq P$ *and, under Assumption 5,* $\sigma_{\alpha_\star,\beta_\star} > 0$ *whenever* $P_{\alpha_\star} \neq P$ *or* $Q_{\beta_\star} \neq P$*.*

Proof First, a simple calculation shows that

$$\tilde{\sigma}^2_{\alpha_n} = \hat{\sigma}^2_{\alpha_n} + \left(\frac{n-2}{n-1} - 1\right)\mathcal{U}_n g(\cdot;\alpha_n) + \frac{4}{n-1}\mathcal{U}_n h(\cdot;\alpha_n)^2,$$

where $\mathcal{U}_n h(\cdot;\alpha)^2$ is a U-statistic of degree two with U-kernel $h(\cdot;\alpha)^2$ acting on the sample $(X_i,U_i)_{i=1,\ldots,n}$. The boundedness of k implies that there exists a constant $C>0$ such that $|g(\cdot,\alpha)| \le C$ and $|h^2(\cdot,\alpha)| \le C$, which gives $|\mathcal{U}_n g(\cdot;\alpha_n)| \le C$ and $|\mathcal{U}_n h(\cdot;\alpha_n)^2| \le C$. Thus,

$$\left|\left(\frac{n-2}{n-1} - 1\right)\mathcal{U}_n g(\cdot;\alpha_n) + \frac{4}{n-1}\mathcal{U}_n h(\cdot;\alpha_n)^2\right| \le \frac{5}{n-1}C = O_{\mathbb{P}}(n^{-1}),$$

proving that $\tilde{\sigma}^2_{\alpha_n} = \hat{\sigma}^2_{\alpha_n} + O_{\mathbb{P}}(n^{-1})$. Similar arguments show that $\tilde{\sigma}^2_{q,\alpha_n} = \hat{\sigma}^2_{q,\alpha_n} + O_{\mathbb{P}}(n^{-1})$, $\tilde{\sigma}^2_{\alpha_n,\beta_n} = \hat{\sigma}^2_{\alpha_n,\beta_n} + O_{\mathbb{P}}(n^{-1})$ and $\tilde{\sigma}^2_{q,\alpha_n,\beta_n} = \hat{\sigma}^2_{q,\alpha_n,\beta_n} + O_{\mathbb{P}}(n^{-1})$.

To prove that $\sigma_{\alpha_\star} > 0$ whenever $P_{\alpha_\star} \neq P$, assume $\sigma_{\alpha_\star} = 0$ and $\mathrm{MMD}(P_{\alpha_\star},P) > 0$, seeking a contradiction. Note that $0 = \sigma^2_{\alpha_\star} = \mathrm{Var}\big(\tilde{h}(X_1,U_1;\alpha_\star)\big)$ implies $\tilde{h}(X_1,U_1;\alpha_\star) = C$ a.s., and, w.l.o.g., assume $C \ge 0$. Since $\mathrm{MMD}(P_{\alpha_\star},P) = \mathbb{E}_{X,U}[\tilde{h}(X,U;\alpha_\star)]$, $\tilde{h}(X_1,U_1;\alpha_\star) = 0$ a.s. implies $\mathrm{MMD}(P_{\alpha_\star},P) = 0$, which would be a contradiction. Thus $C > 0$. Moreover, observe that we can rewrite $\tilde{h}(X_1,U_1,\alpha_\star) = T_1(X_1) + T_2(U_1)$, for some maps T_1 and T_2. Since X_1 and U_1 are independent, there exist two constants C_1 and C_2 s.t. $T_1(X_1) = C_1$ and $T_2(U_1) = C_2$ a.s., and at least one of C_1 and C_2 is non-zero. W.l.o.g. assume $C_1 > 0$ which gives

$$C_1 = \mathbb{E}_{X_2}\big[k(X_2,X_1)\big] - \mathbb{E}_{U_2}\big[k(X_1,F(U_2,\alpha_\star))\big] =: s_1(X_1) - s_2(X_1).$$

This implies $s_1(x) - s_2(x) = C_1$ for almost every $x \in \mathrm{supp}(P)$. By definition of T_2, we have $C_2 = T_2(U_1) = -s_1\big(F(U_1,\alpha_\star)\big) + s_2\big(F(U_1,\alpha^*)\big)$ a.s. Therefore, since $\mathrm{supp}(P) \cap \mathrm{supp}(P_{\alpha_\star}) \neq \varnothing$, this yields $C_2 = -C_1$ and, then, we have $\tilde{h}(X_1,U_1,\alpha_\star) = 0$ a.s., which is contradiction to $\mathrm{MMD}(P_{\alpha_\star},P) > 0$. Thus $\sigma_{\alpha_\star} > 0$ when $P_{\alpha_\star} \neq P$. The proof of $\sigma_{\alpha_\star,\beta_\star} > 0$ whenever $P_{\alpha_\star} \neq P$ or $Q_{\beta_\star} \neq P$ is similar and is thus omitted. ∎

A.2 Proof of Theorem 1 (Model Specification)

We first state a general lemma which is of interest per se.

Lemma 2 *Suppose that Assumptions 1-4 hold.*

(i) *If* $\sqrt{n}\big[\widehat{\mathrm{MMD}}^2(P_{\alpha_\star},P)-\mathrm{MMD}^2(P_{\alpha_\star},P),\ \alpha_n-\alpha_\star\big]$ *weakly tends to a real-valued random vector* $[Z_{\alpha_\star},V_{\alpha_\star}]$ *then*

$$\sqrt{n}\big\{\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P)-\mathrm{MMD}^2(P_{\alpha_\star},P)\big\}\xrightarrow{law} Z_{\alpha_\star}+\nabla_{\alpha^\intercal}\,\mathrm{MMD}^2(P_\alpha,P)|_{\alpha=\alpha_\star}V_{\alpha_\star},$$

where $Z_{\alpha_\star}\sim\mathcal{N}\big(0,\sigma^2_{\alpha_\star}\big)$. *Moreover,* $\tilde{\sigma}^2_{\alpha_n}\to\sigma^2_{\alpha_\star}$ *in probability.*

(ii) *If* $\sqrt{n}(\alpha_n-\alpha_\star)=O_{\mathbb{P}}(1)$ *and* $P_{\alpha_\star}=P$, *i.e., if* $\mathcal{H}_{0,\mathcal{M}}$ *is satisfied, then* $\sqrt{n}\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P)$ $=O_{\mathbb{P}}(n^{-1/2})$ *and* $\tilde{\sigma}^2_{\alpha_n}=O_{\mathbb{P}}(n^{-1})$.

(iii) *If* $\sqrt{n}\big[\widehat{\mathrm{MMD}}^2_q(P_{\alpha_\star},P)-\mathrm{MMD}^2(P_{\alpha_\star},P),\ \alpha_n-\alpha_\star\big]$ *weakly tends to a real-valued random vector* $[Z_{q,\alpha_\star},V_{\alpha_\star}]$ *then*

$$\sqrt{n}\big\{\widehat{\mathrm{MMD}}^2_q(P_{\alpha_n},P)-\mathrm{MMD}^2(P_{\alpha_\star},P)\big\}\xrightarrow{law} Z_{q,\alpha_\star}+\nabla_{\alpha^\intercal}\,\mathrm{MMD}^2(P_\alpha,P)|_{\alpha=\alpha_\star}V_{\alpha_\star},$$

where $Z_{q,\alpha_\star}\sim\mathcal{N}\big(0,\sigma^2_{q,\alpha_\star}\big)$. *Moreover,* $\tilde{\sigma}^2_{q,\alpha_n}\to\sigma^2_{q,\alpha_\star}>0$ *in probability.*

In the latter lemma, the weak convergence of the two centered MMD estimators is guaranteed by standard U-statistics results (through Hájek projections). Here, the main point is their joint convergence with $\sqrt{n}(\alpha_n-\alpha_\star)$, which is usually guaranteed when $\alpha_n-\alpha_\star$ can be approximated by an i.i.d. expansion, a typical situation with M-estimators (van der Vaart, 2000, Section 5.3). For instance, when a log density is sufficiently regular (twice differentiable, in particular) w.r.t. its parameter, this is most often the case for maximum likelihood estimators. Such i.i.d. expansions directly appear when α_n are sample averages, differentiable functionals of them, or sample quantiles. Note that the variance of $Z_{\alpha_\star}$ is zero in the degenerate case $P_{\alpha_\star}=P$, which will be the case under $\mathcal{H}_{0,\mathcal{M}}$.

Proof [of Lemma 2]

(i): A first order Taylor expansion yields

$$\begin{aligned}\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P) &= \frac{1}{n(n-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{n} h\big([\boldsymbol{X},\boldsymbol{U}]_{i,j};\alpha_\star\big)+O_{\mathbb{P}}(\|\alpha_n-\alpha_\star\|^2)\\ &\quad+\frac{1}{n(n-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{n}\nabla_{\alpha^\intercal}h\big([\boldsymbol{X},\boldsymbol{U}]_{i,j};\alpha_\star\big)\cdot(\alpha_n-\alpha_\star).\end{aligned}\tag{25}$$

Note that the remainder term $O_{\mathbb{P}}(\|\alpha_n-\alpha_\star\|^2)$ is due to Assumption 4 and is $O_{\mathbb{P}}(n^{-1})$. The weak convergence of $\sqrt{n}\big\{\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P)-\mathrm{MMD}^2(P_{\alpha_\star},P)\big\}$ is a direct consequence of weak convergence assumption, in addition to Leibniz's Theorem applied to $\nabla_\alpha h$ at $\alpha = \alpha_\star$ (second part of Assumption 4). Recalling (20), the convergence of the estimated variance $\hat{\sigma}^2_{\alpha_n}$ is easily obtained by a first order Taylor expansion of the maps $\alpha \mapsto g\big([\boldsymbol{x},\boldsymbol{u}]_{i,j,k};\alpha\big)$ defined in (13) around $\alpha_\star$, for every triplet of indices (i,j,k), $i \neq j \neq k$. Therefore, $\hat{\sigma}^2_{\alpha_n} = \mathcal{U}_n g(\cdot;\alpha_\star) - 4\big\{\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P)\big\}^2 + o_{\mathbb{P}}(1)$ that tends to $\sigma^2_{\alpha_\star}$ by the consistency of usual U-statistics. Note that we have again invoked Assumption 4 to manage the remainder term. This immediately implies the consistency of $\tilde{\sigma}^2_{\alpha_n}$ by Lemma 1.

(ii): With obvious notations, rewrite (25) as

$$\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P) = \mathcal{U}_n h(\cdot;\alpha_\star) + \mathcal{U}_n \nabla_{\alpha^\top} h(\cdot;\alpha_\star).(\alpha_n - \alpha_\star) + O_{\mathbb{P}}(n^{-1}).$$

Since $\mathcal{U}_n h(\cdot;\alpha_\star)$ is a degenerate U-statistic, it is $O_{\mathbb{P}}(n^{-1})$. Moreover, the mean of the U-statistic $\mathcal{U}_n \nabla_{\alpha^\top} h(\cdot;\alpha_\star)$ is zero because

$$\mathbb{E}\big[\nabla_\alpha h\big([\boldsymbol{X},\boldsymbol{U}]_{i,j};\alpha_\star\big)\big] = \nabla_\alpha \mathbb{E}\big[h\big([\boldsymbol{X},\boldsymbol{U}]_{i,j};\alpha_\star\big)\big] = \nabla_\alpha \mathrm{MMD}^2(P_\alpha,P)|_{\alpha=\alpha_\star} = 0.$$

Thus, the U-statistic $\mathcal{U}_n \nabla_{\alpha^\top} h(\cdot;\alpha_\star)$ is $O_{\mathbb{P}}(n^{-1/2})$. This implies $\sqrt{n}\,\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P) = O_{\mathbb{P}}(n^{-1/2})$, the announced result.

Concerning the asymptotic variance, a Taylor expansion of $\alpha \mapsto g(\cdot;\alpha)$ around $\alpha_\star$ and Assumption 4 yield

$$\begin{aligned}\hat{\sigma}^2_{\alpha_n} &= \mathcal{U}_n g(\cdot;\alpha_n) - 4\big\{\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P\big\}^2 \\ &= \mathcal{U}_n g(\cdot;\alpha_\star) + \mathcal{U}_n \nabla_{\alpha^\top} g(\cdot;\alpha_\star)(\alpha_n - \alpha_\star) + O_{\mathbb{P}}\big(\|\alpha_n - \alpha_\star\|^2\big) + O_{\mathbb{P}}(n^{-2}).\end{aligned}$$

Since it can be checked that $\mathcal{U}_n g(\cdot;\alpha_\star)$ is a degenerate U-statistic, this term is $O_{\mathbb{P}}(n^{-1})$. Moreover, $\mathcal{U}_n \nabla_\alpha g(\cdot;\alpha_\star)$ tends in probability to $\nabla_\alpha \mathbb{E}\big[g\big([\boldsymbol{X},\boldsymbol{U}]_{1:3};\alpha_\star\big)\big]$ (Assumption 4), that is zero because $\alpha \mapsto \mathbb{E}\big[g\big([\boldsymbol{X},\boldsymbol{U}]_{1:3};\alpha\big)\big]$ is minimized at $\alpha = \alpha_\star$. This implies that the U-statistic $\mathcal{U}_n \nabla_\alpha g(\cdot;\alpha_\star)$ is $O_{\mathbb{P}}(n^{-1/2})$. Globally, we get $\hat{\sigma}^2_{\alpha_n} = O_{\mathbb{P}}(n^{-1})$, which immediately implies $\tilde{\sigma}^2_{\alpha_n} = O_{\mathbb{P}}(n^{-1})$ by Lemma 1.

(iii): The positivity of $\sigma_{q,\alpha_\star}$ had already been noted at the end of Section 3.2. The rest can be proved exactly as for *(i)*.

■

The proof of Theorem 1 follows from Lemma 2, with a few adjustments: in the statement of Lemma 2, the term $\nabla_{\alpha^\top} \mathrm{MMD}^2(P_\alpha,P)|_{\alpha=\alpha_\star} V_{\alpha_\star}$ accounts for the influence of parameter estimation on the asymptotic distribution of $\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P)$ and $\widehat{\mathrm{MMD}}^2_q(P_{\alpha_n},P)$. This will no longer be the case under $\mathcal{H}_{0,\mathcal{M}}$, because then $\nabla_{\alpha^\top} \mathrm{MMD}^2(P_\alpha,P)|_{\alpha=\alpha_\star} = 0$. This explains why only $\sqrt{n}(\alpha_n - \alpha_\star) = O_{\mathbb{P}}(1)$ is required to prove Theorem 1. Moreover, since $\sqrt{n}\big\{\widehat{\mathrm{MMD}}^2(P_{\alpha_\star},P) - \mathrm{MMD}^2(P_{\alpha_\star},P)\big\}$ and $\sqrt{n}\big\{\widehat{\mathrm{MMD}}^2_q(P_{\alpha_\star},P) - \mathrm{MMD}^2(P_{\alpha_\star},P)\big\}$ are usual U-statistics, they are jointly weakly convergent.

