Advances in Non-parametric Hypothesis Testing with Kernels

Wenkai Xu

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Gatsby Computational Neuroscience Unit

University College London

2021-10-13

Declaration

I, Wenkai Xu, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Non-parametric statistical hypothesis testing procedures aim to distinguish the null hypothesis against the alternative with minimal assumptions on the model distributions. In recent years, the maximum mean discrepancy (MMD) has been developed as a measure to compare two distributions, which is applicable to two-sample problems and independence tests. With the aid of reproducing kernel Hilbert spaces (RKHS) that are rich-enough, MMD enjoys desirable statistical properties including characteristics, consistency, and maximal test power. Moreover, MMD receives empirical successes in complex tasks such as training and comparing generative models.

Stein's method also provides an elegant probabilistic tool to compare *unnormalised* distributions, which commonly appear in practical machine learning tasks. Combined with rich-enough RKHS, the kernel Stein discrepancy (KSD) has been developed as a proper discrepancy measure between distributions, which can be used to tackle one-sample problems (or goodness-of-fit tests).

The existing development of KSD applies to a limited choice of domains, such as Euclidean space or finite discrete sets, and requires complete data observations, while the current MMD constructions are limited by the choice of simple kernels where the power of the tests suffer, e.g. high-dimensional image data. The main focus of this thesis is on the further advancement of kernel-based statistics for hypothesis testings. Firstly, Stein operators are developed that are compatible with broader data domains to perform the corresponding goodness-of-fit tests. Goodness-of-fit tests for general unnormalised densities on Riemannian manifolds, which are of the non-Euclidean topology, have been developed. In addition, novel non-parametric goodness-of-fit tests for data with censoring are studied. Then the tests for data observations with left truncation are studied, e.g. times of entering the hospital always happen before death time in the hospital, and we say the death time is truncated by the entering time. We test the notion of independence beyond truncation by proposing a kernelised measure for *quasi-independence*. Finally, we study the deep kernel architectures to improve the two-sample testing performances.

To NaN, in the memory of Prime

Acknowledgements

First of all, I would like to express my deep gratitude to my supervisor, Professor Arthur Gretton for bringing me into the field of kernel methods and non-parametric hypothesis testing. Arthur has been a great supervisor pushing my limit and broadening my scope in statistics and machine learning; and helped me in the every crucial moment of the journey. What makes Arthur an even better supervisor, he inspired me not only academically and but also non-academically to lay a strong foundation for my long-term pursuit of research excellence. I feel really fortunate to be nurtured in this particular way, to entitle the Doctor of Philosophy Degree.

I would like to thank the Gatsby charitable foundation, the Gatsby Computational Neuroscience Unit, for providing the greatest research environment at all times, making the challenging Ph.D. journey fruitful and resilient. In particular, I would like to thank Professor Aapo Hyvärinen for his guidance and exemplification of research philosophy, which would have a life-long impact on my future. I would devote my particular appreciation to Dr. Wittawat Jitkrittum, who not only mentored my first research work; encouraged me on my research progress; but also most importantly, enlightened me on searching for my inner peace. Though the process takes a longer time, Dr. Jitkrittum's patience and continuous demonstration make this gradually happen. A special appreciation to Kirsty McNaught for her continuous support, help and demonstration during my journey, on the discouraged and flourish moments alike.

I would like to thank Professor Maneesh Sahani, Professor Peter Orbanz, Professor Danica Sutherland and Dr. Ricardo Pio Monti for their encouragement, help, and support, in one way or another, that made me leap over several obstacles before completing the degree. I would also like to thank my fantastic collaborators, Dr. Tamara Fernandez and Dr. Nicolas Rivera, who has been great mentors for research and friends for life. I also want to thank my office-mates, Dr. Kevin Wenliang Li, Heishiro Kanagawa, Dr. Feng Liu and Dr. Sanjeevan Ahilan for the intellectual stimulating discussions and various support during my Ph.D. Also thank Barry Fong who not only helped and encouraged myself during the course but also making the Gatsby unit a smooth and homely place.

Moreover, I would like to thank Professor Kenji Fukumizu for his continuous guidance and support during my Ph.D. journey. Fukumizu sensei has always been patient to listen to my ideas and give visionary hint when I get stuck. I thank Professor Masashi Sugiyama, and Dr. Takafumi Kajihara for hosting me in Japan which was an unforgettable experience during my Ph.D. I was fortunate to meet many great friends in RIKEN and I thank all their help in different aspect for my Ph.D.

I really thank Dr. Gang Niu, for his altruistic help and strong guidance in various aspects in my life, expanding my research scope and pointing out the directions to obtain interesting research problems, as well as making my first journal publication possible. I would also like to thank, in particular, Dr. Takeru Matsuda. During my time in Japan, Matsuda-san introduced to me the directional statistics, Riemannian manifold and consequently information geometry. These topics are not only interesting in terms of technical research in statistics, but also meaningful to me in terms of understanding a broader scope of subject matters, the underlying mechanisms between Mathematics, statistics, social sciences, and religious study. Building such theory is on its own of separate academic interest and deserves another thesis. I feel blessed to have the opportunities to work on these topics during my Ph.D. journey, leading me closer to the title of Doctor of Philosophy.

I devote my deep gratitude to Professor Gesine Reinert for her great advice on the topics of Stein's method and network analysis as well as her enlightenment on philosophy and attitude towards research life, which greatly strengthened my appreciation and believe on topics related to Stein's method and furthered my exploration on the wider spectrum of topics evolved.

I am also very grateful to have Professor Chris Oates, and Professor François-Xavier Briol to examine my thesis and provide invaluable suggestions for improvement and future enlightenment.

I thank all my parents and my sister for their full supports in every aspects of my journey. I would like to thank My Nan, who encouraged me to improve during this Ph.D. and inspired me to seek for my inner peace during my research endeavor. We are the genuine admirer of each other, the strongest critics for each other, the closest family and the best friend. She raised me up on my worst moment and made me more confident and calm. She nurtured me to appreciate the greatest opportunities and life I have encountered. While we did not have a chance to see each other during the period working on the thesis research and finish writing up due to pandemic travel constraints, she shed the light of a novel life approach upon completing this Ph.D. As such, the completion seeks not to an end but a new beginning and we share the same soul to perceive the world of *nirvana*.

Contents

Chapter 1

Introduction

We address the problem of non-parametric hypothesis testing, advancing the existing techniques to larger classes of distributions beyond complete observations[1] on $\mathbb{R}^d$. We consider testing distributions on complex non-Euclidean domains such as directional distributions $\mathbb{S}^{d-1}$ or Riemannian manifold $\mathcal{M}$; testing data with incomplete observations, in particular, censored data and truncated data. In this thesis, we focus on the development of non-parametric hypothesis testing techniques based on functions in reproducing kernel Hilbert spaces (RKHS).

1.1 Hypothesis Testing with Kernels

The task of hypothesis testing involves the distribution comparison with observed samples, embracing the notion of probability of the null hypothesis falling below a predetermined significance level which is also referred to as the test size. The testing procedure can be rephrased as follows: P and Q are two distributions and the null hypothesis reads $H_0 : P = Q$; comparisons between distributions are made to conclude whether P and Q have significant difference w.r.t. the test size, which is referred to as the alternative hypothesis $H_1 : P \neq Q$. Testing scenarios include two-sample testing, independence testing and goodness-of-fit testing. For two-sample testing, two sets of samples are drawn from *unknown* distributions P and Q and $H_0 : P = Q$ is tested, i.e. whether two sets of samples are drawn from the same distribution. Independence testing tests statistical dependence between two random variables X and Y, with joint distribution P_{xy} and marginal distributions P_x and P_y respectively. The null hypothesis reads $H_0 : P_{xy} = P_x P_y$, i.e. the joint distribution can be factorised into the product of marginals. Goodness-of-fit testing examines $H_0 : P = Q$ for *known* distribution P and a set of samples drawn from *unknown* distribution Q, i.e. testing whether a set of observed samples are drawn from a known model. The knowledge of P distinguishes goodness-of-fit testing from two-

[1] complete observation refers to no missing data, e.g. uncensored data

sample testing[2].

"Non-parametric hypothesis testing" refers to the scenario where the assumptions made on the distributions P and Q are minimal. In particular, the distributions in non-parametric testing are not assumed to be in any parametric family. By contrast, parametric tests, such as student t-test or normality test, assume a pre-defined parametric family to be tested against, and usually deal with particular summary statistics such as means or standard deviations, which are more restrictive in terms of comparing the full distributions. Recent advancement of non-parametric tests introduce RKHS functions which can be rich enough to distinguish distributions whenever they differ. The Maximum mean discrepancy (MMD) [Gretton et al., 2012a] has been developed to tackle the two sample problem via the notion of kernel mean embedding [Muandet et al., 2017], mapping a distribution to a function in an RKHS. If such a mapping is injective, the kernel associated with the RKHS is known as a *characteristic* kernel [Sriperumbudur et al., 2011], making MMD a proper discrepancy measure for distributions. Independence tests are studied via the Hilbert-Schmidt Independence Criterion (HSIC) [Gretton et al., 2008] which is derived using the kernel mean embedding for both the joint distribution and the product marginal distributions.

Stein's method [Barbour and Chen, 2005] provides an elegant probabilistic tool for comparing distributions, and was originally developed for approximating the normal distributions [Stein et al., 1972] and Poisson approximations [Barbour et al., 1992]. As Stein's method may only require to access the distributions through the differential (or difference) of the log density functions (or mass functions), it is applicable to deal with unnormalised models [Hyvärinen, 2005], which draw increasing attentions in the statistics and machine learning communities. With rich enough RKHS test functions, kernel Stein discrepancies (KSD) [Gorham and Mackey, 2015; Ley et al., 2017] are developed for goodness-of-fit testing on smooth unnormalised models on Euclidean space $\mathbb{R}^d$ [Chwialkowski et al., 2016; Liu et al., 2016]. Based on the kernelised Stein discrepancies, statistical tests are also studied for discrete distributions [Yang et al., 2018], point processes [Yang et al., 2019], latent variable models [Kanagawa et al., 2019], as well as conditional densities [Jitkrittum et al., 2020]. Additionally, computationally efficient kernel-based tests are also studied in the context of two-sample testing [Gretton et al., 2012b; Jitkrittum et al., 2016a], independence testing [Jitkrittum et al., 2017] and goodness-of-fit testing [Jitkrittum et al., 2017].

[2] As only one set of samples are observed, goodness-of-fit test is also referred to as the one-sample test or the one-sample problem.