1. Note that

$$\sqrt{n}\frac{\widehat{\mathrm{MMD}}^2_{\epsilon_n}(P_{\alpha_n},P)}{\tilde{\sigma}_{\alpha_n}+\epsilon_n\tilde{\sigma}_{q,\alpha_n}} = \frac{\sqrt{n}\,\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P)}{\tilde{\sigma}_{\alpha_n}+\epsilon_n\tilde{\sigma}_{q,\alpha_n}} + \frac{\sqrt{n}\epsilon_n\,\sqrt{n}\,\widehat{\mathrm{MMD}}^2_q(P_{\alpha_n},P)}{\sqrt{n}\tilde{\sigma}_{\alpha_n}+\sqrt{n}\epsilon_n\tilde{\sigma}_{q,\alpha_n}}.$$

Invoking Lemma 2 (ii), we obtain

$$\frac{\sqrt{n}\,\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P)}{\tilde{\sigma}_{\alpha_n}+\epsilon_n\tilde{\sigma}_{q,\alpha_n}} = \frac{O_{\mathbb{P}}(n^{-1/2})}{O_{\mathbb{P}}(n^{-1/2})+\epsilon_n\tilde{\sigma}_{q,\alpha_n}} = \frac{O_{\mathbb{P}}(1)}{O_{\mathbb{P}}(1)+\sqrt{n}\epsilon_n\tilde{\sigma}_{q,\alpha_n}}.$$

Since $\epsilon_n\sqrt{n}\to\infty$ in probability by assumption and $\tilde{\sigma}_{q,\alpha_n}\to\sigma_{q,\alpha_\star}>0$ in probability, this yields

$$\frac{\sqrt{n}\,\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P)}{\tilde{\sigma}_{\alpha_n}+\epsilon_n\tilde{\sigma}_{q,\alpha_n}} = o_{\mathbb{P}}(1).$$

Again, by Lemma 2 and the consistency of $\tilde{\sigma}_{q,\alpha_n}$, we have $\sqrt{n}\epsilon_n\tilde{\sigma}_{q,\alpha_n}\to\infty$ in probability, when $\sqrt{n}\tilde{\sigma}_{\alpha_n}=O_{\mathbb{P}}(1)$. This provides

$$\frac{\sqrt{n}\epsilon_n\sqrt{n}\,\widehat{\mathrm{MMD}}^2_q(P_{\alpha_n},P)}{\sqrt{n}\tilde{\sigma}_{\alpha_n}+\sqrt{n}\epsilon_n\tilde{\sigma}_{q,\alpha_n}} = \frac{\sqrt{n}\,\widehat{\mathrm{MMD}}^2_q(P_{\alpha_n},P)}{\tilde{\sigma}_{q,\alpha_n}} + o_{\mathbb{P}}(1) \xrightarrow{\text{law}} \frac{\mathcal{N}(0,\sigma^2_{q,\alpha_\star})}{\sigma_{q,\alpha_\star}} \sim \mathcal{N}(0,1),$$

which proves the claim.

2. When $P_{\alpha_\star}\neq P$, $\sqrt{n}\big\{\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P)-\mathrm{MMD}^2(P_{\alpha_\star},P)\big\}=Z_{\alpha_\star}+O_{\mathbb{P}}(1)$ where $Z_{\alpha_\star}$ is a non-degenerate random variable and $\tilde{\sigma}_{\alpha_n}\to\sigma_{\alpha_\star}>0$ in probability, due to Lemmas 1 and 2 (i). Therefore, the denominator of $\mathcal{T}_n(\mathcal{M},P)$ tends to $\sigma_{\alpha_\star}$ in probability and

$$\frac{\sqrt{n}\,\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P)}{\tilde{\sigma}_{\alpha_n}+\epsilon_n\tilde{\sigma}_{q,\alpha_n}} = \sqrt{n}\frac{\mathrm{MMD}^2(P_{\alpha_\star},P)}{\sigma_{\alpha_\star}} + O_{\mathbb{P}}(1).$$

Similarly, Lemma 2 (iii) and $\sqrt{n}(\alpha_n-\alpha_\star)=O_{\mathbb{P}}(1)$ yield

$$\epsilon_n\frac{\sqrt{n}\,\widehat{\mathrm{MMD}}^2_q(P_{\alpha_n},P)}{\tilde{\sigma}_{\alpha_n}+\epsilon_n\tilde{\sigma}_{q,\alpha_n}} = \epsilon_n\sqrt{n}\frac{\mathrm{MMD}^2(P_{\alpha_\star},P)}{\sigma_{\alpha_\star}} + o_{\mathbb{P}}(1).$$

This provides

$$\sqrt{n}\frac{\widehat{\mathrm{MMD}}^2_{\epsilon_n}(P_{\alpha_n},P)}{\tilde{\sigma}_{\alpha_n}+\epsilon_n\tilde{\sigma}_{q,\alpha_n}} = O_{\mathbb{P}}(1)+\sqrt{n}(1+\epsilon_n)\frac{\mathrm{MMD}^2(P_{\alpha_\star},P)}{\sigma_{\alpha_\star}},$$

which implies the claim since $\mathrm{MMD}^2(P_{\alpha_\star},P)>0$.

A.3 Technical Assumptions and Proof of Theorem 2 (Model Specification)

For a given $\delta>0$ and in addition to (18) and (19), define the classes of symmetric functions

$$\mathcal{G}:=\Big\{[\boldsymbol{x},\boldsymbol{u}]_{1:3}\mapsto g\big([\boldsymbol{x},\boldsymbol{u}]_{1:3};\alpha\big)\Big|\ \alpha\in B_\delta(\alpha_\star)\Big\},\ \text{ and}$$

$$\mathcal{Q} := \Big\{ [\boldsymbol{x}, \boldsymbol{u}]_{1:6} \mapsto \xi\big([\boldsymbol{x}, \boldsymbol{u}]_{1:6}; \alpha\big) \Big| \ \alpha \in B_\delta(\alpha_\star) \Big\},$$

where $g(\cdot, \alpha)$ and $\xi(\cdot, \alpha)$ were defined in (13) and (21), respectively.

In this section, we will assume that certain centered empirical U-processes weakly converge to their appropriate limits in some function spaces: for some families $\mathcal{L}$ as above, we will assume that

$$\big(n^{r/2} (\mathcal{U}_n^{(m)} - \mathbb{P}) \ell \big)_{\ell \in \mathcal{L}} \overset{law}{\longrightarrow} (\mathbb{G}_r \ell)_{\ell \in \mathcal{L}} \ \text{ in } \ (L_\infty(\mathcal{L}), \|\cdot\|_\infty), \tag{26}$$

where the non-negative integer $r-1$ denotes the degree of degeneracy of the class of functions $\mathcal{L}$ and $\mathbb{G}_r$ denotes a stochastic process on $\mathcal{L}$ with bounded and uniformly continuous sample paths w.r.t. the L_2-norm on the appropriate product space which is in accordance with the arguments of $\ell \in \mathcal{L}$. When (26) is satisfied, we say that $n^{r/2}(\mathcal{U}_n^{(m)} - \mathbb{P})$ indexed by $\mathcal{L}$ weakly converges. For example, when $r = 1$, $\big(n^{1/2} (\mathcal{U}_n^{(2)} - \mathbb{P}) f \big)_{f \in \mathcal{F}_q}$ weakly converges to a Gaussian process $\big(\mathbb{G}_1 f \big)_{f \in \mathcal{F}_q}$ in $L_\infty(\mathcal{F}_q)$ which has uniformly continuous sample paths w.r.t. $\|\cdot\|_{L_2(\otimes_{i=1}^4 P \otimes P_U)}$.

Sufficient conditions ensuring the weak convergence in (26) are, among others, provided in Arcones and Giné (1993, 1994), van der Vaart and Wellner (1996). In particular, there exist many sufficient conditions that do not rely on any differentiability property of the functions in $\mathcal{L}$, which makes (26) particularly useful when considering non-differentiable generating functions. Moreover, as another strength of the functional convergence (26), it also ensures that the centered empirical U-process satisfies some asymptotic equicontinuity property, as follows:

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}^* \left(\sup_{f, g \in \mathcal{L}, \|f - g\|_{L_2} < \delta} |n^{r/2} (\mathcal{U}_n^{(m)} - \mathbb{P})(f - g)| > \epsilon \right) = 0,$$

denoting $\mathbb{P}^*$ the outer probability associated with $\mathbb{P}$ (van der Vaart and Wellner, 1996, Section 1). To illustrate, when $r = 1$, the latter asymptotic equicontinuity property will allow us to "replace" expressions of the form $\sqrt{n}(\mathcal{U}_n - \mathbb{P}) h(\cdot; \alpha_n)$ by $\sqrt{n}(\mathcal{U}_n - \mathbb{P}) h(\cdot; \alpha_\star)$, since $\sqrt{n}(\mathcal{U}_n - \mathbb{P}) \left(h(\cdot; \alpha_n) - h(\cdot; \alpha_\star) \right)$ vanishes in probability due to asymptotic equicontinuity.

We impose the following assumption, which ensures that the centered empirical U-processes of interest are asymptotically equicontinuous.

Assumption 6 *There exists some $\delta > 0$ such that the empirical U-processes*

$$\big(\sqrt{n} (\mathcal{U}_n - \mathbb{P}) f \big)_{f \in \mathcal{F}} \ , \ \big(\sqrt{n} (\mathcal{U}_n^{(2)} - \mathbb{P}) q \big)_{q \in \mathcal{F}_q} \ , \ \big(\sqrt{n} (\mathcal{U}_n - \mathbb{P}) g \big)_{g \in \mathcal{G}} \ , \ \big(\sqrt{n} (\mathcal{U}_n^{(2)} - \mathbb{P}) \xi \big)_{f \in \mathcal{Q}}$$

weakly converge to their appropriate limits in the functional sense. Moreover,

$$\|h(\cdot; \alpha_n) - h(\cdot; \alpha_\star)\|_{L_2(\otimes_{i=1}^2 P \otimes P_U)} \ , \ \|q(\cdot; \alpha_n) - q(\cdot; \alpha_\star)\|_{L_2(\otimes_{i=1}^4 P \otimes P_U)} \ , \textit{ and}$$

$$\|g(\cdot; \alpha_n) - g(\cdot; \alpha_\star)\|_{L_2(\otimes_{i=1}^3 P \otimes P_U)} \ , \textit{and } \|\xi(\cdot; \alpha_n) - \xi(\cdot; \alpha_\star)\|_{L_2(\otimes_{i=1}^6 P \otimes P_U)}$$

tend to zero with n in probability.

Instead of providing technical conditions that ensure the functional convergence of our centered empirical U-processes, we have opted to impose the latter rather "high-level" assumption to avoid an excessively technical discussion. Nonetheless, we provide explicit sufficient conditions for Assumption 6 in Appendix B and verify these conditions in a ReLu-type generative neural network example in Appendix C.

To later determine the exact rates of convergence of some quantities related to our model specification test, we need to further introduce the auxiliary map

$$\tilde{g}(x,y;\alpha):\ \alpha \mapsto \frac{4}{3}\tilde{h}^2(x,y;\alpha) + \frac{8}{3}\mathbb{E}\Big[h\big((X,F(U;\alpha)),(x,y)\big)\tilde{h}\big(X,F(U;\alpha);\alpha\big)\Big], \tag{27}$$

recalling (9). Note that $\tilde{g}(\cdot;\alpha_\star) = 0$ when $P = P_{\alpha_\star}$ since $\tilde{h}(\cdot;\alpha_\star) = 0$ in this case. We impose the following regularity conditions on $\tilde{h}(x,y;\alpha)$ and $\tilde{g}(x,y;\alpha)$.

Assumption 7 *The maps $\alpha \mapsto \tilde{h}(x,y;\alpha)$ and $\alpha \mapsto \tilde{g}(x,y;\alpha)$ are twice continuously differentiable in a neighborhood of $\alpha = \alpha_\star$, for every $(x,y) \in \mathcal{S}^2$. Additionally, the maps $\iota : \alpha \mapsto \mathbb{E}\Big[\nabla_\alpha \tilde{h}\big(X,F(U;\alpha);\alpha_\star\big)\Big]$ and $\zeta : \alpha \mapsto \mathbb{E}\Big[\nabla_\alpha \tilde{g}\big(X,F(U;\alpha);\alpha_\star\big)\Big]$ are differentiable in a neighborhood of $\alpha = \alpha_\star$. Moreover, $\mathbb{E}\left[\nabla_\alpha \tilde{h}\big(X,F(U;\alpha_\star);\alpha_\star\big)\right] = \nabla_\alpha \mathbb{E}\left[\tilde{h}\big(X,F(U;\alpha_\star);\alpha_\star\big)\right]$ and $\mathbb{E}\left[\nabla_\alpha \tilde{g}\big(X,F(U;\alpha_\star);\alpha_\star\big)\right] = \nabla_\alpha \mathbb{E}\left[\tilde{g}\big(X,F(U;\alpha_\star);\alpha_\star\big)\right]$ and there exists a real constant $\delta > 0$ such that*

$$\begin{cases} \mathbb{E}\Big[\sup_{\alpha_1,\alpha_2 \in B_\delta(\alpha_\star)} \|\nabla^2_{\alpha,\alpha^\top}\tilde{h}\big(X,F(U;\alpha_2);\alpha_1\big)\|_2^2\Big] < \infty, \text{ and} \\ \mathbb{E}\Big[\sup_{\alpha_1,\alpha_2 \in B_\delta(\alpha_\star)} \|\nabla^2_{\alpha,\alpha^\top}\tilde{g}\big(X,F(U;\alpha_2);\alpha_1\big)\|_2^2\Big] < \infty. \end{cases} \tag{28}$$

For any $\alpha_0 \in \Theta_1$, the map $\nabla_\alpha \tilde{h}\big(x,F(u;\alpha_0);\alpha_\star\big)$ denotes hereafter the derivative of $\alpha \mapsto \tilde{h}\big(x,F(u;\alpha_0);\alpha\big)$ at $\alpha = \alpha_\star$. Thus, $\nabla_\alpha \tilde{h}\big(x,F(u;\alpha_\star);\alpha_\star\big)$ should not be confused with the derivative of the map $\alpha \mapsto \tilde{h}\big(x,F(u;\alpha);\alpha\big)$ at $\alpha = \alpha_\star$, that may not exist because $\alpha \mapsto F(u;\alpha)$ may not be differentiable. Similarly, $\nabla_\alpha \tilde{g}\big(x,F(u;\alpha_0);\alpha_\star\big)$ denotes the derivative of $\alpha \mapsto \tilde{g}\big(x,F(u;\alpha_0);\alpha\big)$ at $\alpha = \alpha_\star$.

Even if the map $\alpha \mapsto F(u;\alpha)$ might not be differentiable, Assumption 7 imposes that $\tilde{h}(x,y;\alpha)$ and its derivatives are sufficiently smooth w.r.t. α for any (x,y). This is legitimate because $\mathbb{E}\left[h\big((x,y),(X,F(U;\alpha))\big)\right]$ might be interpreted as a smoothing of the (possibly) non-differentiable function $h\big((x_1,y),(x_2,F(u;\alpha))\big)$. A similar statement applies to $\tilde{g}(x,y;\alpha)$. Moreover, we denote the classes of functions

$$\begin{aligned} \mathcal{F}_c := \Big\{[\boldsymbol{x},\boldsymbol{u}]_{1:2} \ &\mapsto \ h\big([\boldsymbol{x},\boldsymbol{u}]_{1:2};\alpha\big) - \tilde{h}\big(x_1,F(u_1;\alpha);\alpha\big) - \tilde{h}\big(x_2,F(u_2;\alpha);\alpha\big) \\ &\quad + \mathrm{MMD}^2(P_\alpha,P)\,\Big|\, \alpha \in B_\delta(\alpha_\star)\Big\}, \\ \mathcal{G}_c := \Big\{[\boldsymbol{x},\boldsymbol{u}]_{1:3} \ &\mapsto \ g([\boldsymbol{x},\boldsymbol{u}]_{1:3};\alpha) - \tilde{g}\big(x_1,F(u_1;\alpha);\alpha\big) - \tilde{g}\big(x_2,F(u_2;\alpha);\alpha\big) \\ &\quad - \tilde{g}\big(x_3,F(u_3;\alpha);\alpha\big) + 2\mathbb{P}g(\cdot;\alpha)\,\Big|\, \alpha \in B_\delta(\alpha_\star)\Big\}, \end{aligned}$$

$$\tilde{\mathcal{F}} := \Big\{(x,u) \mapsto \nabla_\alpha \tilde{h}\big(x,F(u;\alpha);\alpha_\star\big)\,\Big|\, \alpha \in B_\delta(\alpha_\star)\Big\}, \text{ and}$$

$$\tilde{\mathcal{G}} := \Big\{(x,u) \mapsto \nabla_\alpha \tilde{g}\big(x, F(u;\alpha);\alpha_\star\big) \,\Big|\, \alpha \in B_\delta(\alpha_\star)\Big\}.$$

We assume that the centered empirical (U-)processes indexed by the classes of functions $\mathcal{F}_c$, $\mathcal{G}_c$, $\tilde{\mathcal{F}}$ and $\tilde{\mathcal{G}}$ converge to their appropriate functional limits.

Assumption 8 *There exists some $\delta > 0$ such that the centered empirical U-processes $\big(n(\mathcal{U}_n - \mathbb{P})f\big)_{f\in\mathcal{F}_c}$ and $\big(n(\mathcal{U}_n - \mathbb{P})g\big)_{g\in\mathcal{G}_c}$, and the empirical processes $\big(\sqrt{n}(\mathbb{P}_n - \mathbb{P})f\big)_{f\in\tilde{\mathcal{F}}}$ and $\big(\sqrt{n}(\mathbb{P}_n - \mathbb{P})g\big)_{g\in\tilde{\mathcal{G}}}$ weakly converge. Further, the map $\alpha \mapsto \sigma_\alpha$ is continuously differentiable in a neighborhood of $\alpha_\star$. Finally,*

$$\mathbb{E}_{X,U}\left[\|\nabla_\alpha \tilde{h}\big(X, F(U;\alpha_n);\alpha_\star\big) - \nabla_\alpha \tilde{h}\big(X, F(U;\alpha_\star);\alpha_\star\big)\|_2^2\right], \text{ as well as}$$

$$\mathbb{E}_{X,U}\left[\|\nabla_\alpha \tilde{g}\big(X, F(U;\alpha_n);\alpha_\star\big) - \nabla_\alpha \tilde{g}\big(X, F(U;\alpha_\star);\alpha_\star\big)\|_2^2\right]$$

tend to zero in probability.

Again, we provide some sufficient conditions to ensure the functional convergence of the latter centered empirical U-processes and empirical processes in Appendix B. Now, we can derive the asymptotic distribution of our estimators of the MMD with estimated parameters.

Lemma 3 *Under Assumptions 1-2 and 6-8, the statements of Lemma 2 are valid.*

Proof [of Lemma 3]
We prove the statements (i)-(iii) of Lemma 2 successively.

(i): Note that

$$\begin{aligned}
&\sqrt{n}\big\{\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P) - \mathrm{MMD}^2(P_{\alpha_\star},P)\big\}\\
&= \sqrt{n}\big\{\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P) - \mathrm{MMD}^2(P_{\alpha_n},P) - \widehat{\mathrm{MMD}}^2(P_{\alpha_\star},P) + \mathrm{MMD}^2(P_{\alpha_\star},P)\\
&\quad + \widehat{\mathrm{MMD}}^2(P_{\alpha_\star},P) - \mathrm{MMD}^2(P_{\alpha_\star},P) + \mathrm{MMD}^2(P_{\alpha_n},P) - \mathrm{MMD}^2(P_{\alpha_\star},P)\big\}\\
&= \sqrt{n}(\mathcal{U}_n - \mathbb{P})\big(h(\cdot;\alpha_n) - h(\cdot;\alpha_\star)\big) + \sqrt{n}(\mathcal{U}_n - \mathbb{P})h(\cdot;\alpha_\star)\\
&\quad + \sqrt{n}\nabla_{\alpha^\top}\mathrm{MMD}^2(P_\alpha,P)|_{\alpha=\alpha_\star}(\alpha_n - \alpha_\star) + \sqrt{n}\, o_{\mathbb{P}}(\|\alpha_n - \alpha_\star\|),
\end{aligned}$$

by a Taylor expansion of $\alpha \mapsto \mathrm{MMD}^2(P_\alpha,P)$. Since the process $\sqrt{n}(\mathcal{U}_n - \mathbb{P})$, indexed by $\mathcal{F}$, is asymptotically equicontinuous by Assumption 6 and the convergence $\|h(\cdot;\alpha_n) - h(\cdot;\alpha_\star)\|_{L_2(P\otimes P_U\otimes P\otimes P_U)} \to 0$ in probability holds, we get that $\sqrt{n}(\mathcal{U}_n - \mathbb{P})\big(h(\cdot;\alpha_n) - h(\cdot;\alpha_\star)\big)$ is $o_{\mathbb{P}}(1)$. Moreover, by Arcones and Giné (1993, Theorem 4.9), $\sqrt{n}(\mathcal{U}_n - \mathbb{P})h(\cdot;\alpha_\star)$ converges in law towards a normal random variable whose variance is

$$\sigma^2_{\alpha_\star} = \mathrm{Var}\Big(2\mathbb{E}_{X_2,U_2}\big[h\big([\boldsymbol{X},\boldsymbol{U}]_{1:2};\alpha_\star\big)|\, X_1,U_1\big]\Big).$$

By assumption, $\sqrt{n}\nabla_{\alpha^\top}\mathrm{MMD}^2(P_{\alpha_\star},P)(\alpha_n - \alpha_\star)$ converges to $\nabla_{\alpha^\top}\mathrm{MMD}^2(P_{\alpha_\star},P)V_{\alpha_\star}$ and the joint convergence follows from our assumptions. The fourth term is clearly $o_{\mathbb{P}}(1)$, since $\|\alpha_n - \alpha_\star\| = O_{\mathbb{P}}(n^{-1/2})$. Thus, we have

$$\sqrt{n}\big\{\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P) - \mathrm{MMD}^2(P_{\alpha_\star},P)\big\} \xrightarrow{\text{law}} Z_{\alpha_\star} + \nabla_{\alpha^\top}\mathrm{MMD}^2(P_\alpha,P)|_{\alpha=\alpha_\star}V_{\alpha_\star}.$$

Furthermore, $\hat{\sigma}^2_{\alpha_n} = \mathcal{U}_n g(\cdot;\alpha_n) - 4\Big\{\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P)\Big\}^2$. We have just deduced the convergence of $\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P)$ to $\mathrm{MMD}^2(P_{\alpha_\star},P)$ in probability. Similarly, since $\sqrt{n}(\mathcal{U}_n - \mathbb{P})$ indexed by $\mathcal{G}$ converges to a Gaussian limit and $\|g(\cdot;\alpha_n) - g(\cdot;\alpha_\star)\|_{L_2(\otimes_{i=1}^3 P\otimes P_U)} \to 0$ in probability, we have

$$\begin{aligned}&\mathcal{U}_n g(\cdot;\alpha_n) = \mathbb{E}_{[\boldsymbol{X}',\boldsymbol{U}']_{1:3}}\big[g(\cdot;\alpha_n)\big([\boldsymbol{X}',\boldsymbol{U}']_{1:3}\big)\big] + o_{\mathbb{P}}(1), \text{ and}\\ &\Big|\mathbb{E}_{[\boldsymbol{X}',\boldsymbol{U}']_{1:3}}\big[g(\cdot;\alpha_n)\big([\boldsymbol{X}',\boldsymbol{U}']_{1:3}\big) - g\big([\boldsymbol{X}',\boldsymbol{U}']_{1:3};\alpha_\star\big)\big]\Big|\\ &\le \|g(\cdot;\alpha_n) - g(\cdot;\alpha_\star)\|_{L_2(\otimes_{i=1}^3 P\otimes P_U)} \xrightarrow{\mathbb{P}} 0,\end{aligned}$$

where $[\boldsymbol{X}',\boldsymbol{U}']_{1:3}$ is an independent copy of $[\boldsymbol{X},\boldsymbol{U}]_{1:3}$. Thus, this yields

$$\mathcal{U}_n g(\cdot;\alpha_n) = \mathbb{E}\big[g\big([\boldsymbol{X},\boldsymbol{U}]_{1:3};\alpha_\star\big)\big] + o_{\mathbb{P}}(1),$$

proving that $\hat{\sigma}^2_{\alpha_n} \to \sigma^2_{\alpha_\star}$ in probability. Therefore, we also have $\tilde{\sigma}^2_{\alpha_n} \to \sigma^2_{\alpha_\star}$ in probability by Lemma 1.

(ii): Obviously, when $P_{\alpha_\star} = P$, $\nabla_\alpha \mathrm{MMD}^2(P_\alpha,P)|_{\alpha=\alpha_\star} = 0$ since $\alpha_\star$ belongs to the interior of Θ_1 and is an argmin of $\alpha \mapsto \mathrm{MMD}^2(P_\alpha,P)$. Moreover, $\mathcal{U}_n h(\cdot;\alpha_\star) = O_{\mathbb{P}}(n^{-1})$ by standard U-statistic arguments, because $h(\cdot;\alpha_\star)$ is now a degenerate U-statistic kernel. Note that, by Assumption 2, $\sqrt{n}\,\mathrm{MMD}^2(P_{\alpha_n},P) = O_{\mathbb{P}}(n^{-1/2})$ since

$$\begin{aligned}\mathrm{MMD}^2(P_{\alpha_n},P) &= (\alpha_n - \alpha_\star)^\top \nabla^2_{\alpha,\alpha^\top}\mathrm{MMD}^2(P_\alpha,P)|_{\alpha=\alpha_\star}(\alpha_n-\alpha_\star) + o_{\mathbb{P}}\left(\|\alpha_n - \alpha_\star\|^2\right)\\ &= O_{\mathbb{P}}\left(\|\alpha_n-\alpha_\star\|^2\right),\end{aligned}$$

which is $O_{\mathbb{P}}(n^{-1})$. Let us decompose the quantity of interest as

$$\sqrt{n}\,\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P) = \sqrt{n}\,\mathcal{U}_n h(\cdot;\alpha_n) = \sqrt{n}\,\mathcal{U}_n\big(h(\cdot;\alpha_n) - h(\cdot;\alpha_\star)\big) + O_{\mathbb{P}}(n^{-1/2}).$$

Thus, it remains to prove that $\sqrt{n}\,\mathcal{U}_n\big(h(\cdot;\alpha_n) - h(\cdot;\alpha_\star)\big) = O_{\mathbb{P}}(n^{-1/2})$. To this aim, we will use the asymptotic equicontinuity of degenerate U-statistic kernels. Obviously, for any $\alpha \neq \alpha_\star$, $h(\cdot;\alpha) - h(\cdot;\alpha_\star)$ is generally not a degenerate U-statistic kernel. Nonetheless, we can rewrite

$$\begin{aligned}\sqrt{n}\,\mathcal{U}_n\big(h(\cdot;\alpha_n) - h(\cdot;\alpha_\star)\big) &= \frac{\sqrt{n}}{n(n-1)}\sum_{\substack{i,j=1\\ i\neq j}}^n \big\{\psi\big([\boldsymbol{X},\boldsymbol{U}]_{i,j};\alpha_n\big) - \psi\big([\boldsymbol{X},\boldsymbol{U}]_{i,j};\alpha_\star\big)\big\}\\ &\quad + \frac{2\sqrt{n}}{n}\sum_{i=1}^n \tilde{h}\big(X_i, F(U_i;\alpha_n);\alpha_n\big) + O_{\mathbb{P}}(n^{-1/2}),\end{aligned} \tag{29}$$

$$\psi\big([\boldsymbol{x},\boldsymbol{u}]_{1:2};\alpha\big) := h\big([\boldsymbol{x},\boldsymbol{u}]_{1:2};\alpha\big) - \tilde{h}\big(x_1,F(u_1;\alpha);\alpha\big) - \tilde{h}\big(x_2,F(u_2;\alpha);\alpha\big) + \mathrm{MMD}^2(P_\alpha,P).$$

Note that $\psi\big([\boldsymbol{x},\boldsymbol{u}]_{1:2};\alpha_\star\big) = h\big([\boldsymbol{x},\boldsymbol{u}]_{1:2};\alpha_\star\big)$ because $\tilde{h}(x,y;\alpha_\star) = 0$ for every (x,y), when $P = P_{\alpha_\star}$. Moreover, the U-statistic $\mathcal{U}_n\psi(\cdot;\alpha)$ is now degenerate, for every $\alpha \in \Theta_1$.