Motivations and Contributions Non-parametric hypothesis tests for complex domains are not yet well developed. For instance, in directional statistics, statistical tests are mainly based on parametric tests such as the Rayleigh test or Kuiper test [Mardia and Jupp, 1999], and tests on general Riemannian manifolds are based on the Sobolev test of uniformity, which is limited to estimating complicated transformations of functions [Jupp et al., 2005, 2008]. In addition, density estimations on Riemannian manifolds generally suffer from intractable normalisation terms. While attempts have been made to estimate these normalisation terms [Mardia et al., 2016; Jupp and Kume, 2018], they usually suffer from low accuracy or high computational cost. The success of non-parametric kernel-based statistical tests motivates us to develop kernel-based non-parametric goodness-of-fit tests for non-Euclidean data, directly dealing with unnormalised models. Moreover, following the kernel-based statistical tests for censored data [Fernandez and Gretton, 2019; Fernandez et al., 2019; Fernandez and Rivera, 2019] achieving state-of-the-art performance, we are also inspired to further improve the kernel-based goodness-of-fit tests, and to extend the kernel-based tests for (in)dependence for data with both truncation and censoring.

1.2 Hypothesis Testing with Deep Neural Networks

In kernel-based hypothesis testing, a particular kernel may work well for a given observed set of finite samples while not being so useful w.r.t. another sample set, especially when the data can be complex and dimensions become higher. The initial proposal to tackle the kernel choice problem in two-sample testing [Gretton et al., 2012a] suggests using the median Euclidean distance between data as bandwidth for Gaussian kernel, which yields state-of-the-art performances on relatively simple data. However, statistical testing with more complex data adaptive kernels [Gretton et al., 2012b; Jitkrittum et al., 2016a, 2017] has been proposed to better distinguish the distributions from the particular finite sample observations.

Deep neural networks have been increasingly studied in recent years with successful performance on modern machine learning tasks involving high dimensional datasets. The expanding capacity and rich representations of modern deep neural network architectures facilitate the extraction of useful features and modelling complex functions. Combining the elegant functional and statistical properties of RKHS functions with the flexibility of deep neural networks, the *deep kernel* approach [Wilson et al., 2016] has been devised to tackle various problems including density estimation [Wenliang et al., 2018], semi-supervised learning [Jean et al.,

2018] and training generative models [Li et al., 2017; Sutherland et al., 2016].

Our goal is to develop novel approaches for learning deep kernel representations from observed datasets which can further improve the test power, even with limited sample size. We have further investigated the features extracted from the deep neural networks for interpretability purposes.

1.3 Structure of the Thesis

We start with a review, in Chapter 2, of existing kernel-based hypothesis testing techniques, which address two-sample testing, independence testing and goodness-of-fit testing. In Chapter 3, we develop a novel goodness-of-fit test based on appropriate Stein operators for non-Euclidean data, including directional data; and we provide a more general discussion on smooth Riemannian manifolds. In Chapter 4, we study Stein operators that can deal with censoring, and derive the corresponding goodness-of-fit testing procedures. In Chapter 5, we consider data with both truncation[3] and censoring. We define the notion of quasi-independence and develop the kernel-based testing procedure for quasi-independence. In Chapter 6, we develop and discuss the deep kernel learning for hypothesis testing.

The four main thesis chapters are based on the following publications which appeared over the course of this Ph.D.[4]

1. Chapter 3 Goodness-of-fit Tests on non-Euclidean Data

   Xu, W. & Matsuda, T. (2020) A Stein Goodness-of-fit Test for Directional Distributions. International Conference on Artificial Intelligence and Statistics (AISTATS).

   Xu, W. & Matsuda, T. (2021) Interpretable Stein Goodness-of-fit Tests on Riemannian Manifolds. International Conference on Machine Learning (ICML).

2. Chapter 4 Goodness-of-fit Tests for Censored Data

   Fernandez, T.*, Rivera, N.*, **Xu, W.*** & Gretton, A. (2020) Kernelized Stein Discrepancy Tests of Goodness-of-fit for Time-to-Event Data. International Conference on Machine Learning (ICML).

3. Chapter 5 A Kernel Test for Quasi-independence

[3] truncation refers to ordered data $X \leq Y$ and we say Y is left truncated by X.
[4] * denotes equal contributions.

Fernandez, T., **Xu, W.**, Ditzhaus, M., & Gretton, A. (2020). A Kernel Test for Quasi-independence. In Advances in Neural Information Processing Systems (NeurIPS).

4. Chapter 6 Deep Kernels for Hypothesis Testing

   Liu, F.*, **Xu, W.***, Lu, J., Zhang, G., Gretton, A., & Sutherland, D. J. (2020). Learning Deep Kernels for Non-Parametric Two-Sample Tests. International Conference on Machine Learning (ICML).

Other Contributions Works done during the course of this thesis that are not included are

- Jitkrittum, W., **Xu, W.**, Szabó, Z., Fukumizu, K., & Gretton, A. (2017). A Linear-time Kernel Goodness-of-fit Test. In Advances in Neural Information Processing Systems (NIPS).
- **Xu, W.**, Niu, G., Hyvärinen, A., & Sugiyama, M. (2019). Direction Matters: On Influence-Preserving Graph Summarization and Max-cut Principle for Directed Graphs. arXiv preprint arXiv:1907.09588. (accepted for Neural Computation)
- Wu, X. Z., **Xu, W.**, Liu, S., & Zhou, Z. H. (2020). Model Reuse with Reduced Kernel Mean Embedding Specification. arXiv preprint arXiv:2001.07135.
- **Xu, W.** & Reinert, G. (2021) A Stein Goodness-of-fit test for Exponential Random Graph Models. International Conference on Artificial Intelligence and Statistics (AISTATS).

Chapter 2

Kernel-based Hypothesis Testing

In this chapter, we provide a brief review of existing kernel-based non-parametric statistical procedures for two-sample problems, independence tests and goodness-of-fit tests, which includes preliminary knowledge and notations for the other chapters of the thesis.

2.1 Maximum Mean Discrepancy (MMD)

2.1.1 Reproducing Kernel Hilbert Space

The study of RKHS functions and related methods attract attention in statistics and machine learning community from the evolution and the success of kernel-based methods in tackling problems ranging from predictions to classifications, which exploits the rich representation of the nonlinear *feature map*. There are a few equivalent definitions for RKHS, where we present the simplest version that is sufficient for our presentation. For more detailed treatment on RKHS theory, see Berlinet and Thomas [2004] and Steinwart and Christmann [2008].

Definition 2.1 (Reproducing kernel Berlinet and Thomas [2004] Definition 1)**.** *Let $\mathcal{H}$ be a Hilbert space of real-valued functions defined on a non-empty set $\mathcal{X}$ and associated with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be a* reproducing kernel *of $\mathcal{H}$ if and only if*

1. *(inclusion)* $\forall x \in \mathcal{X}$, $k(x, \cdot) \in \mathcal{H}$
2. *(reproducing property)* $\forall x \in \mathcal{X}$, $\forall f \in \mathcal{H}$, $\langle f, k(x, \cdot) \rangle_{\mathcal{H}} = f(x)$.

The reproducing property states that the function value of f at point x is *reproduced* by the inner product of f with function $k(x, \cdot)$, which allows the evaluation of the bivariate kernel function:

$$k(x, \tilde{x}) = \langle k(x, \cdot), k(\tilde{x}, \cdot) \rangle_{\mathcal{H}}, \quad \forall (x, \tilde{x}) \in \mathcal{X} \times \mathcal{X}, \tag{2.1}$$

Throughout this thesis, such bivariate function is symmetric, i.e. $\forall(x, \tilde{x}) \in \mathcal{X} \times \mathcal{X}$,

$$k(x, \tilde{x}) = \langle k(x, \cdot), k(\tilde{x}, \cdot)\rangle_{\mathcal{H}} = \langle k(\tilde{x}, \cdot), k(x, \cdot)\rangle_{\mathcal{H}} = k(\tilde{x}, x)$$

A Hilbert space equipped with the symmetric reproducing kernel k defined in Definition 2.1 is called a *reproducing kernel Hilbert space* (RKHS) [Steinwart and Christmann, 2008], or a proper Hilbert space [Berlinet and Thomas, 2004].

We sometimes use $\mathcal{H}_k$ to explicitly refer to the RKHS $\mathcal{H}$ equipped with kernel k. Apart from being a function in $\mathcal{H}_k$, $k(x, \cdot)$ can be also interpreted as a vector in $\mathcal{H}_k$ (recall that $\mathcal{H}$ is a vector space). The reproducing property means that the evaluation of function $f(\cdot)$ at x, i.e. $f(x)$, is given by the inner product between a feature vector $k(x, \cdot)$ of x, and a feature representation of the function $f(\cdot)$, which is denoted by f. This interpretation means that $f \in \mathcal{H}$ can be seen as a parameter vector of the function $x \mapsto \langle f, k(x, \cdot)\rangle$, and consequently $\mathcal{H}$ is a space of parameter vectors which can be used to define real-valued functions. We call the function $\varphi : x \mapsto k(x, \cdot)$ the *feature map* (or *canonical feature map*) [Steinwart and Christmann, 2008, Lemma 4.19] of $\mathcal{H}_k$, which can be also seen as a parameter vector as above.

To further illustrate the concept of reproducing kernel, we use a finite dimensional feature vector as a concrete example. Let $\mathcal{X} = \mathbb{R}$ and $\varphi(x) := (\cos(x), \sin(x))^\top$. Denote $\langle \cdot, \cdot \rangle_{\mathbb{R}^2}$ as the standard dot product in $\mathbb{R}^2$, then the reproducing kernel is defined from Eq.(2.1) as

$$k(x, \tilde{x}) = \langle \varphi(x), \varphi(\tilde{x})\rangle_{\mathbb{R}^2} = \cos(x)\cos(\tilde{x}) + \sin(x)\sin(\tilde{x}). \tag{2.2}$$

While the dimension of the feature map is not restricted to be finite, it is not necessary to specify the feature map to define the kernel. Instead, one can define the evaluation of the real-valued bivariate function in Eq.(2.1) to specify the kernel. To further understand how the kernel function correspond to RKHS, we introduce the concept of positive definite (or positive type) function [Berlinet and Thomas, 2004].

Definition 2.2 (Positive-definite functions: Berlinet and Thomas [2004] Definition 2)**.** *A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called* positive definite *if $\forall n \geq 1$, $\forall a_1, \ldots, a_n \in \mathbb{R}$ and all $x_1, \ldots, x_n \in \mathcal{X}$,*

$$\sum_{i,j} a_i a_j k(x_i, x_j) \geq 0. \tag{2.3}$$

Positive definite functions are useful in learning methods such kernel ridge regression and support vector machines (SVM). Common positive definite kernel functions defined in $\mathbb{R}^d$ include linear kernels $k_0(x, \tilde{x}) = x^\top \tilde{x}$; polynomial ker-

nels $k_P(x, \tilde{x}) = (x^\top \tilde{x} + c)^d, c \geq 0, d \in \mathbb{N}$; or exponential kernels $k_E(x, \tilde{x}) = \exp(\beta x^\top \tilde{x}), \beta > 0$;.

Theorem 2.1 (Positive definite kernel and RKHS). *Assume that $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite. The following statements hold.*

1. *There exist a map $\varphi : \mathcal{X} \to \mathcal{H}$ (not unique) that for all $x, \tilde{x} \in \mathcal{X}$, we have $k(x, \tilde{x}) = \langle \varphi(x), \varphi(\tilde{x}) \rangle_{\mathcal{H}}$.*

2. *(Moore-Aronszajn theorem [Aronszajn, 1950]) There is a unique Hilbert space $\mathcal{H}$ of functions on $\mathcal{X}$ where k is the reproducing kernel.*

As such, we know that the RKHS, $\mathcal{H}$, induced from a positive definite reproducing kernel is unique. On the other hand, if a Hilbert space of functions on a non-empty set $\mathcal{X}$ is equipped with a reproducing kernel, then such kernel is also unique [Steinwart and Christmann, 2008, Theorem 4.20].

In the rest of the thesis, we denote *kernel* as the positive definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The reproducing property of the positive definite kernel enables the reformulation of learning objective that depends on the (nonlinear) feature map only through its inner product, which is known as the *kernel trick*. The infinite-dimensional nonlinear features can be powerful tools in learning procedures, and we refer the procedure of using RKHS function $f \in \mathcal{H}$, instead of linearly parametrised function $f_\theta(x)$ (θ refers to the linear weights) to perform relevant tasks as *kernelised* procedure. A commonly used kernel corresponding to an infinite-dimensional RKHS is the Gaussian kernel, also known as the squared exponential kernel or radial basis function (RBF) kernel

$$k_G(x, \tilde{x}) = \exp\left(-\frac{\|x - \tilde{x}\|^2}{2\sigma^2}\right) \tag{2.4}$$

where $\sigma > 0$ is referred to as the bandwidth or length-scale. Different choices of σ define different kernels, thus different corresponding RKHS.

2.1.2 Kernel Mean Embedding

With the rich-enough feature map, RKHS functions are used to construct mappings to represent probability distribution via kernel mean embedding [Muandet et al., 2017]. Let P be a probability distribution on $\mathcal{X}$ and k be the kernel associated with RKHS $\mathcal{H}$, the *mean embedding* of P induced by k is defined as

$$\mu_P := \mathbb{E}_{x \sim P}[k(x, \cdot)] \in \mathcal{H}, \tag{2.5}$$

whenever μ_P exist. Intuitively, the mean embedding is the expectation of the feature map under the distribution P.

Lemma 2.1 (Gretton et al. [2012a] Lemma 3)**.** *If* $\mathbb{E}_{x\sim P}[\sqrt{k(x,x)}] < \infty$*, then* μ_P *in Eq.*(2.5) *exist,* $\mu_P \in \mathcal{H}$ *and* $\mathbb{E}_{x\sim P}[f(x)] = \langle f, \mu_P\rangle_{\mathcal{H}}$*,* $\forall f \in \mathcal{H}$*.*

If the kernel k is assumed to be bounded, i.e. $\sup_{x\in\mathcal{X}} k(x,x) < \infty$ the mean embedding is well defined for all Borel probability measures [Sriperumbudur et al., 2010, Proposition 2]. For instance, the commonly used Gaussian kernel in Eq. (2.4) is upper bounded by 1. In this thesis, we assume all the reproducing kernels[1] we study are bounded.

Lemma 2.2 (Boundedness of kernels Steinwart and Christmann [2008] Lemma 4.23)**.** *Let* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *be a kernel associated with RKHS* $\mathcal{H}$*. Then* k *is bounded if and only if* $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)| < \infty, \forall f \in \mathcal{H}$*.*

Hence, the boundedness of kernel k implies the boundedness of any RKHS function in $\mathcal{H}$ and vice versa.

Kernel mean embedding allows representing the distributions as a single point in RKHS, which facilitates the comparison between distributions. For two distributions P and Q with mean embedding μ_P and μ_Q respectively, a natural way to perceive how far P is from Q might be transferred to the RKHS distance between two points $\mu_P, \mu_Q \in \mathcal{H}$,

$$D(P,Q) = \|\mu_P - \mu_Q\|^2_{\mathcal{H}}, \tag{2.6}$$

as $\mu_P - \mu_Q \in \mathcal{H}$. We will explain later that the measure of distributional difference in Eq.(2.6) actually corresponds to MMD [Gretton et al., 2012a] derived from taking supremum of functions over unit ball in the RKHS.

The kernel mean embedding in Eq.(2.5) can be estimated empirically from independent and identically distributed (i.i.d.) samples, $x_1, \ldots, x_n \sim P$:

$$\widehat{\mu}_P := \frac{1}{n}\sum_{i=1}^{n} k(x_i, \cdot) \tag{2.7}$$

replacing P by its empirical counterpart $\widehat{P} = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ where δ_{x_i} denotes the Dirac measure at $x_i \in \mathcal{X}$. The empirical mean embedding is a $\sqrt{n}$-consistent estimator for μ_P in RKHS norm depending on the i.i.d. samples [Tolstikhin et al., 2017], i.e., $\|\mu_P - \widehat{\mu}_P\|_{\mathcal{H}} = O_P(n^{-\frac{1}{2}})$ w.r.t. sample size n of the of i.i.d. samples.

[1] The reproducing kernel here need to be distinguished from the functions we study later with Stein operator, where sometimes termed as "Stein kernel".

2.1.3 Two-sample Tests with MMD

Two sample tests (or test of homogeneity) are hypothesis tests aiming to determine whether two sets of samples are drawn from the same distribution. Traditional methods such as t-tests [Student, 1908] and Kolmogorov-Smirnov (KS) tests [Lilliefors, 1967; Smirnov, 1948] are mainstays of statistical applications. However, these tests may either require strong parametric assumptions about the distributions being studied (t-tests) or may only be effective on data in extremely low-dimensional spaces (KS test[2]). Recent work in statistics and machine learning has focused on relaxing these assumptions, with methods either generally applicable or specific to various more complex domains [Gretton et al., 2012a; Jitkrittum et al., 2016a; Ramdas et al., 2017]. Other related kernel-based methods include kernel Fisher discriminant analysis [Eric et al., 2007] and tests based on checking for differences in mean embedding evaluated at specific locations such as mean embedding (ME) test or smooth characteristic function test (SCF) test [Chwialkowski et al., 2015; Jitkrittum et al., 2016a]. These tests are non-parametric and achieve very good test performance with asymptotically maximal power and well controlled type-I error, with the appropriate choice of kernels. We now introduce a popular class of the kernel-based non-parametric two-sample tests based on the *maximum mean discrepancy* (MMD) which is constructed from the kernel mean embedding introduced above.

Maximum mean discrepancy MMD has been introduced to compare two distributions [Gretton et al., 2007], utilising the reproducing property of RKHS functions and the rich representation of the kernel mean embedding. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be the kernel associated with RKHS $\mathcal{H}$. The MMD between two distributions P and Q is defined as

$$\begin{aligned} \mathrm{MMD}(P, Q; \mathcal{H}_k) &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}_{x \sim P}[f(x)] - \mathbb{E}_{x' \sim Q}[f(x')] && (2.8) \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{H}} && (2.9) \\ &= \|\mu_P - \mu_Q\|_{\mathcal{H}} && (2.10) \end{aligned}$$

The second line Eq.(2.9) follows from the reproducing property from Lemma 2.1 and the last line Eq.(2.10) follows from the fact that the inner product achieves its supremum when two vectors align and the constraint that f is in the unit ball of $\mathcal{H}$.

[2]KS test requires computation of cumulative distribution function (c.d.f.) that is applicable for univariate distribution.

From Eq.(2.9), the RKHS function that attains such supremum

$$f^*(v) \propto \mu_P(v) - \mu_Q(v) = \mathbb{E}_{x\sim P}[k(x,v)] - \mathbb{E}_{x'\sim Q}[k(x',v)], \qquad \forall v \in \mathcal{X}$$

is referred to as *witness function* [Gretton et al., 2012a, Section 2.3]. As such, the squared version of MMD corresponds to the difference between two distributions measured in RKHS norm stated in Eq.(2.6). We note that the MMD defined from the supremum notion Eq.(2.8) is an instance of an *integral probability metric* (IPM) [Müller, 1997]. More discussions and examples of IPM can be found in Sriperumbudur et al. [2010]. MMD is generally a pseudo-metric on the space of probability measures and is a metric when k is characteristic that is to be defined below. In the testing context, the characteristic notion on the kernel is very useful to impose.

Definition 2.3 (Characteristic kernels [Sriperumbudur et al., 2011])**.** *A kernel k is said to be characteristic if the mean map $P \mapsto \mu_P$ is injective on the set of Borel probability measures $\mathcal{P}$. For any distributions $P, Q \in \mathcal{P}$ and the corresponding mean embedding μ_P, μ_Q induced from characteristic kernel k,*

$$\mathrm{MMD}(P, Q; \mathcal{H}_k) = \|\mu_P - \mu_Q\|_{\mathcal{H}_k} = 0 \Longleftrightarrow P = Q.$$

The injectivity of the the mean map with characteristic kernels ensures that distinct distribution are mapped to distinct points in RKHS, allowing MMD to depart from 0 if and only if two distributions are not equal. In the two-sample testing context, such notion enables the construction of the following consistent tests against any alternatives in the distribution class $\mathcal{P}$. Consider the squared version of MMD:

$$\begin{aligned}\mathrm{MMD}^2(P, Q; \mathcal{H}_k) &= \|\mu_P - \mu_Q\|^2_{\mathcal{H}_k} = \langle \mu_P, \mu_P \rangle + \langle \mu_Q, \mu_Q \rangle - 2 \langle \mu_P, \mu_Q \rangle \\ &= \mathbb{E}_{x,x'\sim P} k(x, x') + \mathbb{E}_{y,y'\sim Q} k(y, y') - \mathbb{E}_{x\sim P, y\sim Q} k(x, y) \quad (2.11)\end{aligned}$$

Given two sets of independent identically distributed (*i.i.d.*) samples $S_P = \{x_1, \ldots, x_m\} \overset{i.i.d.}{\sim} P$ and $S_Q = \{y_1, \ldots, y_n\} \overset{i.i.d.}{\sim} Q$, an unbiased estimator of Eq.(2.11), based on the empirical estimate of kernel mean embedding in Eq.(2.7), is given by

$$\begin{aligned}&\mathrm{MMD}^2_u(S_P, S_Q; \mathcal{H}_k) \\ &= \frac{1}{m(m-1)} \sum_{i \neq i'} k(x_i, x_{i'}) + \frac{1}{n(n-1)} \sum_{j \neq j'} k(y_j, y_{j'}) - \frac{2}{mn} \sum_{ij} k(x_i, y_j).\end{aligned} \quad (2.12)$$