For any $\epsilon > 0$ and any positive constants δ_k, $k \in \{1,2\}$, we get

$$\begin{aligned}
&\mathbb{P}\Big(\Big|\frac{\sqrt{n}}{n(n-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{n}\big\{\psi\big([\boldsymbol{X},\boldsymbol{U}]_{i,j};\alpha_n\big)-\psi\big([\boldsymbol{X},\boldsymbol{U}]_{i,j};\alpha_\star\big)\big\}\Big| > n^{-1/2}\epsilon\Big)\\
&\leq \mathbb{P}\Big(\sup_{f,g\in\mathcal{F}_c \|f-g\|_{L_2(\otimes_{i=1}^2 P\otimes P_u)}<\delta_2}\big|n(\mathcal{U}_n-\mathbb{P})(f-g)\big| > \epsilon\Big)+\mathbb{P}(\|\alpha_n-\alpha_\star\|\geq\delta_1)\\
&\quad+\mathbb{P}\Big(\|\psi(\cdot;\alpha_n)-\psi(\cdot;\alpha_\star)\|_{L_2(\otimes_{i=1}^2(P\otimes P_U))}\geq\delta_2, \|\alpha_n-\alpha_\star\|<\delta_1\Big) =: T_1+T_2+T_3.
\end{aligned}$$

By the asymptotic equicontinuity of $n(\mathcal{U}_n-\mathbb{P})$ indexed by $\mathcal{F}_c$ (Assumption 8), there exists δ_2 s.t. T_1 is arbitrarily small for n sufficiently large. Since

$$\|\psi(\cdot;\alpha_n)-\psi(\cdot;\alpha_\star)\|_{L_2(\otimes_{i=1}^2(P\otimes P_U))}\leq 4\|h(\cdot;\alpha_n)-h(\cdot;\alpha_\star)\|_{L_2(\otimes_{i=1}^2(P\otimes P_U))}$$

that tends to zero in probability, δ_1 can be chosen so that T_2 and T_3 are arbitrarily small for sufficiently large n. Globally, we have obtained that

$$\frac{\sqrt{n}}{n(n-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{n}\big\{\psi\big([\boldsymbol{X},\boldsymbol{U}]_{i,j};\alpha_n\big)-\psi\big([\boldsymbol{X},\boldsymbol{U}]_{i,j};\alpha_\star\big)\big\} = o_{\mathbb{P}}(n^{-1/2}). \tag{30}$$

Moreover, since $\tilde{h}(x,y;\alpha)$ is twice continuously differentiable w.r.t. α by Assumption 7, we obtain by a limited expansion

$$\begin{aligned}
&\frac{2\sqrt{n}}{n}\sum_{i=1}^{n}\tilde{h}\big(X_i,F(U_i;\alpha_n);\alpha_n\big)\\
&=\frac{2\sqrt{n}}{n}\sum_{i=1}^{n}\Big\{\tilde{h}\big(X_i,F(U_i;\alpha_n);\alpha_\star\big)+\nabla_{\alpha^\top}\tilde{h}\big(X_i,F(U_i;\alpha_n);\alpha_\star\big)(\alpha_n-\alpha_\star)\\
&\quad+2^{-1}(\alpha_n-\alpha_\star)^\top\nabla^2_{\alpha,\alpha^\top}\tilde{h}\big(X_i,F(U_i;\alpha_n);\tilde{\alpha}_n\big)(\alpha_n-\alpha_\star)\Big\},
\end{aligned}$$

for some random parameter $\tilde{\alpha}_n$ that satisfies $\|\tilde{\alpha}_n-\alpha_\star\|<\|\alpha_n-\alpha_\star\|$. Due to (28) from Assumption 7 and since $\tilde{h}_{\alpha_\star}=0$ under the null, this yields

$$\begin{aligned}
&\frac{2\sqrt{n}}{n}\sum_{i=1}^{n}\tilde{h}\big(X_i,F(U_i;\alpha_n);\alpha_n\big)\\
&= \frac{2\sqrt{n}}{n}\sum_{i=1}^{n}\nabla_{\alpha^\top}\tilde{h}\big(X_i,F(U_i;\alpha_n);\alpha_\star\big)(\alpha_n-\alpha_\star)+O_{\mathbb{P}}(n^{-1/2})\\
&= \sqrt{n}(\alpha_n-\alpha_\star)^\top\frac{2}{n}\sum_{i=1}^{n}\Big\{\nabla_\alpha\tilde{h}\big(X_i,F(U_i;\alpha_n);\alpha_\star\big)-\nabla_\alpha\tilde{h}\big(X_i,F(U_i;\alpha_\star);\alpha_\star\big)\Big\}\\
&\quad+\sqrt{n}(\alpha_n-\alpha_\star)^\top\frac{2}{n}\sum_{i=1}^{n}\nabla_\alpha\tilde{h}\big(X_i,F(U_i;\alpha_\star);\alpha_\star\big)+O_{\mathbb{P}}(n^{-1/2})\\
&= 2\sqrt{n}(\alpha_n-\alpha_\star)^\top\mathbb{P}_n\Big\{\nabla_\alpha\tilde{h}\big(\cdot,F(\cdot;\alpha_n);\alpha_\star\big)-\nabla_\alpha\tilde{h}\big(\cdot,F(\cdot;\alpha_\star);\alpha_\star\big)\Big\}\\
&\quad+2\sqrt{n}(\alpha_n-\alpha_\star)^\top\mathbb{P}_n\nabla_\alpha\tilde{h}\big(\cdot,F(\cdot;\alpha_\star);\alpha_\star\big)+O_{\mathbb{P}}(n^{-1/2}).
\end{aligned}$$

Observe that $\mathbb{E}\Big[\nabla_\alpha \tilde{h}\big(X, F(U;\alpha_\star);\alpha_\star\big)\Big] = \nabla_\alpha \mathbb{E}\Big[\tilde{h}\big(X, F(U;\alpha_\star);\alpha_\star\big)\Big] = \nabla_\alpha 0 = 0$ due to Assumption 7 and $P_{\alpha_\star} = P$. Moreover, as the map $\alpha \mapsto \iota(\alpha) = \mathbb{E}\Big[\nabla_\alpha \tilde{h}\big(X, F(U;\alpha);\alpha_\star\big)\Big]$ is assumed to be differentiable at $\alpha = \alpha_\star$ (Assumption 7), a Taylor expansion provides

$$\|\iota(\alpha_n)\|_1 = \|\iota(\alpha_\star) + (\alpha_n - \alpha_\star)^\top \nabla_\alpha \iota(\alpha_\star)\|_1 + o_{\mathbb{P}}(\|\alpha_n - \alpha_\star\|) = 0 + O_{\mathbb{P}}(n^{-1/2}).$$

Deduce

$$\|\mathbb{P}\Big\{\nabla_\alpha \tilde{h}\big(\cdot, F(\cdot;\alpha_n);\alpha_\star\big) - \nabla_\alpha \tilde{h}\big(\cdot, F(\cdot;\alpha_\star);\alpha_\star\big)\Big\}\|_1 = \|\iota(\alpha_n) - \iota(\alpha_\star)\|_1 = O_{\mathbb{P}}(n^{-1/2}).$$

Therefore, we get

$$\begin{aligned}&\frac{2\sqrt{n}}{n}\sum_{i=1}^n \tilde{h}\big(X_i, F(U_i;\alpha_n);\alpha_n\big) = 2\sqrt{n}(\alpha_n - \alpha_\star)^\top (\mathbb{P}_n - \mathbb{P})\nabla_\alpha \tilde{h}\big(\cdot, F(\cdot;\alpha_\star);\alpha_\star\big)\\ &+ 2\sqrt{n}(\alpha_n - \alpha_\star)^\top (\mathbb{P}_n - \mathbb{P})\Big\{\nabla_\alpha \tilde{h}\big(\cdot, F(\cdot;\alpha_n);\alpha_\star\big) - \nabla_\alpha \tilde{h}\big(\cdot; F(\cdot;\alpha_\star);\alpha_\star\big)\Big\} + O_{\mathbb{P}}(n^{-1/2}).\end{aligned}$$

Since $\mathbb{E}_{X,U}\Big[\|\nabla_\alpha \tilde{h}\big(X, F(U;\alpha_n);\alpha_\star\big) - \nabla_\alpha \tilde{h}\big(X, F(U;\alpha_\star);\alpha_\star\big)\|_2^2\Big] \overset{\mathbb{P}}{\to} 0$ and $\sqrt{n}\big((\mathbb{P}_n - \mathbb{P})f\big)_{f\in\tilde{\mathcal{F}}}$ converges to a tight Gaussian limit (Assumption 8), the asymptotic equicontinuity of the latter process yields

$$2\sqrt{n}(\alpha_n - \alpha_\star)^\top (\mathbb{P}_n - \mathbb{P})\Big\{\nabla_\alpha \tilde{h}\big(\cdot, F(\cdot;\alpha_n);\alpha_\star\big) - \nabla_\alpha \tilde{h}\big(\cdot, F(\cdot;\alpha_\star);\alpha_\star\big)\Big\} = o_{\mathbb{P}}(n^{-1/2}).$$

Moreover, the usual CLT yields $\sqrt{n}(\alpha_n - \alpha_\star)^\top (\mathbb{P}_n - \mathbb{P})\nabla_\alpha \tilde{h}\big(\cdot, F(\cdot;\alpha_\star);\alpha_\star\big) = O_{\mathbb{P}}(n^{-1/2})$. In other words, we have obtained

$$\frac{2\sqrt{n}}{n}\sum_{i=1}^n \tilde{h}\big(X_i, F(U_i;\alpha_n);\alpha_n\big) = O_{\mathbb{P}}(n^{-1/2}). \tag{31}$$

Therefore, by (29), (30) and (31), we get $\sqrt{n}\,\mathcal{U}_n\big(h(\cdot;\alpha_n) - h(\cdot;\alpha_\star)\big) = O_{\mathbb{P}}(n^{-1/2})$ and $\widehat{\mathrm{MMD}}^2(P_{\alpha_n}, P) = O_{\mathbb{P}}(n^{-1})$.

Furthermore, note that we have also just proven that

$$\hat{\sigma}^2_{\alpha_n} = \mathcal{U}_n g(\cdot;\alpha_n) - 4\big(\widehat{\mathrm{MMD}}^2(P_{\alpha_n}, P)\big)^2 = \mathcal{U}_n g(\cdot;\alpha_n) + O_{\mathbb{P}}(n^{-2}).$$

Thus, it remains to show that $\mathcal{U}_n g(\cdot;\alpha_n) = O_{\mathbb{P}}(n^{-1})$ when $P_{\alpha_\star} = P$. To this aim, observe that

$$\begin{aligned}&3\mathbb{E}\left[g\big([\boldsymbol{X},\boldsymbol{U}]_{1:3};\alpha_\star\big) \mid X_1, U_1\right]/4 = \tilde{h}^2\big(X_1, F(U_1;\alpha_\star);\alpha_\star\big)\\ &\quad + \mathbb{E}\left[\mathbb{E}\left[h\big([\boldsymbol{X},\boldsymbol{U}]_{2,1};\alpha_\star\big)h\big([\boldsymbol{X},\boldsymbol{U}]_{2,3};\alpha_\star\big) \mid [\boldsymbol{X},\boldsymbol{U}]_{1,2}\right] \mid X_1, U_1\right]\\ &\quad + \mathbb{E}\left[\mathbb{E}\left[h\big([\boldsymbol{X},\boldsymbol{U}]_{3,2};\alpha_\star\big)h\big([\boldsymbol{X},\boldsymbol{U}]_{3,1};\alpha_\star\big) \mid [\boldsymbol{X},\boldsymbol{U}]_{1,3}\right] \mid X_1, U_1\right]\\ &= \tilde{h}^2\big(X_1, F(U_1;\alpha_\star);\alpha_\star\big) + 2\mathbb{E}\left[h\big([\boldsymbol{X},\boldsymbol{U}]_{2,1};\alpha_\star\big)\tilde{h}\big(X_2, F(U_2;\alpha_\star);\alpha_\star\big) \mid X_1, U_1\right] = 0,\end{aligned}$$

since $\tilde{h}(\cdot,\cdot;\alpha_\star) = 0$ when $P_{\alpha_\star} = P$, i.e., $g(\cdot;\alpha_\star)$ is a degenerate U-statistic kernel. Therefore, as a degenerate U-statistic, $\big(\mathcal{U}_n - \mathbb{P}\big)g(\cdot;\alpha_\star)$ is $O_{\mathbb{P}}(n^{-1})$. Since $\mathbb{P}g(\cdot;\alpha_n) = \sigma^2_{\alpha_n} + O_{\mathbb{P}}(n^{-1})$, deduce

$$\mathcal{U}_n g(\cdot;\alpha_n) = (\mathcal{U}_n - \mathbb{P})g(\cdot;\alpha_\star) + (\mathcal{U}_n - \mathbb{P})(g(\cdot;\alpha_n) - g(\cdot;\alpha_\star)) + \sigma^2_{\alpha_n} + O_{\mathbb{P}}(n^{-1}).$$

Since $\alpha \mapsto \sigma_\alpha$ is continuously differentiable in a neighborhood of $\alpha_\star$, there exists $\bar{\alpha} \in \Theta_1$ s.t. $\sigma_\alpha = (\alpha - \alpha_\star)^\top \nabla_\alpha \sigma_{\bar{\alpha}}$ and $\|\bar{\alpha} - \alpha_\star\| < \|\alpha - \alpha_\star\|$. This yields $\sigma^2_{\alpha_n} = O_{\mathbb{P}}\big(\|\alpha_n - \alpha_\star\|^2\big) = O_{\mathbb{P}}(n^{-1})$. We have obtained $\mathcal{U}_n g(\cdot;\alpha_n) = (\mathcal{U}_n - \mathbb{P})(g(\cdot;\alpha_n) - g(\cdot;\alpha_\star)) + O_{\mathbb{P}}(n^{-1})$. By the same type of decomposition as above, we get

$$\begin{aligned}\sqrt{n}(\mathcal{U}_n - \mathbb{P})\big(g(\cdot;\alpha_n) - g(\cdot;\alpha_\star)\big) &= \frac{\sqrt{n}}{n(n-1)(n-2)} \sum_{\substack{i,j,k=1\\ i\neq j\neq k}}^n \Big\{ g\big([\boldsymbol{X},\boldsymbol{U}]_{i,j,k};\alpha_n\big) \\ &- \tilde{g}\big(X_i, F(U_i;\alpha_n);\alpha_n\big) - \tilde{g}\big(X_j, F(U_j;\alpha_n);\alpha_n\big) - \tilde{g}\big(X_k, F(U_k;\alpha_n);\alpha_n\big) + 2\mathbb{P}g(\cdot;\alpha_n) \\ &- g\big(X_i, F(U_i;\alpha_\star), X_j, F(U_j;\alpha_\star), X_k, F(U_k;\alpha_\star)\big)\Big\} - 3\sqrt{n}\mathbb{P}g(\cdot;\alpha_n) \\ &+ \frac{3\sqrt{n}}{n}\sum_{i=1}^n \tilde{g}\big(X_i, F(U_i;\alpha_n);\alpha_n\big).\end{aligned}$$

By the asymptotic equicontinuity of the degenerate process $n(\mathcal{U}_n - \mathbb{P})$ indexed by $\mathcal{G}_c$, we obtain

$$\sqrt{n}(\mathcal{U}_n - \mathbb{P})\big(g(\cdot;\alpha_n) - g(\cdot;\alpha_\star)\big) = \frac{3\sqrt{n}}{n}\sum_{i=1}^n \tilde{g}\big(X_i, F(U_i;\alpha_n);\alpha_n\big) + O_{\mathbb{P}}(n^{-1/2}).$$

Now, since $\alpha \mapsto \tilde{g}_\alpha(x,y)$ is twice continuously differentiable, a limited expansion yields

$$\begin{aligned}\frac{3\sqrt{n}}{n}\sum_{i=1}^n \tilde{g}\big(X_i, F(U_i;\alpha_n);\alpha_n\big) &= \sqrt{n}(\alpha_n - \alpha_\star)\frac{3}{n}\sum_{i=1}^n \nabla_\alpha \tilde{g}\big(X_i, F(U_i;\alpha_\star);\alpha_\star\big) \\ &+ \frac{3(\alpha_n - \alpha_\star)}{\sqrt{n}}\sum_{i=1}^n \Big\{\nabla_\alpha \tilde{g}\big(X_i, F(U_i;\alpha_n);\alpha_\star\big) - \nabla_\alpha \tilde{g}\big(X_i, F(U_i;\alpha_\star);\alpha_\star\big)\Big\} + O_{\mathbb{P}}(n^{-1/2}).\end{aligned}$$

Note we have invoked Assumption 7 to manage the remainder term. Finally, due to the asymptotic equicontinuity of the empirical U process indexed by $\tilde{\mathcal{G}}$ and the usual CLT, we obtain $\sum_{i=1}^n \tilde{g}\big(X_i, F(U_i;\alpha_n);\alpha_n\big)/\sqrt{n} = O_{\mathbb{P}}(n^{-1/2})$. This yields $\hat{\sigma}^2_{\alpha_n} = O_{\mathbb{P}}(n^{-1})$ and therefore $\tilde{\sigma}^2_{\alpha_n} = O_{\mathbb{P}}(n^{-1})$ also by Lemma 1.