Recall that the two-sample problem aim to test the null hypothesis $H_0 : P = Q$ against the alternative hypothesis $H_1 : P \neq Q$. We would desire the sample sizes m and n are of the same order for nice asymptotic behaviours. Without loss of generality, assume $m = n$. It has been shown that the asymptotic distribution of $n \cdot \mathrm{MMD}_u^2(S_P, S_Q; \mathcal{H}_k)$ under the null ($P = Q$) follows infinite weighted sum of χ^2-distribution [Gretton et al., 2012a, Theorem 12], where under the alternative ($P \neq Q$), $\sqrt{n} \cdot \mathrm{MMD}_u^2(S_P, S_Q; \mathcal{H}_k)$ is asymptotically normally distributed with the mean centered at $\mathrm{MMD}(P, Q; \mathcal{H}_k) > 0$. As such, $n \cdot \mathrm{MMD}_u^2(S_P, S_Q; \mathcal{H}_k)$ is then considered as a test statistic to be compared against the *rejection threshold*. If the test statistic exceeds the rejection threshold, the empirical estimation of the MMD is thought to exhibit significant departure from the null hypothesis so that H_0 is rejected. A common choice of rejection threshold is the $(1 - \alpha)$-quantile of the null distribution with α being the predetermined level of significance of the test (or test size). As the null distribution is given by infinite weighted sum of χ^2 random variables which does not have closed form expression, the null distribution can be simulated via a permutation procedure [Gretton et al., 2008] where the rejection threshold can be determined. Alternatively, simulating the null distribution via truncated weights via eigendecomposition [Gretton et al., 2009a] or wild-bootstrap procedure [Chwialkowski et al., 2014] is also possible. More discussions on existing testing procedures can be found in Section 6.2.

Beyond the application on two-sample testing and independence testing, MMD, being a discrepancy measure between distributions, also allows applications in a wide range of machine learning problems including feature extractions [Jitkrittum et al., 2016a], covariate shift [Gretton et al., 2009b], distribution regression [Szabó et al., 2016] and generative modelling [Bińkowski et al., 2018; Li et al., 2017].

2.1.4 Hilbert-Schmidt Independence Criterion (HSIC)

Statistical tests for (in)dependence is another crucial aspect of hypothesis testings, i.e. for joint random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, testing whether a joint distribution P_{xy} factorises into the product of marginals $P_x P_y$, with the null hypothesis that $H_0 : X$ and Y are independent versus the alternative hypothesis $H_1 : X$ and Y are *not* independent. The test is performed through some dependence measure, the simplest being the correlation coefficients, which captures the dependence through *linear* features. The MMD statistic in Eq.(2.8) measures the difference between distribution over a broad class of *nonlinear* functions. As such, dependence measure through nonlinear features have been developed including kernel canonical corre-

lation (KCC) [Fukumizu et al., 2007a], constrained covariance (COCO) [Gretton et al., 2005c], normalized cross-covariance (NOCCO) [Fukumizu et al., 2007a] and complete orthogonal systems (COND) [Fukumizu et al., 2007b].

Let $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$, the cross-covariance type of dependence measure are defined through the product of the mean-centered function between f and g:

$$\begin{aligned} &\mathbb{E}_{P_{xy}}\Big[\big(f(X) - \mathbb{E}_{P_x}[f(X)]\big)\big(g(Y) - \mathbb{E}_{P_y}[g(Y)]\big)\Big] \\ &= \mathbb{E}_{P_{xy}}\Big[f(X)g(Y)\Big] - \mathbb{E}_{P_x}\Big[f(X)\Big]\mathbb{E}_{P_y}\Big[g(Y)\Big] \end{aligned} \tag{2.13}$$

The sensitivity of the statistic to dependence between X and Y heavily depends on the choice of features through functions f and g[3]. While the good choices of such features may not be easily pre-defined, Gretton et al. [2005a] developed the dependence measure, named as Hilbert-Schmidt Independence Criterion (HSIC), which is defined as the Hilbert-Schmidt norm of the cross covariance operator. Denote $k_x \in \mathcal{H}_\mathcal{X}$, $k_y \in \mathcal{H}_\mathcal{Y}$ as the kernels and associate RKHS for X and Y respectively; denote $\mu_x \in \mathcal{H}_\mathcal{X}$, $\mu_y \in \mathcal{H}_\mathcal{Y}$ as the corresponding mean embedding for marginals P_x and P_y; and denote $\otimes$ as the tensor product as in [Gretton et al., 2005a, Eq.(6)]. The cross-covariance operator $C_{xy} = \mathbb{E}_{P_{xy}}[(k_x(X,\cdot) - \mu_x) \otimes (k_y(Y,\cdot) - \mu_y)] \in \mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}$.

$$\mathrm{HSIC}^2(P_{xy}, \mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{Y}) = \|C_{xy}\|^2_{HS} \tag{2.14}$$

where $\|\cdot\|_{HS}$ denotes the Hilbert-Schmidt norm [Gretton et al., 2005a, Definition 1]. It turned out to be that the Definition in Eq.(2.14) can be reformulated from the dependence measure in Eq.(2.13) by taking the supremum of $f \otimes g$ over the joint tensor space of RKHS $\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}$,

$$\mathrm{HSIC}(P_{xy}, \mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{Y}) = \sup_{\|f\otimes g\|_{\mathcal{H}_\mathcal{X}\otimes\mathcal{H}_\mathcal{Y}} \le 1} \mathbb{E}_{P_{xy}}\Big[f(X)g(Y)\Big] - \mathbb{E}_{P_x}\Big[f(X)\Big]\mathbb{E}_{P_y}\Big[g(Y)\Big] \tag{2.15}$$

$$= \Big\|\mathbb{E}_{P_{xy}}\Big[(k_x(X,\cdot) - \mu_x) \otimes (k_y(Y,\cdot) - \mu_y)\Big]\Big\|^2_{\mathcal{H}_\mathcal{X}\otimes\mathcal{H}_\mathcal{Y}}, \tag{2.16}$$

where the RKHS norm on the tensor space $\|f \otimes g\|_{\mathcal{H}_\mathcal{X}\otimes\mathcal{H}_\mathcal{Y}} = \|f\|_{\mathcal{H}_\mathcal{X}}\|g\|_{\mathcal{H}_\mathcal{Y}}$. With the supremum notion, HSIC can be seen as an application of kernel mean embed-

[3]Choosing f and g be linear functions on X and Y respectively degenerates to covariance measure, where the dependence measure is through linear interaction.

ding where the distributions between the joint distribution P_{xy} and the product of marginals $P_x P_y$ are compared. Eq.(2.16) is equivalent to

$$\mathrm{HSIC}^2(P_{xy}, \mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{Y}) = \|\mu_{xy} - \mu_x \otimes \mu_y\|^2_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}} ,$$

where $\mu_{xy}(\cdot, *) = \mathbb{E}_{P_{xy}}[k_x(X, \cdot) \otimes k_y(Y, *)] \in \mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}$ denotes the mean embedding w.r.t the joint distribution P_{xy}.

Independence Tests with HSIC With promising statistical properties, similar to MMD, HSIC is used as a statistic to test independence. When k_x and k_y are both *characteristic kernels*, $\mathrm{HSIC} = 0$ if and only if $P_{xy} = P_x P_y$ where the independence hypothesis holds. Given a set of joint i.i.d. samples $S_{P_{xy}} = \{(x_i, y_i)\}_{i=1}^n \sim P_{xy}$, the empirical estimate of HSIC can be computed similarly as MMD in Eq.(2.12),

$$\mathrm{HSIC}^2_u(S_{P_{xy}}, \mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{Y}) = \frac{1}{(n)_2} \sum_{i \neq j} k_x(x_i, x_j) k_y(y_i, y_j) \tag{2.17}$$

$$+ \frac{1}{(n)_4} \sum_{(i,j,q,r) \in \mathbf{i}_4^n} k_x(x_i, x_j) k_y(y_q, y_r) - 2 \frac{1}{(n)_3} \sum_{(i,j,r) \in \mathbf{i}_3^n} k_x(x_i, x_j) k_y(y_i, y_r) \tag{2.18}$$

where $(n)_m := \frac{n!}{(n-m)!}$ and $\mathbf{i}_m^n$ denotes the set index of all m-tuples drawn from the index set $[n] = \{1, \ldots, n\}$ without replacement. To efficiently estimate the *unknown* null distribution, the permutation-based tests of independence by HSIC has been proposed [Gretton et al., 2008]. Subsequently, more efficient implementation of the HSIC-based independence tests have also been proposed [Jitkrittum et al., 2016b; Zhang et al., 2018].

2.2 Kernel Stein Discrepancy (KSD)

2.2.1 Stein's Method for Comparing Distributions

Stein's method [Barbour and Chen, 2005] provides an elegant probabilistic tool for comparing distributions, which has been used to tackle various problems in statistical inference, random graph theory, computational biology, etc. As Stein's method may only require to access the distributions through the differential (or difference) of the log density functions (or mass functions), it is applicable to deal with unnormalised models [Hyvärinen, 2005], which have become increasingly popular in the machine learning. Stein's method has recently caught the attention from the machine learning community [Anastasiou et al., 2021], and a variety of practical applications have been developed, including variational methods [Liu and Wang,

2016], approximate inference [Huggins and Mackey, 2018], learning implicit models [Li and Turner, 2017], non-convex optimisations [Sedghi and Anandkumar, 2014], sampling techniques [Chen et al., 2018; Gorham and Mackey, 2015], control variates [Oates et al., 2019].

Combining the Stein's method with rich representation of RKHS test functions, the kernel Stein discrepancy (KSD) [Gorham and Mackey, 2015] has been studied for comparing distributions in $\mathbb{R}^d$. Using KSD as test statistics, kernel-based non-parametric goodness-of-fit tests [Chwialkowski et al., 2016; Liu et al., 2016] have been developed, which is capable of dealing with unnormalised densities. In addition, with appropriately chosen Stein operators, KSD goodness-of-fit tests have been extended to various settings such as discrete variable models [Yang et al., 2018], point process [Yang et al., 2019], latent variable models [Kanagawa et al., 2019], and conditional densities [Jitkrittum et al., 2020]. Computationally efficient tests [Jitkrittum et al., 2017, 2018; Huggins and Mackey, 2018] have also been developed.