(iii): Follows exactly by the same arguments as the proof of *(i)*.

■

Note that the proof of Theorem 1 solely relied on Lemma 2, which is why Theorem 2 immediately follows from Lemma 3.

A.4 Proof of Theorem 3 (Model Comparison)

If the maps $\alpha \mapsto F(u;\alpha)$ and $\beta \mapsto G(v;\beta)$ are twice differentiable for every $u \in \mathcal{U}$ and $v \in \mathcal{V}$, the same arguments as in Section 3.3 can be invoked to obtain the asymptotic behaviors of $\mathcal{T}_n(\mathcal{M}_1, \mathcal{M}_2, P)$'s numerator and denominator. As for Lemmas 2 and 3, we establish a more general result than we need. To this aim, we require the following joint convergence assumption.

Assumption 9 *When $n \to \infty$,*

$$\sqrt{n}\begin{pmatrix} \widehat{\mathrm{MMD}}^2(P_{\alpha_\star}, P) - \mathrm{MMD}^2(P_{\alpha_\star}, P) \\ \widehat{\mathrm{MMD}}^2(Q_{\beta_\star}, P) - \mathrm{MMD}^2(Q_{\beta_\star}, P) \\ \widehat{\mathrm{MMD}}_q^2(P_{\alpha_\star}, P) - \mathrm{MMD}^2(P_{\alpha_\star}, P) \\ \widehat{\mathrm{MMD}}_q^2(Q_{\beta_\star}, P) - \mathrm{MMD}^2(Q_{\beta_\star}, P) \\ \alpha_n - \alpha_\star \\ \beta_n - \beta_\star \end{pmatrix} \xrightarrow{law} \begin{pmatrix} Z^{(1)}_{\alpha_\star} \\ Z^{(2)}_{\beta_\star} \\ Z^{(1)}_{q,\alpha_\star} \\ Z^{(2)}_{q,\beta_\star} \\ V^{(1)}_{\alpha_\star} \\ V^{(2)}_{\beta_\star} \end{pmatrix}.$$

When $\alpha_\star$ (resp. $\beta_\star$) minimizes the distance $\mathrm{MMD}(P_\alpha, P)$ *over Θ_1 (resp.* $\mathrm{MMD}(Q_\beta, P)$ *over Θ_2), the weak convergence of $\sqrt{n}(\alpha_n - \alpha_\star)$ (resp. $\sqrt{n}(\beta_n - \beta_\star)$) is no longer required and replaced by the weaker requirement $\sqrt{n}(\alpha_n - \alpha_\star) = O_{\mathbb{P}}(1)$ (resp. $\sqrt{n}(\beta_n - \beta_\star) = O_{\mathbb{P}}(1)$).*

Lemma 4 *Assume that Assumptions 1-5 and 9 are satisfied by the competing models $\mathcal{M}_1$ and $\mathcal{M}_2$. Then the following is true:*

(i)

$$\sqrt{n}\{\widehat{\mathrm{MMD}}^2(P_{\alpha_n}, P) - \mathrm{MMD}^2(P_{\alpha_\star}, P) - \widehat{\mathrm{MMD}}^2(Q_{\beta_n}, P) + \mathrm{MMD}^2(Q_{\beta_\star}, P)\}$$
$$\xrightarrow{law} W_{\alpha_\star,\beta_\star} + \nabla_{\alpha^\top} \mathrm{MMD}^2(P_\alpha, P)|_{\alpha=\alpha_\star} V^{(1)}_{\alpha_\star} - \nabla_{\beta^\top} \mathrm{MMD}^2(Q_\beta, P)|_{\beta=\beta_\star} V^{(2)}_{\beta_\star},$$

where $W_{\alpha_\star,\beta_\star} := Z^{(1)}_{\alpha_\star} - Z^{(2)}_{\beta_\star} \sim \mathcal{N}\big(0, \sigma^2_{\alpha_\star,\beta_\star}\big)$. Moreover, $\tilde{\sigma}^2_{\alpha_n,\beta_n} \to \sigma^2_{\alpha_\star,\beta_\star}$ in probability.

(ii) If $P_{\alpha_\star} = Q_{\beta_\star} = P$, then $\sigma_{\alpha_\star,\beta_\star} = 0$. Moreover, $\tilde{\sigma}^2_{\alpha_n,\beta_n} = O_{\mathbb{P}}(n^{-1})$ and

$$\sqrt{n}\{\widehat{\mathrm{MMD}}^2(P_{\alpha_n}, P) - \widehat{\mathrm{MMD}}^2(Q_{\beta_n}, P)\} = O_{\mathbb{P}}\big(n^{-1/2}\big).$$

(iii)

$$\sqrt{n}\{\widehat{\mathrm{MMD}}_q^2(P_{\alpha_n}, P) - \mathrm{MMD}^2(P_{\alpha_\star}, P) - \widehat{\mathrm{MMD}}_q^2(Q_{\beta_n}, P) + \mathrm{MMD}^2(Q_{\beta_\star}, P)\}$$
$$\xrightarrow{law} W_{q,\alpha_\star,\beta_\star} + \nabla_{\alpha^\top} \mathrm{MMD}^2(P_\alpha, P)|_{\alpha=\alpha_\star} V^{(1)}_{\alpha_\star} - \nabla_{\beta^\top} \mathrm{MMD}^2(Q_\beta, P)|_{\beta=\beta_\star} V^{(2)}_{\beta_\star},$$

where $W_{q,\alpha_\star,\beta_\star} := Z^{(1)}_{q,\alpha_\star} - Z^{(2)}_{q,\beta_\star} \sim \mathcal{N}\big(0, \sigma^2_{q,\alpha_\star,\beta_\star}\big)$. If the samples $(U_i)_{i\geq 1}$ and $(V_i)_{i\geq 1}$ are independent, then $\sigma^2_{q,\alpha_\star,\beta_\star} > 0$. Finally, $\tilde{\sigma}^2_{q,\alpha_n,\beta_n} \to \sigma^2_{q,\alpha_\star,\beta_\star}$ in probability.

Proof [of Lemma 4] As in Lemma 2, the proof is essentially based on two Taylor expansions. Recall that $\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P)-\widehat{\mathrm{MMD}}^2(Q_{\beta_n},P)=\mathcal{U}_n h(\cdot;\alpha_n,\beta_n)$. By a Taylor expansion of the U-kernel $h(\cdot;\alpha_n,\beta_n)$ around $(\alpha_\star,\beta_\star)$ and due to Assumptions 3-4 for both models, we have

$$\begin{aligned} h(\cdot;\alpha_n,\beta_n) &= h(\cdot;\alpha_\star,\beta_\star)+\nabla_{\alpha^\intercal} h(\cdot;\alpha_\star,\beta_\star).(\alpha_n-\alpha_\star)+\nabla_{\beta^\intercal} h(\cdot;\alpha_\star,\beta_\star).(\beta_n-\beta_\star) \\ &\quad + O_{\mathbb{P}}(\|\alpha_n-\alpha_\star\|^2)+O_{\mathbb{P}}(\|\beta_n-\beta_\star\|^2). \end{aligned} \tag{32}$$

The convergence of $\sqrt{n}\big\{\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P)-\mathrm{MMD}^2(P_{\alpha_\star},P)-\widehat{\mathrm{MMD}}^2(Q_{\beta_n},P)+\mathrm{MMD}^2(Q_{\beta_\star},P)\big\}$ immediately follows from Assumption 9. We prove the convergence of the estimated variance $\hat\sigma^2_{\alpha_n,\beta_n}$ defined in (22) using a first order Taylor expansion of the map $(\alpha,\beta)\mapsto g\big([\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{i,j,k};\alpha,\beta\big)$ defined in (23) around $(\alpha_\star,\beta_\star)$, for every triplet of indices (i,j,k), $i\neq j\neq k$. By Lemma 1, we get the convergence of $\tilde\sigma_{\alpha_n,\beta_n}$ to $\sigma_{\alpha_\star,\beta_\star}$. This yields (i).

In the case (ii), $P_{\alpha_\star}=P=Q_{\beta_\star}$ and $\alpha_\star$ and $\beta_\star$ are minimizers of $\alpha\mapsto\mathrm{MMD}(P_\alpha,P)$ and $\beta\mapsto\mathrm{MMD}(Q_\beta,P)$ respectively. By direct calculation and recalling (14), we easily obtain $\sigma_{\alpha_\star,\beta_\star}=0$. Then, (32) provides $\mathcal{U}_n h(\cdot;\alpha_n,\beta_n)=O_{\mathbb{P}}(n^{-1})$ because

$$\mathbb{E}\Big[\nabla_\alpha h([X,U,V]_{1,2};\alpha_\star,\beta_\star)\Big]=\nabla_\alpha\mathbb{E}\Big[h([X,U,V]_{1,2};\alpha_\star,\beta_\star)\Big]=\nabla_\alpha\mathrm{MMD}(P_{\alpha_\star},P)=0,$$

and similarly $\mathbb{E}\left[\nabla_\beta h([X,U,V]_{1,2};\alpha_\star,\beta_\star)\right]=0$. Moreover, a Taylor expansion of the U-kernel $g(\cdot;\alpha_n,\beta_n)$ around $(\alpha_\star,\beta_\star)$ yields that $\mathcal{U}_n g(\cdot;\alpha_n,\beta_n)=O_{\mathbb{P}}(n^{-1})$ when $P_{\alpha_\star}=P=Q_{\beta_\star}$ since $\mathcal{U}_n g(\cdot;\alpha_\star,\beta_\star)$ is a degenerate U-statistic, as in the proof of Lemma 2 (ii). This implies $\hat\sigma^2_{\alpha_n,\beta_n}=\mathcal{U}_n g(\cdot;\alpha_n,\beta_n)-4\big(\widehat{\mathrm{MMD}}^2(P_{\alpha_n},P)-\widehat{\mathrm{MMD}}^2(Q_{\beta_n},P)\big)^2=O_{\mathbb{P}}(n^{-1})$.

The same reasonings apply to prove (iii), noting that $\widehat{\mathrm{MMD}}^2_q(P_{\alpha_n},P)-\widehat{\mathrm{MMD}}^2_q(Q_{\beta_n},P)=\mathcal{U}_n^{(2)}q(\cdot;\alpha_n,\beta_n)$. In particular, mimic the same arguments as above for the U-kernels $q(\cdot;\alpha_n,\beta_n)$ and $\xi(\cdot;\alpha_n,\beta_n)$, recalling (15) and (24). This yields $\sqrt{n}\big\{\mathcal{U}_n^{(2)}q(\cdot;\alpha_n,\beta_n)-\big(\mathrm{MMD}(P_{\alpha_\star},P)-\mathrm{MMD}(Q_{\beta_\star},P)\big)\big\}\to\mathcal{N}(0,\sigma^2_{q,\alpha_\star,\beta_\star})$ and $\hat\sigma_{q,\alpha_n,\beta_n}\to\sigma_{q,\alpha_\star,\beta_\star}>0$. ■

The proof of Theorem 3 is a direct consequence of Lemma 4, mimicking the proof of Theorem 1. Note that we only need the joint weak convergence of the first four components of the random vector in Assumption 9, that is obtained by the Cramer-Wold device and usual Hájek projections of U-statistics based on the sample (X_i,U_i,V_i). Contrary to Theorem 1, one must differentiate the case $P_{\alpha_\star}=P=Q_{\beta_\star}$ and the case $\mathrm{MMD}(P_{\alpha_\star},P)=\mathrm{MMD}(Q_{\beta_\star})$, but $P_{\alpha_\star}\neq P\neq Q_{\beta_\star}$ under the null. When $P_{\alpha_\star}=P=Q_{\beta_\star}$, the proof is identical to the proof of Theorem 1. When $\mathrm{MMD}(P_{\alpha_\star},P)=\mathrm{MMD}(Q_{\beta_\star})$ but $P_{\alpha_\star}\neq P\neq Q_{\beta_\star}$ the proof structure is still identical to the latter case, but with the sole differences that $\sqrt{n}\,\mathcal{U}_n h(\cdot;\alpha_n,\beta_n)\to\mathcal{N}(0,\sigma^2_{\alpha_\star,\beta_\star})$, $\tilde\sigma_{\alpha_n,\beta_n}\to\sigma_{\alpha_\star,\beta_\star}>0$, $\epsilon_n\tilde\sigma_{q,\alpha_n,\beta_n}\to 0$, as well as $\sqrt{n}\epsilon_n\big\{\widehat{\mathrm{MMD}}^2_q(P_{\alpha_n},P)-\widehat{\mathrm{MMD}}^2_q(Q_{\beta_n},P)\big\}\to 0$. The latter arguments yield (i) of Theorem 3 and the result follows.

Remark 3 *In Lemma 4, it is possible that $\sigma^2_{q,\alpha_\star,\beta_\star}=0$ when the random variables U and V are "perfectly dependent" and $P_{\alpha_\star}=Q_{\beta_\star}$, i.e., when $F(U_i;\alpha_\star)=G(V_i;\beta_\star)$ a.s. for every* *i. To be specific, by simple calculations, we always have*

$$\begin{aligned}\sigma^2_{q,\alpha_\star,\beta_\star}/8 := \mathrm{Var}\Big(&\mathbb{E}[k\big(X, G(V_2;\beta_\star)\big)|V_2] - \mathbb{E}[k\big(X, F(U_2;\alpha_\star)\big)|U_2]\\ &+ \mathbb{E}[k\big(X_2, G(V;\beta_\star)\big)|X_2] - \mathbb{E}[k\big(X_2, F(U;\alpha_\star)\big)|X_2]\\ &+ \mathbb{E}[k\big(F(U_1;\alpha_\star), F(U;\alpha_\star)\big)|U_1] - \mathbb{E}[k\big(G(V_1;\beta_\star), G(V;\beta_\star)\big)|V_1]\Big).\end{aligned}$$

When $F(U_i;\alpha_\star) = G(V_i;\beta_\star)$ a.s., which is for example the case when both generators of the optimal models are identical and U_i and V_i are generated by the same source of randomness, then it becomes clear that $\sigma^2_{q,\alpha_\star,\beta_\star} = 0$. However, this undesired phenomenon is avoided when U_i and V_i are chosen to be independent, i.e., when the two competing models are generated by independent sources of randomness, which is why we have imposed this assumption.