2.2.2 The Stein Operator on $\mathbb{R}^d$

We introduce, a Stein operator in $\mathbb{R}^d$ [Gorham and Mackey, 2015; Ley et al., 2017; Barp et al., 2019], which is the core ingredients for deriving the KSD. Let $\mathcal{X} \subset \mathbb{R}^d$ and $f_i : \mathcal{X} \to \mathbb{R}$ for $i = 1, \ldots, d$ be scalar-valued functions on $\mathcal{X}$. $\mathbf{f}(x) = (f_1(x), \ldots, f_d(x))^\top \in \mathbb{R}^d$ defines a vector valued function $\mathbf{f}$. Let q be a smooth probability density on $\mathbb{R}^d$ which *vanishes at infinity*. For a bounded smooth function $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^d$, the Stein operator $\mathcal{T}_q$ is defined by

$$\mathcal{T}_q \mathbf{f}(x) = \sum_{i=1}^{d} \left(f_i(x) \frac{\partial}{\partial x^i} \log q(x) + \frac{\partial}{\partial x^i} f_i(x) \right). \tag{2.19}$$

The operator $\mathcal{T}_q$ is called a Stein operator if the Stein's identity holds for *all* f in an appropriate class of functions

$$\mathbb{E}_q[\mathcal{T}_q \mathbf{f}] = 0. \tag{2.20}$$

The operator in Eq.(2.19) can be rewrite in the form

$$\mathcal{T}_q f(x) = \sum_i \frac{1}{q(x)} \frac{\partial}{\partial x_i} \left(f_i(x) q(x) \right)$$

and is a Stein operator due to integration by parts on $\mathbb{R}^d$,

$$\mathbb{E}_q[\mathcal{T}_q\mathbf{f}] = \int_{\mathbb{R}^d} \mathcal{T}_q\mathbf{f}(x)q(x)dx = \sum_i \int_{\mathbb{R}^d} \frac{\partial}{\partial x^i}\left(f_i(x)q(x)\right)dx = 0.$$

The last equality holds since $f_i(x)q(x)$ vanishes at infinity. The function $\mathbf{f}$ here is referred to as *test function*. The class of functions that Eq.(2.20) holds, is called *Stein class* of q. Since Stein operator $\mathcal{T}_q$ depends on the density q only through the derivatives of $\log q$, it does not involve the normalisation constant of q, which is a useful property for dealing with unnormalised models [Hyvärinen, 2005].

Stein operator can be used to compare two distributions via a class of test functions. Let p, q be two smooth densities in $\mathbb{R}_d$, vanishing at infinity, the Stein discrepancy between p, q through function $\mathbf{f}$ is

$$\mathrm{SD}(p\|q; f) = \mathbb{E}_p[\mathcal{T}_q\mathbf{f}] - \mathbb{E}_q[\mathcal{T}_q\mathbf{f}] = \mathbb{E}_p[\mathcal{T}_q\mathbf{f}], \tag{2.21}$$

where the last equality holds by Stein's identity. We also know that, for any test functions in the Stein class, $p = q$ implies $\mathrm{SD}(p\|q; f) = 0$.

Assume the test function class to be a RKHS, i.e. $f_i \in \mathcal{H}, \forall i$, where each scalar valued functions lives in the same RKHS [Chwialkowski et al., 2016; Liu et al., 2016]. The kernel Stein discrepancy (KSD) between density p and q is defined via taking the supremum over the unit ball RKHS function class, similar to Eq.(2.8) :

$$\mathrm{KSD}(p\|q) = \sup_{\|\mathbf{f}\|_{\mathcal{H}^d} \leq 1} \mathbb{E}_p[\mathcal{T}_q\mathbf{f}]. \tag{2.22}$$

As shown by the boundedness property of kernels in Lemma 2.2, a bounded kernel corresponds to a bounded RKHS test function class. From properties of Stein discrepancy, we know that if $p = q$, we have $\mathrm{KSD}(q\|p) = 0$. However, in the testing procedure, a desirable property of the discrepancy measure is that $\mathrm{KSD}(q\|p) = 0$ if and only if $p = q$. This correspond to the *characteristic* notion in MMD construction in Definition 2.3. As such, we require our RKHS to be sufficiently large to capture any possible discrepancies between p and q. Under mild regularity conditions,

- $\mathcal{H}$ equipped with C_0-universal kernel k [Carmeli et al., 2010, Definition 4.1]
- $\mathbb{E}_{x\sim q}\left[\langle \mathcal{T}_q k(x,\cdot), \mathcal{T}_q k(x,\cdot)\rangle_{\mathcal{H}}\right] < \infty$
- $\mathbb{E}_q\left\|\sum_i \frac{\partial}{\partial x_i}\log\frac{p(x)}{q(x)}\right\|^2 < \infty$

it is shown that $\mathrm{KSD}(p\|q) \geq 0$ and $\mathrm{KSD}(p\|q) = 0$ if and only if $p = q$ [Chwialkowski et al., 2016, Theorem 2.2]. Thus, KSD is a proper discrepancy measure between densities.

2.2.3 Goodness-of-fit Tests with KSD

The goodness-of-fit testing procedure aims to check the hypothesis $H_0 : q = p$, where q is the target distribution required up to normalization constant; and p is the *unknown* data distribution only accessible from samples, $x_1, \ldots, x_n \sim p$. Since p is unknown, algebraic manipulations produce the following form of $\mathrm{KSD}(p\|q)$:

$$\mathrm{KSD}^2(p\|q) = \mathbb{E}_{x,\tilde{x}\sim p}[h_q(x,\tilde{x})], \tag{2.23}$$

where $h_q(x,\tilde{x}) = \langle \mathcal{T}_q k(x,\cdot), \mathcal{T}_q k(\tilde{x},\cdot)\rangle$ does not involve p. $k(x,\cdot)$ denotes the kernel associated with $\mathcal{H}$. As such, KSD can be empirically estimated via Eq.(2.23) using U-statistics or V-statistics. The critical value is determined by bootstrap based on the theory of U-statistics or V-statistics. In this way, a goodness-of-fit testing procedure on $\mathbb{R}^d$ is obtained, which is applicable to unnormalised models.

Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) on $\mathbb{R}^d$ and $\mathcal{H}^d$ be its product. By using Stein operator, kernel Stein discrepancy (KSD) [Chwialkowski et al., 2016; Liu et al., 2016] between two densities p and q is defined as $\mathrm{KSD}(p\|q) = \sup_{\|\mathbf{f}\|_{\mathcal{H}^d}\leq 1} \mathbb{E}_p[\mathcal{T}_q \mathbf{f}]$.

Now, suppose we have samples $x_1, \ldots, x_n$ from unknown density p on $\mathbb{R}^d$. Then, an empirical estimate of $\mathrm{KSD}^2(p\|q)$ is obtained by using Eq.(2.23) in the form of U-statistics, and this estimate can be used to test the hypothesis $H_0 : p = q$, where the critical value is determined by bootstrap. In this way, a general method of goodness-of-fit test on $\mathbb{R}^d$ is obtained, which does not require computation of the normalisation constant.

The kernel Stein discrepancy (KSD) [Gorham and Mackey, 2015; Ley et al., 2017] is a discrepancy measure between distributions that is based on Stein method [Barbour and Chen, 2005; Chen et al., 2010] and reproducing kernel Hilbert space (RKHS) theory [Berlinet and Thomas, 2004]. KSD provides a general procedure for goodness-of-fit testing that does not require computation of the normalization constant, and it has shown state-of-the-art performance in various scenarios including Euclidean data [Chwialkowski et al., 2016; Liu et al., 2016], discrete data [Yang et al., 2018], point processes [Yang et al., 2019], censored data [Fernandez et al., 2020] and directional data [Xu and Matsuda, 2020]. In addition, by using the technique of kernel mean embedding [Muandet et al., 2017], KSD test also enables extraction of distributional features to perform model criticism [Jitkrittum et al.,

2017, 2018; Kanagawa et al., 2019; Jitkrittum et al., 2020]. We note that Stein's method has recently been extended to Riemannian manifolds and applied to numerical integration [Barp et al., 2018] and Bayesian inference [Liu and Zhu, 2018].

Chapter 3

Goodness-of-fit Tests on non-Euclidean Data

Summary We address the problem of goodness-of-fit testing and model criticism on non-Euclidean data such as hyperspheres, torus or rotation groups. Due to the different topologies, the standard statistical procedures for multivariate data in $\mathbb{R}^d$ are not applicable to such data. We first derive goodness-of-fit testing and model criticism methods for directional distribution, and then study its generalisation to testing and comparing distributions on smooth Riemannian manifolds, especially for those with an intractable normalisation constant. The proposed methods are based on Stein operators on Riemannian manifolds. Simulation results and real data applications show the superior test performances and useful interpretability of the proposed methods.

3.1 Introduction

In many scientific and machine learning applications, data is obtained in the form of directions and they are naturally identified with a vector on the unit hypersphere $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d \mid \|x\| = 1\} \subset \mathbb{R}^d$. For example, wind direction is represented by a vector on the unit circle $\mathbb{S}^1 \subset \mathbb{R}^2$ [Genton and Hering, 2007; Hering and Genton, 2010]; while the protein structure is described by vectors on the unit sphere $\mathbb{S}^2 \subset \mathbb{R}^3$ [Hamelryck et al., 2006]. In addition, usual multivariate data in $\mathbb{R}^d$ is transformed to directional data by applying normalisation, and such transformation is useful to analyse scale-invariant features. For example, text document and gene expression data are transformed into directional data and applied model-based clustering [Banerjee et al., 2005]. Moreover, it has been showed that projecting face images to a unit hypersphere can improve face recognition performance by convolutional neural networks [Wang et al., 2017]. Statistical methods for such directional data have been widely studied in the field of directional statistics [Mardia and Jupp, 1999; Ley and Verdebout, 2017], and many statistical models of directional distributions have been proposed.

One characteristic feature of directional distributions is that they often involve an intractable normalisation constant. For example, the Fisher-Bingham distribution [Kent, 1982] is defined by an unnormalised density

$$p(x \mid A, b) \propto \exp(x^\top A x + b^\top x), \quad x \in \mathbb{S}^{d-1} \tag{3.1}$$

and its normalisation constant is not represented in closed form. Such intractable normalisation constant makes statistical inferences for directional distributions computationally difficult. While directional data are becoming increasingly important in many applications such as bioinformatics, meteorology, chronobiology, and text/image analysis, to the best of our knowledge, non-parametric goodness-of-fit testing procedures for general directional distributions is not well established. Despite statistical inference methods have been developed to directly deal with unnormalised models in Euclidean space $\mathbb{R}^d$ [Hyvärinen, 2005], they are not readily applicable to non-Euclidean data, i.e., applying the non-parametric goodness-of-fit tests described in Section 2.2 [Chwialkowski et al., 2016; Liu and Wang, 2016] are not consistent and do not produce controlled type-I error.

Generalising from directional distributions, data also commonly appear in the domains described by Riemannian manifolds. For example, structures of biological molecules can be described by a pair of angular variables, which is identified with a point on the torus [Singh et al., 2002]. In computer vision, the orientation of a camera is represented by a 3×3 rotation matrix, which gives rise to data on the rotation group [Song et al., 2009]. Other examples include the orbit of a comet [Jupp et al., 1979] and the vectorcardiogram data [Downs, 1972]. In addition, shape analysis [Dryden and Mardia, 2016] and compositional data analysis [Pawlowsky-Glahn and Buccianti, 2011] also deal with complex data defined on Riemannian manifolds. Since the usual statistical procedures for Euclidean data are not applicable, many studies have developed statistical models and methods tailored for data on Riemannian manifolds [Chikuse, 2003, 2012; Hoff, 2009].

Statistical models on Riemannian manifolds are also often given in the form of unnormalised densities where the normalisation constants remain computationally intractable. For example, the Fisher distribution on the rotation group [Chikuse, 2012; Sei et al., 2013] is defined by

$$p(X \mid \Theta) \propto \exp(\mathrm{tr}(\Theta^\top X)), \quad X, \Theta \in \mathbb{R}^{3\times 3} \tag{3.2}$$

and the normalisation constant involves integrating over $\mathbb{R}^{3\times 3}$ which are unable to express in closed form. Similar to the directional distributions, statistical inference

with such models are computationally intensive due to the intractable normalisation constant. Thus, statistical methods on Riemannian manifolds that do not require computation of the normalisation constant have been developed for several tasks such as parameter estimations [Mardia et al., 2016] and sampling techniques on the manifold [Byrne and Girolami, 2013; Ma et al., 2015]. However, non-parametric goodness-of-fit testing for general distributions on Riemannian manifolds are not yet established.

Contributions In this chapter, we develop and analyse novel non-parametric goodness-of-fit testing procedures for general Riemannian manifold distributions including directional distributions by extending kernel Stein discrepancy. We first derive a Stein operator and the corresponding KSD test that is applicable to Riemannian manifold, using directional distributions as a motivating example. Then we discuss other possible Stein operators and the kernel-based goodness-of-fit tests for distributions on general Riemannian manifolds. Comparison of relative test efficiencies between these tests, in terms of Bahadur slope, are provided. In addition, we perform model criticism on Riemannian manifold based on a modified finite-set Stein discrepancy (FSSD) statistic, which is learned from maximising test power. We show that the proposed methods control type-I error well and have better test power compared to existing alternative tests via simulations. Real data applications show the usefulness of extracting interpretable features via the proposed model criticism procedure.

3.2 Unnormalised Distributions

3.2.1 Directional Distributions

The density models defined on the unit hyperspheres $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d \mid \|x\| = 1\}$ are used to describe the directional data and are referred to as *directional distributions* [Mardia and Jupp, 1999]. Here we present two representative directional distributions alongside with the uniform distributions which are widely used: the von-Mises-Fisher distribution and the Fisher-Bingham distribution. Figure 3.1 shows illustrative examples of these distributions via samples on $\mathbb{S}^2$.

We define the probability density of directional distributions by taking the uniform distribution on $\mathbb{S}^{d-1}$ as base measure. Namely, the unnormalised density of the uniform distribution is constant, $p(x) \propto 1, \forall x \in \mathbb{S}^{d-1}$. The von-Mises-Fisher (or von-Mises when $d = 2$) distribution is a directional counterpart of the isotropic

(a) Uniform Distribution (b) von-Mises-Fisher (c) Fisher-Bingham

Figure 3.1: Samples from directional distributions on $\mathbb{S}^2$

Gaussian distribution on $\mathbb{R}^d$. Its unnormalised density is given by

$$p(x \mid \mu, \kappa) = \frac{1}{C_d(\kappa)} \exp(\kappa \mu^\top x), \tag{3.3}$$

for $x \in \mathbb{S}^{d-1}$, where $\mu \in \mathbb{S}^{d-1}$, $\kappa > 0$,

$$C_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)},$$

and I_v is the modified Bessel function of the first kind and order v. It is a unimodal distribution with peak at μ[1] and degree of concentration specified by κ[2].

The Fisher-Bingham (or Kent) distribution is an extension of the von-Mises-Fisher distribution [Kent, 1982], where the log-likelihood includes second order terms. Its unnormalised density is given by

$$p(x \mid A, b) = \frac{1}{Z(A, b)} \exp(x^\top A x + b^\top x), \tag{3.4}$$

for $x \in \mathbb{S}^{d-1}$, where $A \in \mathbb{R}^{d \times d}$ is symmetric and $b \in \mathbb{R}^d$. The normalisation constant $Z(A, b)$ is not represented in closed form and hard to compute in general.

Kent distribution sometimes specifically refers to the so-called 5-parameter Fisher-Bingham distribution (FB5) [Kent et al., 1979], where for specific class of parameters A and b in Eq. (3.4), the unnormalised density can be simplified the the form of

$$p(x) \propto \exp(\kappa u^\top x + \beta((\gamma_1^\top x)^2 + \gamma_2^\top x)^2), \quad x \in \mathbb{S}^{d-1}.$$

The 8-parameter Fisher-Bingham distribution (FB8) are also studied [Yuan, 2019],

[1]The parameter μ is analogous to the "mean parameter" in Gaussian density in $\mathbb{R}^d$.
[2]The parameter κ is analogous to the inverse of "variance parameter" in Gaussian density in $\mathbb{R}^d$.

where the unnormalised density can be simplified as

$$p(x) \propto \exp(\kappa \nu^\top \Gamma x + \beta_1((\gamma_1^\top x)^2 + \beta_2(\gamma_2^\top x)^2).$$

An example of the Fisher-Bingham on $\mathbb{S}^2$ is shown in Figure.3.1(c).

The goodness-of-fit testing procedures for directional distributions are mainly parametric and limited to testing for specific distributions such as uniform [Figueiredo, 2007; García-Portugués and Verdebout, 2018; Mardia and Jupp, 1999] and von-Mises-Fisher [Figueiredo, 2012; Mardia et al., 1984]. Although [Boente et al., 2014] proposed testing procedures based on the kernel density estimator which are non-parametric, they are difficult to apply to unnormalised models such as the Fisher-Bingham distribution in Eq. (3.4) due to the requirement of the normalisation constant from the null model to calculate the L^p test statistics.

3.2.2 Distributions on General Riemannian Manifolds

A Riemannian manifold describes a topological manifold, $\mathcal{M}$, with additional geometric structure called *Riemannian metric*, g, a smoothly varying inner product on tangent space on manifold. More specifically, g is a covariant 2-tensor field on $\mathcal{M}$ whose value at each point $\mathrm{p} \in \mathcal{M}$ defines a positive definite inner product on the tangent space $\mathrm{T_p}\mathcal{M}$ at p. A manifold is said to be compact if its underlying topological space is compact. $\mathcal{M}$ may have non-empty boundary, denoted by $\partial\mathcal{M}$. See Kobayashi and Nomizu [1963] for details on Riemannian geometry. Denote a smooth Riemannian manifold by $(\mathcal{M}, g)$. Lee [2018] shows that that every smooth manifold admits a Riemannian metric. In this chapter, we use manifold to refer a compact smooth Riemannian manifold when no ambiguity arises. We note that the hypersphere can be seen as such a manifold. Here, we give two additional examples of manifold that are commonly seen in practice. Note that we define the probability density of each distribution by its Radon–Nikodym derivative with respect to the volume element of $(\mathcal{M}, g)$.

Torus The torus $\mathbb{S}^1 \times \mathbb{S}^1$ is the direct product of two circles $\mathbb{S}^1$ and the bivariate circular data $(x_1, x_2) \in [0, 2\pi)^2$ can be viewed as data on the torus $\mathbb{S}^1 \times \mathbb{S}^1$, where we identify $(\cos x, \sin x) \in \mathbb{S}^1$ with $x \in [0, 2\pi)$. To describe dependence between the two circular variables[3], Singh et al. [2002] proposed the bivariate von-Mises

[3]We note that the torus, $\mathbb{S}^1 \times \mathbb{S}^1$, describing two circular variables has different topological structure than a unit sphere $\mathbb{S}^2$ where the domain $(\tilde{x}_1, \tilde{x}_2) \in [0, 2\pi) \times [0, \pi)$ for $(\tilde{x}_1, \tilde{x}_2) \in \mathbb{S}^2$.

distribution:

$$p(x_1, x_2 \mid \xi) \propto \exp(\kappa_1 \cos(x_1 - \mu_1) + \kappa_2 \cos(x_2 - \mu_2) + \lambda_{12} \sin(x_1 - \mu_1) \sin(x_2 - \mu_2)), \tag{3.5}$$

where $\xi = (\kappa_1, \kappa_2, \mu_1, \mu_2, \lambda_{12})$, $\kappa_1 \geq 0$, $\kappa_2 \geq 0$, $0 \leq \mu_1 < 2\pi$ and $0 \leq \mu_2 < 2\pi$. The normalisation constant can not be represented in closed form for general ξ. We will apply this model to wind direction data in Section 3.5.

Rotation group The rotation group $\mathrm{SO}(m)$ is defined as

$$\mathrm{SO}(m) = \{X \in \mathbb{R}^{m \times m} \mid X^\top X = I_m, \det X = 1\},$$

where I_m is the m-dimensional identity matrix. The Fisher distribution [Chikuse, 2012; Sei et al., 2013] on $\mathrm{SO}(m)$ is defined as

$$p(X \mid \Theta) \propto \exp(\mathrm{tr}(\Theta^\top X)), \tag{3.6}$$

for which the normalisation constant is not given in closed form. We will apply this model to vectorcardiogram data in Section 3.5. More discussions on general form of exponential-trace type of distribution for matrix-valued manifold can be found in Chikuse [2003]; Hoff [2009][4].

Statistical methods have been developed to deal with models for the Riemannian manifold. For sampling procedure, MCMC techniques on Riemannian manifolds have been developed [Byrne and Girolami, 2013; Ma et al., 2015; Hoff, 2019]. For parameter estimation, score matching [Hyvärinen, 2005] has been extended to specific Riemannian manifold [Mardia et al., 2016; Mardia, 2018]. The goodness-of-fit testing procedures on general Riemannian manifolds are less investigated, even with the parametric models. For the special case of the uniform distribution, testing procedures were developed by [Chikuse and Jupp, 2004; Chikuse, 2012] and they have been extended the Sobolev test for uniformity [Giné, 1975] based on estimating the model parameters Jupp et al. [2005, 2008]. However, these methods are not applicable to general cases. To overcome this, the Sobolev test has been extended to general cases Jupp et al. [2005] and transformation based test via

[4]For instance, similar to directional distribution, the Bingham-von-Mises-Fisher (BMF) distribution for matrix-valued variable [Hoff, 2009] has unnormalised density of the form,

$$p(X \mid A, B, F) \propto \exp(\mathrm{tr}(F^\top X + B X^\top A X)).$$

sliced cumulative distribution function has been developed Jupp and Kume [2018], For tests of uniformity, several methods have been proposed such as the Sobolev test [Chikuse and Jupp, 2004; Giné, 1975; Jupp et al., 2008]. However, they are not readily applicable to general distributions. Although there are a few methods applicable to general distributions [Jupp et al., 2005; Jupp and Kume, 2018], they require computation of the normalisation constant, which is often computationally intensive. Also, existing methods cannot be applied to model criticism [Jitkrittum et al., 2016a], which would provide an intuitive clarification of the discrepancy between model and data.