A.5 Technical Assumptions and Proof of Theorem 4 (Model Comparison)

Define $\tilde{h}(x,y,z;\alpha,\beta) := \tilde{h}^{(1)}(x,y;\alpha) - \tilde{h}^{(2)}(x,z;\beta)$, where $\tilde{h}^{(1)}(x,y;\alpha)$ and $\tilde{h}^{(2)}(x,z;\beta)$ are defined according to (9). Moreover, define

$$\begin{aligned}\tilde{g}(x,y,z;\alpha,\beta) :=& \frac{8}{3}\mathbb{E}\big[h\left((X,F(U;\alpha),G(V;\beta)),(x,y,z)\right)\tilde{h}\left(X,F(U,\alpha),G(V;\beta);\alpha,\beta\right)\big]\\ &+ \frac{4}{3}\tilde{h}^2(x,y,z;\alpha,\beta).\end{aligned}$$

Define some classes of functions as in Section A.3, but now indexed by (α,β) instead of α only, as

$$\mathcal{F}^{(\mathcal{M})} := \Big\{[\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{1:3} \mapsto h\big([\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{1:3};\alpha,\beta\big)\Big|\ \alpha \in B_\delta(\alpha_\star), \beta \in B_\delta(\beta_\star)\Big\},$$

$$\mathcal{G}^{(\mathcal{M})} := \Big\{[\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{1:3} \mapsto g\big([\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}]_{1:3};\alpha,\beta\big)\Big|\ \alpha \in B_\delta(\alpha_\star), \beta \in B_\delta(\beta_\star)\Big\},$$

and, in the same spirit, the analogs $\mathcal{F}_q^{(\mathcal{M})}$, $\mathcal{Q}^{(\mathcal{M})}$, $\mathcal{F}_c^{(\mathcal{M})}$, $\mathcal{G}_c^{(\mathcal{M})}$, $\tilde{\mathcal{F}}^{(\mathcal{M})}$ and $\tilde{\mathcal{G}}^{(\mathcal{M})}$ of $\mathcal{F}_q$, $\mathcal{Q}$, $\mathcal{F}_c$, $\mathcal{G}_c$, $\tilde{\mathcal{F}}$ and $\tilde{\mathcal{G}}$. We need to adapt the regularity Assumptions 7 and 8 to the new framework.

Assumption 10 *The maps $(\alpha,\beta) \mapsto \tilde{h}(x,y,z;\alpha,\beta)$ and $(\alpha,\beta) \mapsto \tilde{g}(x,y,z;\alpha,\beta)$ are twice continuously differentiable in a neighborhood of $(\alpha_\star,\beta_\star)$ at every $(x,y,z) \in \mathcal{S}\times\mathcal{S}\times\mathcal{S}$. Additionally, the maps*

$$(\alpha,\beta) \mapsto \mathbb{E}\Big[\nabla_{(\alpha,\beta)}\tilde{h}\big(X,F(U;\alpha),G(V;\beta);\alpha_\star,\beta_\star\big)\Big],\ \textit{and}$$

$$(\alpha,\beta) \mapsto \mathbb{E}\Big[\nabla_{(\alpha,\beta)}\tilde{g}\big(X,F(U;\alpha),G(V;\beta);\alpha_\star,\beta_\star\big)\Big]$$

are differentiable in a neighborhood of $(\alpha_\star,\beta_\star)$. We assume

$$\mathbb{E}\left[\nabla_{(\alpha,\beta)}\tilde{h}\big(X,F(U;\alpha_\star),G(V,\beta_\star);\alpha_\star,\beta_\star\big)\right] = \nabla_{(\alpha,\beta)}\mathbb{E}\left[\tilde{h}\big(X,F(U;\alpha_\star),G(V;\beta_\star);\alpha_\star,\beta_\star\big)\right] \textit{ and}$$

$$\mathbb{E}\left[\nabla_{(\alpha,\beta)}\tilde{g}\big(X,F(U;\alpha_\star),G(V;\beta_\star);\alpha_\star,\beta_\star\big)\right] = \nabla_{(\alpha,\beta)}\mathbb{E}\left[\tilde{g}\big(X,F(U;\alpha_\star),G(V;\beta_\star);\alpha_\star,\beta_\star\big)\right].$$

Moreover,

$$\|h(\cdot;\alpha_n,\beta_n) - h(\cdot;\alpha_\star,\beta_\star)\|_{L_2(\otimes_{i=1}^2 P\otimes P_U\otimes P_V)}\ ,\ \|q(\cdot;\alpha_n,\beta_n) - q(\cdot;\alpha_\star,\beta_\star)\|_{L_2(\otimes_{i=1}^4 P\otimes P_U\otimes P_V)}\ ,$$

$$\|g(\cdot;\alpha_n,\beta_n) - g(\cdot;\alpha_\star,\beta_\star)\|_{L_2(\otimes_{i=1}^3 P\otimes P_U\otimes P_V)}\ ,\text{and}\ \|\xi(\cdot;\alpha_n,\beta_n) - \xi(\cdot;\alpha_\star,\beta_\star)\|_{L_2(\otimes_{i=1}^6 P\otimes P_U\otimes P_V)}$$

tend to zero in probability for $n\to\infty$. Finally, for some real constant $\delta > 0$, we have

$$\mathbb{E}\Big[\sup_{\substack{(\alpha_1,\beta_1)\in B_\delta(\alpha_\star)\times B_\delta(\beta_\star),\\ (\alpha_2,\beta_2)\in B_\delta(\alpha_\star)\times B_\delta(\beta_\star)}} \|\nabla^2_{(\alpha,\beta),(\alpha,\beta)^\top}\tilde{h}\big(X,F(U;\alpha_2),G(V;\beta_2);\alpha,\beta\big)|_{(\alpha,\beta)=(\alpha_1,\beta_1)}\|_2^2\Big]$$
$$+\,\mathbb{E}\Big[\sup_{\substack{(\alpha_1,\beta_1)\in B_\delta(\alpha_\star)\times B_\delta(\beta_\star),\\ (\alpha_2,\beta_2)\in B_\delta(\alpha_\star)\times B_\delta(\beta_\star)}} \|\nabla^2_{(\alpha,\beta),(\alpha,\beta)^\top}\tilde{g}\big(X,F(U;\alpha_2),G(V;\beta_2);\alpha,\beta\big)|_{(\alpha,\beta)=(\alpha_1,\beta_1)}\|_2^2\Big] < \infty.$$

Assumption 11 *There exists some $\delta > 0$ such that the centered empirical U-processes $\big(\sqrt{n}(\mathcal{U}_n - \mathbb{P})f\big)_{f\in\mathcal{F}(\mathcal{M})}$, $\big(\sqrt{n}(\mathcal{U}_n - \mathbb{P})f\big)_{f\in\mathcal{F}_q(\mathcal{M})}$, $\big(n(\mathcal{U}_n - \mathbb{P})f\big)_{f\in\mathcal{F}_c(\mathcal{M})}$, $\big(\sqrt{n}(\mathcal{U}_n^{(2)} - \mathbb{P})f\big)_{f\in\mathcal{Q}(\mathcal{M})}$, $\big(\sqrt{n}(\mathcal{U}_n - \mathbb{P})g\big)_{g\in\mathcal{G}(\mathcal{M})}$ and $\big(n(\mathcal{U}_n - \mathbb{P})g\big)_{g\in\mathcal{G}_c(\mathcal{M})}$ as well as the empirical processes $\big(\sqrt{n}(\mathbb{P}_n - \mathbb{P})f\big)_{f\in\tilde{\mathcal{F}}(\mathcal{M})}$ and $\big(\sqrt{n}(\mathbb{P}_n - \mathbb{P})g\big)_{g\in\tilde{\mathcal{G}}(\mathcal{M})}$ weakly converge. Finally, the map $(\alpha,\beta)\mapsto\sigma_{\alpha,\beta}$ is continuously differentiable in a neighborhood of $(\alpha_\star,\beta_\star)$ and*

$$\mathbb{E}_{X,U,V}\left[\|\nabla_{(\alpha,\beta)}\tilde{h}\big(X,F(U;\alpha_n),G(V;\beta_n);\alpha_\star,\beta_\star\big) - \nabla_{(\alpha,\beta)}\tilde{h}\big(X,F(U;\alpha_\star),G(V;\beta_\star);\alpha_\star,\beta_\star\big)\|_2^2\right]$$

as well as

$$\mathbb{E}_{X,U,V}\left[\|\nabla_{(\alpha,\beta)}\tilde{g}\big(X,F(U;\alpha_n),G(V;\beta_n);\alpha_\star,\beta_\star\big) - \nabla_{(\alpha,\beta)}\tilde{g}\big(X,F(U;\alpha_\star),G(V;\beta_\star);\alpha_\star,\beta_\star\big)\|_2^2\right]$$

converge to 0 in probability.

Again, sufficient conditions to ensure the functional convergence of the considered empirical U-processes can be found in Appendix B.

Lemma 5 *Assume the models $\mathcal{M}_1$ and $\mathcal{M}_2$ satisfy Assumptions 1-2 and 9-11. Then the conclusions of Lemma 4 apply.*

Essentially, the proof of Lemma 5 goes along the same lines as the proof of Lemma 3, replacing each quantity defined for model specification with the corresponding quantity for model selection. Therefore, it has been omitted. Then, assuming Lemma 5 is satisfied, the proof of Theorem 4 follows identically to the proof of Theorem 3.

Appendix B. Sufficient Condition for Functional Convergence of Centered Empirical U-processes

In Section 3.4 and Section 4.3, we require that some centered empirical U-processes indexed by certain classes of functions converge to their appropriate limits in the functional sense (26). In this section, we provide some sufficient conditions that ensure such functional weak convergences.

Consider a generic class $\mathcal{L}$ of symmetric real-valued measurable functions on some product probability space $\big(\mathcal{Z}^q, \otimes_{i=1}^q P_Z\big)$. A class $\mathcal{L}$ is degenerate of order $r-1$, $r \geq 1$, if

$$\mathbb{E}_{(Z_i)_{1\le i\le q-r+1}\sim\otimes_{i=1}^{q-r+1}P_Z}\big[\ell(Z_1,\dots,Z_q)\big] = \text{const} \quad \otimes_{q-r+2}^{q} P_Z\text{-a.s. , and}$$

$$\text{Var}\Big(\mathbb{E}_{(Z_i)_{1\le i\le q-r}\sim\otimes_{i=1}^{q-r}P_Z}\big[\ell(Z_1,\dots,Z_q)\big]\Big) > 0$$

for every $\ell \in \mathcal{L}$. When $r = 1$, this simply means $\mathbb{E}\left[\ell(Z_1,\dots,Z_q)\right] = 0$ and $\mathbb{E}[\ell(Z_1,\dots,Z_q) \mid Z_1]$ is not a constant; thus, the class $\mathcal{L}$ is also called non-degenerate in this case.

Let us recall the usual definition of covering numbers: for a given norm $\|\cdot\|$ on $\mathcal{L}$, define the ϵ-covering number of $(\mathcal{L}, \|\cdot\|)$ as

$$N(\epsilon, \mathcal{L}, \|\cdot\|) := \min\Big\{n : \exists f_1,\dots,f_n \in \mathcal{L} \text{ s.t. } \sup_{f\in\mathcal{L}}\min_{i\le n}\|f - f_i\| \le \epsilon\Big\}.$$

Essentially, the covering number measures the size of $\mathcal{L}$ w.r.t. $\|\cdot\|$ and can be interpreted as a measure of complexity of $\mathcal{L}$ w.r.t. $\|\cdot\|$. Hereafter, we define some regularity condition on a generic class of functions $\mathcal{L}$ that is based on the covering numbers of $\mathcal{L}$. This condition will ensure the weak convergence of the centered empirical U-process $n^{r/2}\big((\mathcal{U}_n - \mathbb{P})\ell\big)_{\ell\in\mathcal{L}}$ and it is mainly based on (simplified) conditions provided by Arcones and Giné (1993).

Definition 1 *Let $\mathcal{L}$ denote a class of symmetric measurable functions on some product probability space $\big(\mathcal{Z}^q, \otimes_{i=1}^q P_Z\big)$ and let r be a positive integer. Then, $\mathcal{L}$ is called r-regular if the following is satisfied:*

1. $S(z_1,\dots,z_q) := \sup_{\ell\in\mathcal{L}}|\ell(z_1,\dots,z_q)| < \infty$ *for all* $(z_1,\dots,z_q) \in \mathcal{Z}^q$*;*
2. $\mathbb{E}_{(Z_i)_{1\le i\le q}\sim\otimes_{i=1}^q P_Z}\left[S(Z_1,\dots,Z_q)^2\right] < \infty$*;*
3. $\lim_{t\to\infty} tP_Z\Big(\mathbb{E}_{(Z_i)_{1\le i\le q-1}\sim\otimes_{i=1}^{q-1}P_Z}\big[S(Z_1,\dots,Z_{q-1},Z)^2\big] > t\Big) \to 0$*;*
4. *$\mathcal{L}$ is image admissible Suslin;*
5. *for all probability measures Q such that* $\mathbb{E}_{(Z_i)_{1\le i\le q}\sim Q}\left[S(Z_1,\dots,Z_q)^2\right] < \infty$*, we have*

$$\int_0^\infty \left\{\sup_Q \log N\Big(\epsilon\sqrt{\mathbb{E}_{(Z_i)_{1\le i\le q}\sim Q}\big[S(Z_1,\dots,Z_q)^2\big]}, \mathcal{L}, \|\cdot\|_{L_2(Q)}\Big)\right\}^{r/2} \text{d}\epsilon < \infty, \tag{33}$$

$$\textit{and} \quad \int_0^\infty \left\{\log N\Big(\epsilon, \mathcal{L}, \|\cdot\|_{L_2(\otimes_{i=1}^q P_Z)}\Big)\right\}^{r/2} \text{d}\epsilon < \infty.$$

Any centered empirical U-process $n^{r/2}\big((\mathcal{U}_n - \mathbb{P})\ell\big)_{\ell\in\mathcal{L}}$ that is degenerate of degree $r-1$ and r-regular weakly converges to a limit process in $L_\infty(\mathcal{L})$, which is asymptotically uniformly equicontinuous w.r.t. the norm $\|\cdot\|_{L_2(\otimes_{i=1}^q P_Z)}$: see Arcones and Giné (1993, p. 1535).