3.3 Stein Operators on Manifold

In this section, we introduce several Stein operators of different type for distributions on Riemannian manifolds by using Stokes' theorem. The operators are categorised via the order of differentials of the test functions[5].

3.3.1 Differential Forms and Stokes' Theorem

To derive Stein operators on Riemannian manifolds, we need to use differential forms and Stokes' theorem. Here, we briefly introduce these concepts. For more detailed and rigorous treatments, see Flanders [1963]; Spivak [2018].

Let $\mathcal{M}$ be a smooth d-dimensional Riemannian manifold and take its local coordinate system $x^1, \ldots, x^d$. We introduce symbols $\mathrm{d}x^1, \ldots, \mathrm{d}x^d$ and an associative and anti-symmetric operation $\wedge$ between them called the wedge product: $\mathrm{d}x^i \wedge \mathrm{d}x^j = -\mathrm{d}x^j \wedge \mathrm{d}x^i$. Note that $\mathrm{d}x^i \wedge \mathrm{d}x^i = 0$. Then, a p-form ω on M $(0 \leq p \leq d)$ is defined as

$$\omega = \sum_{i_1 \cdots i_p} f_{i_1 \cdots i_p} \mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_p}, \tag{3.7}$$

where the sum is taken over all p-tuples $\{i_1, \cdots, i_p\} \subset \{1, \ldots, d\}$ and each $f_{i_1 \cdots i_p}$ is a smooth function on $\mathcal{M}$. For a p-form ω above and a q-form $\eta = \sum_{j_1 \cdots j_q} g_{j_1 \cdots j_q} \mathrm{d}x^{j_1} \wedge \cdots \wedge \mathrm{d}x^{j_q}$ with $p+q \leq d$, their wedge product $\omega \wedge \eta$ is defined as the $(p+q)$-form given by

$$\omega \wedge \eta = \sum_{i_1 \cdots i_p} \sum_{j_q \cdots j_q} f_{i_1 \cdots i_p} g_{j_1 \cdots j_q} \mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_p} \wedge \mathrm{d}x^{j_1} \wedge \cdots \wedge \mathrm{d}x^{j_q}. \tag{3.8}$$

[5]Note that this does not refer to the differentials for the (unnormalised) density functions.

The exterior derivative $\mathrm{d}\omega$ of ω is defined as the $(p+1)$-form given by

$$\mathrm{d}\omega = \sum_{i_1 \cdots i_p} \sum_{i=1}^{d} \frac{\partial f_{i_1 \cdots i_p}}{\partial x^i} \mathrm{d}x^i \wedge \mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_p}, \tag{3.9}$$

where $\mathrm{d}f$ for a function f is the 1-form defined by

$$\mathrm{d}f = \sum_{i=1}^{d} \frac{\partial f}{\partial x^i} \mathrm{d}x^i.$$

For another coordinate system $y^1, \ldots, y^d$ on $\mathcal{M}$, the differential form can be transformed from the coordinate of $x^1, \ldots, x^d$ by

$$\mathrm{d}y^j = \sum_{i=1}^{d} \frac{\partial y^j}{\partial x^i} \mathrm{d}x^i. \tag{3.10}$$

The differential forms and exterior derivatives will be useful for discussing Stokes' theorem on manifold. For example, the volume element, with respect to the coordinate $x^1, \ldots, x^d$, is defined as the d-form given by

$$(\det g)^{1/2} \mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^d,$$

where $g = g(x^1, \ldots, x^d)$ is the $d \times d$ matrix of the Riemannian metric with respect to $x^1, \ldots, x^d$.

The integration of a d-form on a d-dimensional manifold is naturally defined like the usual integration on $\mathbb{R}^d$ and invariant with respect to the coordinate selection. Correspondingly, the integration by parts formula on $\mathbb{R}^d$ is generalised in the form of Stokes' theorem.

Theorem 3.1 (Stokes' theorem)**.** *Let $\partial\mathcal{M}$ be the boundary of $\mathcal{M}$ and ω be a $(d-1)$-form on $\mathcal{M}$. Then,*

$$\int_{\mathcal{M}} \mathrm{d}\omega = \int_{\partial\mathcal{M}} \omega.$$

Corollary 3.1. *If $\partial\mathcal{M}$ is empty, then*

$$\int_{\mathcal{M}} \mathrm{d}\omega = 0$$

for any $(d-1)$-form ω on $\mathcal{M}$.

Coordinate choice In the following, to facilitate the derivation as well as computation of Stein operators, we assume that there exists a coordinate system $\theta^1, \dots, \theta^d$ on $\mathcal{M}$ that covers $\mathcal{M}$ almost everywhere. For example, spherical coordinate system $\theta = (\theta^1, \dots, \theta^{d-1})$ can be considered on hypersphere $\mathbb{S}^{d-1}$, which is defined by

$$\begin{pmatrix} \theta^1 \\ \theta^2 \\ \theta^3 \\ \vdots \\ \theta^{d-1} \end{pmatrix} \mapsto \begin{pmatrix} \cos\theta^1 \\ \sin\theta^1 \cos\theta^2 \\ \sin\theta^1 \sin\theta^2 \cos\theta^3 \\ \vdots \\ \sin\theta^1 \cdots \sin\theta^{d-1} \end{pmatrix} \in \mathbb{S}^{d-1}, \tag{3.11}$$

where $(\theta^1, \dots, \theta^{d-2}) \in [0, \pi)^{d-2}$ and $\theta^{d-1} \in [0, 2\pi)$. In this coordinate system, the volume element [Flanders, 1963] is given by

$$\mathrm{d}S = J(\theta^1, \dots, \theta^{d-1}) \mathrm{d}\theta^1 \wedge \cdots \wedge \mathrm{d}\theta^{d-1},$$

where $J(\theta^1, \dots, \theta^{d-1}) = \sin^{d-2}(\theta^1) \sin^{d-3}(\theta^2) \cdots \sin(\theta^{d-2})$.

Note that $J(\theta^1) = 1$ when $d = 2$. Since the surface area of $\mathbb{S}^{d-1}$ is $S_{d-1} = 2\pi^{d/2}/\Gamma(d/2)$, the uniform distribution on $\mathbb{S}^{d-1}$ corresponds to the $(d-1)$-form η on $\mathbb{S}^{d-1}$, given by

$$\eta = \frac{1}{S_{d-1}} J(\theta^1, \dots, \theta^{d-1}) \mathrm{d}\theta^1 \wedge \cdots \wedge \mathrm{d}\theta^{d-1}.$$

By using the uniform density as the base measure, the directional distribution on $\mathbb{S}^{d-1}$ with density p is represented by the $(d-1)$-form ω given by

$$\omega = p\eta.$$

Thus, expectation of a function g with respect to p is obtained by

$$\mathbb{E}_p[g] = \int_{\mathbb{S}^{d-1}} g\omega = \frac{1}{S_{d-1}} \int_0^{2\pi} \int_0^{\pi} \cdots \int_0^{\pi} g(\theta) p(\theta) J(\theta) \mathrm{d}\theta^1 \cdots \mathrm{d}\theta^{d-1}.$$

Similarly, the polar coordinate system is useful for the torus. In addition, Stereographical projection can also be useful coordinate of choice. The generalised Euler angles [Chikuse, 2012, Section 2.5.1] for the rotation groups, and Givens rotations [Pourzanjani et al., 2017] for the Stiefel manifolds are also useful coordinate system satisfying the above assumption.

3.3.2 First Order Stein Operator

For a smooth probability density q on $\mathcal{M}$ and a smooth function $\mathbf{f} = (f^1, \ldots, f^d) : \mathcal{M} \to \mathbb{R}^d$ and coordinate system $(\theta^1, \ldots, \theta^d)$, define a function $\mathcal{A}_q^{(1)}\mathbf{f} : \mathcal{M} \to \mathbb{R}$ by

$$\mathcal{A}_q^{(1)}\mathbf{f} = \sum_{i=1}^{d} \left(\frac{\partial f^i}{\partial \theta^i} + f^i \frac{\partial}{\partial \theta^i} \log(qJ) \right), \tag{3.12}$$

where $J = (\det g)^{1/2}$ is the volume element.

Theorem 3.2 (Stein's identity)**.** *If $\partial\mathcal{M}$ is empty or $f^1, \ldots, f^d$ vanish on $\partial\mathcal{M}$, then*

$$\mathbb{E}_q[\mathcal{A}_q^{(1)}\mathbf{f}] = 0.$$

The Stein operator only involves the first order differential w.r.t. to test function $\mathbf{f} = (f^1, \ldots, f^d)$. Thus we refer $\mathcal{A}^{(1)}$ as a first order Stein operator.

Proof. Let

$$\omega = \sum_{i=1}^{d} f^i \mathrm{d}\theta^{(-i)},$$

where $\mathrm{d}\theta^{(-i)} = \mathrm{d}\theta^{i+1} \wedge \cdots \wedge \mathrm{d}\theta^d \wedge \mathrm{d}\theta^1 \cdots \wedge \mathrm{d}\theta^{i-1}$ for $i = 1, \ldots, d$. Then,

$$\begin{aligned} \mathrm{d}(qJ\omega) &= \sum_{i=1}^{d} \left(\frac{\partial f^i}{\partial \theta^i} + f^i \frac{\partial}{\partial \theta^i} \log(qJ) \right) qJ \mathrm{d}\theta^1 \wedge \cdots \wedge \mathrm{d}\theta^d \\ &= (qJ\mathcal{A}_q^{(1)}\mathbf{f}) \mathrm{d}\theta^1 \wedge \cdots \wedge \mathrm{d}\theta^d. \end{aligned}$$

Therefore, from Theorem 3.1 and Corollary 3.1,

$$\mathbb{E}_q[\mathcal{A}_q^{(1)}\mathbf{f}] = \int_{\mathcal{M}} \mathrm{d}(qJ\omega) = 0,$$

and the Stein's identity follows. □

The boundary assumption of Theorem 3.2 is in the same fashion as Assumption 4 in Barp et al. [2018] to make the Stein's identity hold. If $\mathcal{M}$ is a closed manifold such as hyperspheres, torus or rotation group, it does not have boundary by definition and thus the assumption of Theorem 3.2 holds. If the boundary of $\mathcal{M}$ is non-empty, $\mathbf{f}(\partial\mathcal{M}) = 0$ can be imposed by choosing specific test function class. A relevant discussion of Theorem 3.2 can be found in density estimation on truncated domain context [Liu and Kanamori, 2019].