In Definition 1, the concept of being image admissible Suslin is a measurability property that may not be familiar to many readers. This is why we briefly discuss some sufficient conditions to ensure the image admissible Suslin property of a considered class of functions $\mathcal{L}$. Assume that $\mathcal{L}$ is parameterized by a vector-valued parameter α (resp. β) which belongs to a neighborhood of $\alpha_\star$ (resp. $\beta_\star$). In other words, we assume any class of functions $\mathcal{L}$ may be rewritten as $\mathcal{L} = \{\ell_\theta : \mathcal{E} \mapsto \mathbb{R} \,|\, \theta \in \Theta_0\}$, for some non-empty compact subset Θ_0 of a finite dimensional Euclidean space and $\mathcal{E}$ is the Cartesian product of $\mathcal{S}, \mathcal{U}$ (resp. $\mathcal{V}$) spaces. From Dudley (1984), p. 101, this implies that $\mathcal{L}$ is image admissible Suslin if

- $\mathcal{E}$ endowed with its Borel $\sigma-$algebra is a Polish space (i.e., is metrizable to become a separable and complete metric space), and
- the map $(x, \theta) \mapsto f_\theta(x)$ from $\mathcal{E} \times \Theta_0$ to $\mathbb{R}$ is jointly measurable.

These conditions are often met in practice and ensure the image admissible Suslin property of $\mathcal{L}$. Moreover, for certain classes of functions, it is well known that (33) is satisfied, such as for VC-subgraph classes of functions or classes of functions that satisfy a Lipschitz-property, see (van der Vaart and Wellner, 1996, Section 2.6 and 2.7). In particular, these conditions may be directly checked provided the kernel k, $\big(F(\cdot;\alpha)\big)_{\alpha\in\Theta_1}$ and $\big(G(\cdot;\beta)\big)_{\beta\in\Theta_2}$.

We are ready to provide sufficient conditions for the weak convergence statements in Assumptions 6, 8 and 11 to hold. A sufficient condition to ensure that there exists some $\delta > 0$ such that the empirical U-processes $\big(\sqrt{n}(\mathcal{U}_n - \mathbb{P})f\big)_{f\in\mathcal{F}}$, $\big(\sqrt{n}(\mathcal{U}_n^{(2)} - \mathbb{P})q\big)_{q\in\mathcal{F}_q}$, $\big(\sqrt{n}(\mathcal{U}_n - \mathbb{P})g\big)_{g\in\mathcal{G}}$ and $\big(\sqrt{n}(\mathcal{U}_n^{(2)} - \mathbb{P})\xi\big)_{f\in\mathcal{Q}}$ weakly converge to their appropriate limits in the functional sense as claimed in Assumption 6 is as follows: *there exists some $\delta > 0$ such that the classes of functions $\mathcal{F}$, $\mathcal{F}_q$, $\mathcal{G}$ and $\mathcal{Q}$ are 1-regular.*

Moreover, a sufficient condition to ensure that there exists $\delta > 0$ such that the centered empirical U-processes $\big(n(\mathcal{U}_n - \mathbb{P})f\big)_{f\in\mathcal{F}_c}$ and $\big(n(\mathcal{U}_n - \mathbb{P})g\big)_{f\in\mathcal{G}_c}$ and the empirical processes $\big(\sqrt{n}(\mathbb{P}_n - \mathbb{P})f\big)_{f\in\tilde{\mathcal{F}}}$ and $\big(\sqrt{n}(\mathbb{P}_n - \mathbb{P})g\big)_{g\in\tilde{\mathcal{G}}}$ converge to their appropriate functional limits as claimed in Assumption 8 is: *there exists some $\delta > 0$ such that the classes of functions $\tilde{\mathcal{F}}$ and $\tilde{\mathcal{G}}$ are 1-regular and that the classes of functions $\mathcal{F}_c$ and $\mathcal{G}_c$ are 2-regular.*

A sufficient condition to ensure that there exists some $\delta > 0$ such that the centered empirical U-processes $\big(\sqrt{n}(\mathcal{U}_n - \mathbb{P})f\big)_{f\in\mathcal{F}^{(\mathcal{M})}}$, $\big(\sqrt{n}(\mathcal{U}_n - \mathbb{P})f\big)_{f\in\mathcal{F}_q^{(\mathcal{M})}}$, $\big(n(\mathcal{U}_n - \mathbb{P})f\big)_{f\in\mathcal{F}_c^{(\mathcal{M})}}$, $\big(\sqrt{n}(\mathcal{U}_n^{(2)} - \mathbb{P})f\big)_{f\in\mathcal{Q}^{(\mathcal{M})}}$, $\big(\sqrt{n}(\mathcal{U}_n - \mathbb{P})g\big)_{g\in\mathcal{G}^{(\mathcal{M})}}$ and $\big(n(\mathcal{U}_n - \mathbb{P})g\big)_{g\in\mathcal{G}_c^{(\mathcal{M})}}$ and the empirical processes $\big(\sqrt{n}(\mathbb{P}_n - \mathbb{P})f\big)_{f\in\tilde{\mathcal{F}}^{(\mathcal{M})}}$ and $\big(\sqrt{n}(\mathbb{P}_n - \mathbb{P})g\big)_{g\in\tilde{\mathcal{G}}^{(\mathcal{M})}}$ weakly converge to their appropriate functional limits as claimed in Assumption 11 is: *there exists some $\delta > 0$ such that the classes of functions $\mathcal{F}^{(\mathcal{M})}$, $\mathcal{F}_q^{(\mathcal{M})}$, $\tilde{\mathcal{F}}^{(\mathcal{M})}$, $\mathcal{Q}^{(\mathcal{M})}$, $\mathcal{G}^{(\mathcal{M})}$ and $\tilde{\mathcal{G}}^{(\mathcal{M})}$ are 1-regular and that the class of functions $\mathcal{F}_c^{(\mathcal{M})}$ and $\mathcal{G}_c^{(\mathcal{M})}$ is 2-regular.*

Note that the 1-regularity of the classes of functions $\tilde{\mathcal{F}}$, $\tilde{\mathcal{G}}$, $\tilde{\mathcal{F}}^{(\mathcal{M})}$ and $\tilde{\mathcal{G}}^{(\mathcal{M})}$ is simply a sufficient condition to ensure that they are Donsker classes (van der Vaart and Wellner, 1996, Section 2.5). Finally, it should again be emphasized that the concept of r-regularity does not rely on any differentiability property of the considered functions. Therefore, it is a suitable concept to ensure the asymptotic convergence of centered empirical U-processes indexed by classes of non-differentiable functions, such as in Section 3.4 and Section 4.3. See Appendix C for an example.

Appendix C. A ReLu-Type Neural Network Model

The goal of this section is to provide a short example which illustrates that the conditions for an application of Theorem 2 are verifiable when the generating function $\alpha \mapsto F(\cdot, \alpha)$ is non-smooth. Similar calculations yield that the assumptions of Theorem 4 are also satisfied for this example, but due to space limitations we leave the detailed calculations to the reader. The generative model with which we conduct our analysis is a special type of a one-layer ReLu neural network given by

$$F(u;\alpha) : u \mapsto \sum_{k=1}^{m} a_k \max(u - b_k, 0) + c, \tag{34}$$

a map that is parametrized by $\alpha := (a_1, \dots, a_m, b_2, \dots, b_m, c)$, since we impose $b_1 = 0$. We try to model P in terms of a generative model $(P_\alpha)_{\alpha \in \Theta_1}$ where the law of any P_α is the law of $F(U;\alpha)$, $U \sim \text{Unif}([0,1])$, and Θ_1 denotes some subset of $\mathbb{R}^{2m}$ with non-empty interior. This model is clearly not differentiable w.r.t. (some of) the parameters b_k. Thus, one cannot rely on the results of Theorem 1. To make the exposition easier we assume that $\alpha_\star = \text{argmin}_\alpha \text{MMD}(P_\alpha, P)$ is unique. This latter property is satisfied when all parameters a_k are positive and $b_1 = 0 < b_1 < \ldots < b_m < 1$ (Lemma 6 below), an identifiability condition that is assumed from now on. Further, we assume our kernel is bounded by one and is twice continuously differentiable with bounded first and second derivatives, which is, e.g., satisfied by the Gaussian kernel. This implies that it is globally Lipschitz continuous and we denote the corresponding Lipschitz constant as L_k.

We will frequently use Theorem 2.10.20 in van der Vaart and Wellner (1996). It implies that sums of Donsker classes are again Donsker. Moreover, products of uniformly bounded Donsker classes are Donsker. Let $\mathcal{L}$ denote a generic class of uniformly bounded functions with an upper bound C. We will show that each $\mathcal{L} \in \{\mathcal{F}, \mathcal{F}_q, \mathcal{Q}, \mathcal{G}, \mathcal{F}_c, \mathcal{G}_c, \tilde{\mathcal{F}}, \tilde{\mathcal{G}}\}$ has polynomially bounded covering number w.r.t. $\|\cdot\|_{L_2(Q)}$, where Q denotes an arbitrary discrete probability measure. This then implies that the entropy integral in (33) is finite, since (van der Vaart and Wellner, 1996, footnote 1 p. 84) implies that the entropy integral in (33) only has to be considered for discrete probability measures. From this it immediately follows that $\mathcal{L}$ is r-regular for $r \in \{1, 2\}$.

First, recall that, for every norm $\|\cdot\|$ on $\mathcal{L}$, the ϵ covering number $N(\epsilon, \mathcal{L}, \|\cdot\|)$ is bounded by the corresponding 2ϵ bracketing number, denoted $N_{[]}(2\epsilon, \mathcal{L}, \|\cdot\|)$ (van der Vaart and Wellner, 1996, p. 84). Further, every ReLu neural network with a bounded parameter set and bounded inputs is globally Lipschitz w.r.t. its parameter α, denoting the corresponding Lipschitz constant as L. Then, we have

$$|k\big(F(U_1;\alpha_1), F(U_2;\alpha_2)\big) - k\big(F(U_1;\alpha_3), F(U_2;\alpha_4)\big)| \le L_k L \|(\alpha_1, \alpha_2) - (\alpha_3, \alpha_4)\|_1, \text{ a.e.}$$

By van der Vaart and Wellner (1996, Theorem 2.7.11), the $2L_k L\epsilon$ bracketing number of the classes $\mathcal{K}_1 := \{k\big(F(\cdot;\alpha_1), F(\cdot;\alpha_2)\big) \mid \alpha_1, \alpha_2 \in \Theta_1\}$ and $\mathcal{K}_2 := \{k(\cdot, F(\cdot;\alpha_2)\big) \mid \alpha \in \Theta_1\}$ w.r.t. an arbitrary norm $\|\cdot\|$ are bounded by the $C\epsilon$ covering number of $\Theta_1 \times \Theta_1$ and Θ_1 w.r.t. $\|\cdot\|_1$, respectively, where C is a suitable constant. These covering numbers are polynomial in ϵ as $\Theta_1 \times \Theta_1$, resp. Θ_1, is contained in a ball of radius $R > 0$ and such covering numbers are bounded by an expression of the form $C_1 \epsilon^{-2(2m+1)}$, $C_1 > 0$ (Vershynin, 2018, Corollary 4.2.13). Now, consider a class $\mathcal{L} \in \{\mathcal{F}, \mathcal{F}_q, \mathcal{Q}, \mathcal{G}, \mathcal{F}_c, \mathcal{G}_c\}$. Any element in $\mathcal{H}$ is written as sums and/or products of uniformly bounded functions from $\mathcal{K}_1$ and $\mathcal{K}_2$. Thus, the proof of van der Vaart and Wellner (1996, Theorem 2.10.20) - in their notation, $\alpha_i = 1$ - implies the existence of $C_2 > 0$ s.t.

$$\sup_Q N(\epsilon, \mathcal{L}, L_2(Q)) \leq \sup_Q N(C_2\epsilon, \mathcal{K}_1, \|\cdot\|_{L_2(Q)})^{j_1} N(C_2\epsilon, \mathcal{K}_2, \|\cdot\|_{L_2(Q)})^{j_2} \quad i = 1, 2,$$

where the powers $j_1, j_2 \in \mathbb{N}$ depend on the respective function class $\mathcal{L}$, and Q ranges over all discrete probability measures. Due to the derivations above,

$$N\big(C_2\epsilon, \mathcal{K}_i, L_2(Q)\big)^{j_i} \leq N_{[]}(2C_2\epsilon, \mathcal{K}_i, L_2(Q))^{j_i} \leq C_3\epsilon^{-2j_i(2m+1)}$$

for a suitable $C_3 > 0$. Thus, $\sup_Q N(\epsilon, \mathcal{L}, L_2(Q))$ is bounded by an expression of the form $C_4\epsilon^{-2(j_1+j_2)(2m+1)}$, $C_4 > 0$, implying that the entropy integrals in (33) are finite for every positive integer r. Thus, $\{\mathcal{F}, \mathcal{F}_q, \mathcal{Q}, \mathcal{G}, \mathcal{F}_c, \mathcal{G}_c\}$ are r-regular for every $r \in \mathbb{N}$ and for every ReLu neural network with bounded inputs and a bounded parameter set.

Let us continue by verifying the existence of moments for the derivatives of $\tilde{h}(x, \alpha)$ and $\tilde{g}(x, y; \alpha)$ w.r.t. α (recall (9) and (27)). It is easy to see that the differentiability of $\alpha \mapsto \tilde{h}(x, y; \alpha)$ and $\alpha \mapsto \tilde{g}(x, y; \alpha)$ follows from the differentiability of terms of the form $\mathbb{E}\big[k\big(x, F(U; \alpha)\big)\big]$, since they are sums, squares and products of these terms. In our model (34) and setting $b_{m+1} = 1$, these terms can be written as

$$\mathbb{E}\big[k\big(x, F(U; \alpha)\big)\big] = \sum_{k=1}^{m} \int_{b_k}^{b_{k+1}} k\big(x, c + \sum_{j=1}^{k} a_j(u - b_j)\big) \mathrm{d}u.$$

Thus, $\mathbb{E}\left[k(x, F(U; \alpha)\right]$ is twice continuously differentiable w.r.t. α in a neighborhood of $\alpha_\star$ for every fixed x, with bounded derivatives. This implies the existence of $\nabla_{\alpha_\star}\tilde{h}(x, y; \alpha_\star)$ and $\nabla^2_{\alpha_\star}\tilde{h}(x, y; \alpha_\star)$ and its respective moment conditions. Similar arguments show that the same is true for $\nabla_{\alpha_\star}\tilde{g}(x, y; \alpha_\star)$ and $\nabla^2_{\alpha_\star}\tilde{g}(x, y; \alpha_\star)$. Therefore, Assumption 7 is satisfied. Note that this also implies that $\alpha \mapsto \sigma^2_\alpha$ and $\alpha \mapsto \sigma^2_{q,\alpha}$ are differentiable in a neighborhood of $\alpha_\star$.

As $\nabla_{\alpha_\star}\tilde{h}(x, y; \alpha_\star)$ and $\nabla_{\alpha_\star}\tilde{g}(x, y; \alpha_\star)$ are continuously differentiable in a neighborhood of $\alpha_\star$ and k has bounded derivatives, we also have that $\tilde{\mathcal{F}}$ and $\tilde{\mathcal{G}}$ are classes of Lipschitz continuous bounded functions indexed by $\alpha \in B_\delta(\alpha_\star)$, choosing a $\delta > 0$ small enough. Therefore, mimicking the arguments above we can show that $\{\tilde{\mathcal{F}}, \tilde{\mathcal{G}}\}$ have polynomially bounded covering numbers and are r-regular for every $r \in \mathbb{N}$. Finally, all conditions mentioned in Assumption 6-8 of the form $\|l(x, \alpha_n) - l(x, \alpha_\star)\|_{L_2} \to 0$ are satisfied for all $l \in \mathcal{L}$. Indeed, $l(x, \alpha_n) - l(x, \alpha_\star) \to 0$ in probability for every consistent estimator α_n of $\alpha_\star$, because all considered functions are bounded and Lipschitz w.r.t. α.

It only remains to spell out conditions under which an estimator α_n for $\alpha_\star$ is asymptotically normal. To this purpose, define

$$\alpha_n = \mathrm{argmin}_{\alpha \in \Theta_1} n^{-1} \sum_{i=1}^{n/2} h\Big(\big(X_{2i}, F(U_{2i}; \alpha)\big), \big(X_{2i-1}, F(U_{2i-1}; \alpha)\big)\Big).$$

Note that α_n is an M-estimator of $\alpha_\star := \mathrm{argmin}_{\alpha \in \Theta_1} \mathrm{MMD}(P_\alpha, P)$. By the same reasoning as above, the map $\alpha \mapsto h\Big(\big(x_1, F(u_1; \alpha)\big), \big(x_2, F(u_2; \alpha)\big)\Big)$ is differentiable in a neighborhood of $\alpha_\star$ for $P \otimes P_U \otimes P \otimes P_U$ almost every (x_1, u_1, x_2, u_2), with Lipschitz continuous derivative w.r.t. α. Additionally, the corresponding Lipschitz constant is independent of (x_1, u_1, x_2, u_2). Since $\alpha \mapsto \text{MMD}(P_\alpha, P)$ is twice continuously differentiable in a neighborhood of $\alpha_\star$ we obtain that α_n is asymptotically normal (van der Vaart, 2000, Theorem 5.23).

Note that the most difficult task in this example is to ensure the uniqueness of $\alpha_\star$, which is necessary to apply standard results to obtain the asymptotic normality of α_n. All other statements above do not require the existence of a unique $\alpha_\star$ and are also valid when there are multiple argmins. We were able to show that, under the assumptions of Lemma 6 below, that our model (34) always has a unique argmin. Thus, under the identifiability conditions from Lemma 6 (a result that is of interest per se), Assumptions 1-4 and Assumptions 6-8 are satisfied and Theorem 2 applies.

Lemma 6 *If a_k, $k \in \{1, \dots, m\}$ are strictly positive and $b_1 = 0 < b_2 < \dots < b_{m-1} < b_m < 1$, then model (34) is identifiable.*

Proof Denote by $\theta := (a_1, \dots, a_m, b_2, \dots, b_m, c)$ the vector of unknown parameters. Let F_θ be the cdf of the random variable $\sum_{k=1}^m a_k \max(U - b_k, 0) + c$. For notational convenience, set $s_k := a_1 + \dots + a_k$ and $v_k := a_1 b_1 + a_2 b_2 + \dots + a_k b_k$, $k \in \{1, \dots, m\}$. Note that $s_k \geq v_k$ for any k. The support of our law is then $\mathcal{D}_\theta := (c, c + s_m - v_m)$. Set $b_{m+1} = 1$. For any t in $\mathcal{D}_\theta$, write

$$\begin{aligned} F_\theta(t) &= \sum_{k=1}^m \mathbb{P}\bigg(\sum_{j=1}^m a_j \max(U - b_j, 0) + c \leq t; b_k \leq U \leq b_{k+1} \bigg) \\ &= \sum_{k=1}^m \mathbb{P}\big(s_k U - v_k + c \leq t; b_k \leq U \leq b_{k+1} \big) \\ &= \sum_{k=1}^m \Big\{ \mathbf{1}(s_k b_k < t - c + v_k < s_k b_{k+1}) \Big(\frac{t - c + v_k}{s_k} - b_k \Big) + \mathbf{1}(t - c + v_k > s_k b_{k+1})(b_{k+1} - b_k) \Big\}. \end{aligned}$$

Any cdf F_θ is piecewise linear, with successive positive linear slopes $1/s_1$, $1/s_2$,...,$1/s_m$. Indeed, the interior of the intervals $I_k := [s_k b_k + c - v_k, s_k b_{k+1} + c - v_k)$ are never empty since $b_{k+1} > b_k$. Moreover, their intersections with $\mathcal{D}_\theta$ are never empty since $s_k b_k - v_k \geq 0$ and

$$s_k b_{k+1} - v_k \leq s_k - v_k \leq s_m - v_m.$$

Note that any I_k starts at $t_k = s_k b_k + c - v_k$, ends at $t'_k = s_k b_{k+1} + c - v_k$, and check that $t'_k = t_{k+1}$, $k \in \{1, \dots, m-1\}$. Since $t_1 = c$ and $t'_m = c + s_m - v_m$, the intervals $I_1, \dots, I_{m-1}$ are disjoint and yield a partition of $\mathcal{D}_\theta$.

Now consider two model parameters $\theta^{(1)}$ and $\theta^{(2)}$ s.t. $F_{\theta^{(1)}}(t) = F_{\theta^{(2)}}(t)$ for any real number t. In particular, assume their supports are the same. With obvious notations, this implies $c^{(1)} = c^{(2)} =: c$. Moreover, their sequences of slopes have to be the same, implying $s_k^{(1)} = s_k^{(2)}$ for every $k = 1, \dots, m$. This implies $a_k^{(1)} = a_k^{(2)}$ for every $k = 1, \dots, m$, now denoted a_k simply and similarly for their sums s_k. Considering the starting points of the upward sloping segments, we have to satisfy $s_k b_k^{(1)} - v_k^{(1)} = s_k b_k^{(2)} - v_k^{(2)}$, $k \in \{2, \dots, m\}$. In particular, $s_2 b_2^{(1)} - v_2^{(1)} = s_2 b_2^{(2)} - v_2^{(2)}$, or $a_1 b_2^{(1)} = a_1 b_2^{(2)}$ equivalently. Since $a_1 > 0$, we get $b_2^{(1)} = b_2^{(2)}$. Recursively, it can be proved that $b_k^{(1)} = b_k^{(2)}$ for any $k \in \{2, \dots, m\}$. Indeed, assume $b_j^{(1)} = b_j^{(2)}$

for $j \in \{1, \ldots, k-1\}$. Then $s_k b_k^{(1)} - v_k^{(1)} = s_k b_k^{(2)} - v_k^{(2)}$ implies $\left(s_j - a_j\right) b_k^{(1)} = \left(s_j - a_j\right) b_k^{(2)}$ and then $b_k^{(1)} = b_k^{(2)}$, proving the result. ■

References

Pierre Alquier and Mathieu Gerber. Universal robust regression via maximum mean discrepancy. *Biometrika*, 111:71–92, 2024.

Pierre Alquier, Badr-Eddine Chérief-Abdellatif, Alexis Derumigny, and Jean-David Fermanian. Estimation of copulas via maximum mean discrepancy. *Journal of the American Statistical Association*, 118:1997–2012, 2023.

Miguel A. Arcones and Evarist Giné. Limit theorems for U-processes. *The Annals of Probability*, 21:1494–1542, 1993.

Miguel A. Arcones and Evarist Giné. U-processes indexed by Vapnik-Červonenkis classes of functions with applications to asymptotics and bootstrap of U-statistics with estimated parameters. *Stochastic Processes and their Applications*, 52:17–38, 1994.

Sylvain Arlot, Alain Celisse, and Zaid Harchaoui. A kernel multiple change-point algorithm via model selection. *Journal of Machine Learning Research*, 20:1–56, 2019.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

Rudolf Beran and Pressley Warwick Millar. A stochastic minimum distance test for multivariate parametric models. *The Annals of Statistics*, 17:125–140, 1989.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer, 2011.

Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G Willcocks. Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:7327–7347, 2021.

Wacha Bounliphone, Eugene Belilovsky, Matthew B Blaschko, Ioannis Antonoglou, and Arthur Gretton. A test of relative similarity for model selection in generative models. *arXiv preprint arXiv:1511.04581*, 2015.

Francois-Xavier Briol, Alessandro Barp, Andrew B Duncan, and Mark Girolami. Statistical inference for generative models with maximum mean discrepancy. *arXiv preprint arXiv:1906.05944*, 2019.

Florian Brück. Generative neural networks for characteristic functions. *Journal of Computational and Graphical Statistics*, pages 1–19, 2025.

Florian Brück, Jean-David Fermanian, and Aleksey Min. A corrected Clarke test for model selection and beyond. *Journal of Econometrics*, 235:105–132, 2023.

Ricardo Cao and Gábor Lugosi. Goodness-of-fit tests based on the kernel density estimator. *Scandinavian Journal of Statistics*, 32:599–616, 2005.

Antoine Chatalic, Nicolas Schreuder, Lorenzo Rosasco, and Alessandro Rudi. Nyström kernel mean embeddings. In *Proceedings of the 39th International Conference on Machine Learning*, pages 3006–3024. PMLR, 2022.

Badr-Eddine Cherief-Abdellatif and Pierre Alquier. MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. In *Proceedings of the 2nd Symposium on Advances in Approximate Bayesian Inference*, pages 1–21, 2020.

Badr-Eddine Chérief-Abdellatif and Pierre Alquier. Finite sample properties of parametric MMD estimation: robustness to misspecification and dependence. *Bernoulli*, 28:181–213, 2022.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 2606–2615. PMLR, 2016.

Nicolae Dinculeanu. *Vector integration and stochastic integration in Banach spaces*. John Wiley & Sons, 2000.

Richard M. Dudley. A course on empirical processes. In *Ecole d'été de Probabilités de Saint-Flour XII-1982*, pages 1–142. Springer, 1984.

James Durbin. *Distribution theory for tests based on the sample distribution function*. SIAM, 1973a.

James Durbin. Weak convergence of the sample distribution function when parameters are estimated. *The Annals of Statistics*, 1:279–290, 1973b.

Gintare Karolina Dziugaite, Daniel M. Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, pages 258–267. AUAI Press, 2015.

Yanqin Fan. Goodness-of-fit tests for a multivariate distribution by the empirical characteristic function. *Journal of Multivariate Analysis*, 62:36–63, 1997.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Kernel dimension reduction in regression. *The Annals of Statistics*, 37:1871–1905, 2009.

Kenji Fukumizu, Le Song, and Arthur Gretton. Kernel Bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14:3753–3783, 2013.

Christian Genest, Kilani Ghoudi, and Louis-Paul Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82:543–552, 1995.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, volume 19. Curran Associates, Inc., 2006.

Arthur Gretton, Kenji Fukumizu, Zaid Harchaoui, and Bharath K. Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Marc Hallin, Gilles Mordant, and Johan Segers. Multivariate goodness-of-fit tests based on Wasserstein distance. *Electronic Journal of Statistics*, 15:1328–1371, 2021.

Marius Hofert, Avinash Prasad, and Mu Zhu. Quasi-random sampling for multivariate distributions via generative neural networks. *Journal of Computational and Graphical Statistics*, 30:647–670, 2021.

Stefanie Jegelka, Arthur Gretton, Bernhard Schölkopf, Bharath K Sriperumbudur, and Ulrike Von Luxburg. Generalized clustering via kernel embeddings. In *KI 2009: Advances in Artificial Intelligence*, pages 144–152, 2009.

Oscar Key, Arthur Gretton, François-Xavier Briol, and Tamara Fernandez. Composite goodness-of-fit tests with kernels. *Journal of Machine Learning Research*, 26:1–60, 2025.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabas Poczos. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1718–1727. PMLR, 2015.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *Proceedings of the 33rd International conference on Machine Learning*, pages 276–284. PMLR, 2016.

James R. Lloyd and Zoubin Ghahramani. Statistical model criticism using kernel two sample tests. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2208–2217. PMLR, 2017.

David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Iliya Tolstikhin. Towards a learning theory of cause-effect inference. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1452–1461. PMLR, 2015.

Alexander J. McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management*. Princeton University Press, 2005.

Ha Quang Minh. Some properties of Gaussian reproducing kernel Hilbert spaces and their implications for function approximation and learning theory. *Constructive Approximation*, 32:307–338, 2010.

Krikamol Muandet, Kenji Fukumizu, Bharath K. Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10:1–141, 2017.

Krikamol Muandet, Wittawat Jitkrittum, and Jonas Kübler. Kernel conditional moment test via maximum moment restriction. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, volume 124 of *Proceedings of Machine Learning Research*, pages 41–50. PMLR, 2020.

Yu Nishiyama and Kenji Fukumizu. Characteristic kernels and infinitely divisible distributions. *Journal of Machine Learning Research*, 17:1–28, 2016.

Mijung Park, Wittawat Jitkrittum, and Dino Sejdinovic. K2-ABC: Approximate Bayesian computation with kernel embeddings. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 398–407. PMLR, 2016.

David Pollard. The minimum distance method of testing. *Metrika*, 27:43–70, 1980.

Susanne M. Schennach and Daniel Wilhelm. A simple parametric model selection test. *Journal of the American Statistical Association*, 112:1663–1674, 2017.

Bernhard Schölkopf, Krikamol Muandet, Kenji Fukumizu, Stefan Harmeling, and Jonas Peters. Computing functions of random variables via reproducing kernel Hilbert space representations. *Statistics and Computing*, 25:755–766, 2015.

Dino Sejdinovic, Heiko Strathmann, Maria Lomeli Garcia, Christophe Andrieu, and Arthur Gretton. Kernel adaptive Metropolis-Hastings. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1665–1673. PMLR, 2014.

Robert J. Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 1980.

Shubhanshu Shekhar, Ilmun Kim, and Aaditya Ramdas. A permutation-free kernel two-sample test. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022.

Carl-Johann Simon-Gabriel and Bernhard Schölkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research*, 19:1–29, 2018.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *Proc. 18th International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th International Conference on Machine Learning*, pages 961–968, 2009.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011.

Danica J. Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representations*, 2017.

Zoltán Szabó and Bharath K Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18:1–29, 2018.

Aad W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 2000.

Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer, 1996.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018.

Quang H. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57:307–333, 1989.

Yunfei Yang, Zhen Li, and Yang Wang. On the capacity of deep generative networks for approximating distributions. *Neural Networks*, 145:144–154, 2022.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, pages 804–813. AUAI Press, 2011.

Zhaokun Zhou, Yuanhong Zhong, Xiaoming Liu, Qiang Li, and Shu Han. DC-MMD-GAN: A new maximum mean discrepancy generative adversarial network using divide and conquer. *Applied Sciences*, 10, 2020.