3.3.3 Second Order Stein Operator

In the context of numerical integration on Riemannian manifolds, Barp et al. [2018] introduced another type of Stein operator $\mathcal{A}_q^{(2)}$, which involves the second order differential operators w.r.t. the test functions. Here, we refer $\mathcal{A}_q^{(2)}$ as the second order Stein operator. Specifically, for a smooth probability density q on $\mathcal{M}$ and a smooth function $\tilde{f} : \mathcal{M} \to \mathbb{R}$, define $\mathcal{A}_q^{(2)}\tilde{f} : \mathcal{M} \to \mathbb{R}$ by

$$\mathcal{A}_q^{(2)}\tilde{f} = \sum_{ij}\left(g^{ij}\frac{\partial^2\tilde{f}}{\partial\theta^i\partial\theta^j} + g^{ij}\frac{\partial\tilde{f}}{\partial\theta^j}\frac{\partial\log J}{\partial\theta^i} + g^{ij}\frac{\partial\tilde{f}}{\partial\theta^j}\frac{\partial\log q}{\partial\theta^i}\right), \tag{3.13}$$

where we denote the inverse matrix of (g_{ij}) by (g^{ij}) following the convention of Riemmanian geometry.

Remark Denote ∂x^i as the basis vector on tangent space of x, $\mathrm{T_x}\mathcal{M}$. In Barp et al. [2018], the Stein operator is defined via the (Riemannian) gradient operator

$$\nabla\tilde{f} = \sum_{i,j}[G^{-1}]_{i,j}\frac{\partial\tilde{f}}{\partial x^j}\partial x^i,$$

where $G \in \mathbb{R}^{d\times d}$ denotes the metric tensor matrix; the divergence operator

$$\nabla\cdot\mathbf{s} = \sum_i \frac{\partial s_i}{\partial x^i} + s_i\frac{\partial}{\partial x^i}\log\sqrt{\det(G)}$$

for $\mathbf{s} = s_1\partial x^1, \ldots, s_d\partial x^d$; and the Laplace–Bertrami operator $\Delta\tilde{f} = \nabla\cdot\nabla\tilde{f}$. The second order Stein operator can be written in the form

$$\widetilde{\mathcal{A}}_q^{(2)}\tilde{f} = \langle\nabla\tilde{f}, \nabla\log q\rangle + \Delta\tilde{f}. \tag{3.14}$$

Theorem 3.3 (Proposition 1 of Barp et al. [2018]). *If $\partial\mathcal{M}$ is empty or $\int_{\partial\mathcal{M}}\sum_{ij} g^{ij}p(\nabla\tilde{f})_i\cdot\mathbf{n}_j i_{\mathbf{n_j}}dV = 0$ for normal vector $\mathbf{n}$ and its associated volume $i_{\mathbf{n}}dV$, then*

$$\mathbb{E}_q[\mathcal{A}_q^{(2)}\tilde{f}] = 0.$$

Theorem 3.3 follows from Theorem 3.2, because the second order Stein operator in Eq.(3.13) can be viewed as a special case of the first order Stein operator in Eq.(3.12) with

$$f^i = \sum_j g^{ij}\frac{\partial\tilde{f}}{\partial\theta^j}.$$

Similar form of the second order Stein operator in Eq. (3.13) (or Eq. (3.14)) has been studied in Liu and Zhu [2018] for Bayesian inference. On the other hand, Le et al. [2020] arrives at a similar second order Stein operator in the context of density approximation by considering an infinitesimal generator of the Feller's diffusion process whose stationary distribution is q.

3.3.4 Zeroth Order Stein Operator

For a smooth probability density q on $\mathcal{M}$ and a function $h : \mathcal{M} \to \mathbb{R}$, define a function $\mathcal{A}_q^{(0)} h : \mathcal{M} \to \mathbb{R}$ by

$$\mathcal{A}_q^{(0)} h = h - \mathbb{E}_q[h].$$

It is easy to see that $\mathbb{E}_q[\mathcal{A}_q^{(0)} h] = 0$. Since $\mathcal{A}_q^{(0)}$ does not involve any differential operators, we refer it as the zeroth order Stein operator. Compared to the first and second order Stein operators, this operator requires the normalisation constant of q, which is often computationally intractable for Riemannian manifolds. We will show later that this operator corresponds to the maximum mean discrepancy [Gretton et al., 2012a]. We will also compare the test performances build on different Stein operators via corresponding manifold Kernel Stein Discrepancy (mKSD) that we now introduce.

3.4 Goodness-of-fit Tests on Manifold

In this section, we propose goodness-of-fit testing and model criticism procedures for distributions on Riemannian manifolds based on the Stein operators introduced in the previous section.

3.4.1 Manifold Kernel Stein Discrepancies (mKSD)

By using Stein operators introduced in the previous section, we extend kernel Stein discrepancy to distributions on Riemannian manifolds.

Let $\mathcal{H}$ be a RKHS on $\mathcal{M}$ with reproducing kernel k and $\mathcal{H}^d$ be its product. We define the manifold kernel Stein discrepancies (mKSD) of the first, second and zeroth order by

$$\begin{aligned}
\mathrm{mKSD}^{(1)}(p\|q) &= \sup_{\|\mathbf{f}\|_{\mathcal{H}^d} \leq 1} \mathbb{E}_p[\mathcal{A}_q^{(1)} \mathbf{f}], \\
\mathrm{mKSD}^{(2)}(p\|q) &= \sup_{\|\tilde{f}\|_{\mathcal{H}} \leq 1} \mathbb{E}_p[\mathcal{A}_q^{(2)} \tilde{f}], \\
\mathrm{mKSD}^{(0)}(p\|q) &= \sup_{\|h\|_{\mathcal{H}} \leq 1} \mathbb{E}_p[\mathcal{A}_q^{(0)} h],
\end{aligned}$$

respectively. We also define the Stein kernels[6] of first, second and zeroth order by

$$h_q^{(1)}(x,\tilde{x}) = \left\langle \mathcal{A}_q^{(1)} k(x,\cdot), \mathcal{A}_q^{(1)} k(\tilde{x},\cdot) \right\rangle_{\mathcal{H}^d},$$
$$h_q^{(2)}(x,\tilde{x}) = \left\langle \mathcal{A}_q^{(2)} k(x,\cdot), \mathcal{A}_q^{(2)} k(\tilde{x},\cdot) \right\rangle_{\mathcal{H}},$$
$$h_q^{(0)}(x,\tilde{x}) = \left\langle \mathcal{A}_q^{(0)} k(x,\cdot), \mathcal{A}_q^{(0)} k(\tilde{x},\cdot) \right\rangle_{\mathcal{H}},$$

respectively. Then, by algebraic manipulation through reproducing property and taking the supremum over unit ball RKHS similar to Eq.(2.23), we obtain the following.

Theorem 3.4. *If p and q are smooth densities on $\mathcal{M}$ and the reproducing kernel k of $\mathcal{H}$ is smooth, then for $c = 0, 1, 2$,*

$$\mathrm{mKSD}^{(c)}(p\|q)^2 = \mathbb{E}_{x,\tilde{x}}[h_q^{(c)}(x,\tilde{x})]. \tag{3.15}$$

From Theorem 3.4, we can estimate mKSD by using samples from p. This is an important property in goodness-of-fit testing. Detailed derivations are shown in Appendix 3.A. The study the conditions that mKSD is a proper discrepancy measure between distributions on Riemannian manifolds, we consider cases $c = 0, 1, 2$. First we consider score function for density ratio of the form $L(x) = (L_1(x), \ldots, L_d)^\top \in \mathbb{R}^d$ with

$$L_i(x) = \frac{\partial}{\partial \theta^i} \log \frac{q(x)}{p(x)}.$$

Then we consider the universality notions for RKHS functions adapted from Carmeli et al. [2010]. Denote $\mathcal{C}(\mathcal{X};\mathbb{R})$ as the space of continuous functions with compact-open topology and $\mathcal{C}_0(\mathcal{X};\mathbb{R})$ as the continuous functions vanishing at infinty with uniform norm (otherwise called infinity norm) $\|f\|_\infty = \max_{x\in\mathcal{X}} \|f(x)\|$. k is Mercer kernel provided that $\mathcal{H}_k$ is a subset of $\mathcal{C}(\mathcal{X};\mathbb{R})$; k is c_0-kernel provided that $\mathcal{H}_k$ is a subset of $\mathcal{C}_0(\mathcal{X};\mathbb{R})$.

Definition 3.1 (Carmeli et al. [2010] Definition 4.2)**.** *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a reproducing kernel for $\mathcal{H}_k$.*

*(i) A c_0-kernel is called **universal** if $\mathcal{H}_k$ is dense in $L^2(X,\mu,\mathbb{R})$ for each probability measure μ.*

*(ii) A Mercer kernel is called **compact-universal** if $\mathcal{H}_k$ is dense in $L^2(X,\mu,\mathbb{R})$ for each probability measure μ with compact support.*

[6] As noted in Chapter 2, we distinguish the Stein kernel defined here h_q from RKHS kernel k. We also note that as h_q depends on both q and k, the boundedness of k does not guarantee the boundedness of h_q.

Theorem 3.5. *Let p and q be smooth densities on $\mathcal{M}$. Assume: 1) $\partial\mathcal{M}$ is empty and kernel k is compact universal in the sense of* [Carmeli et al., 2010, Definition 2 (ii)]*; 2) $\mathbb{E}_{x,\tilde{x}\sim p}[h_q^{(c)}(x,\tilde{x})^2] < \infty$, for $c = 0, 1, 2$; 3) $\mathbb{E}_p\|L(x)\|^2 < \infty$. Then, $\mathrm{mKSD}^{(c)}(p\|q) \geq 0$ and $\mathrm{mKSD}^{(c)}(p\|q) = 0$ if and only if $p = q$.*

Manifold with non-empty boundaries

The characterisation in Theorem 3.5 requires both Stein's identity (for forward direction) and universality of RKHS (for backward direction). The Stein's identity relies on, either vanishing boundary or $f(\partial\mathcal{M}) = 0, f \in \mathcal{H}$. Theorem 3.5 requires $\partial\mathcal{M}$ to be empty for both conditions to hold.

For manifold with non-empty boundary, $f(\partial\mathcal{M}) = 0, f \in \mathcal{H}$ is required for Stein's identity to hold. However, [Barp et al., 2018, Theorem 3] shows that the space is dense requires the function vanishes nowhere. Hence, $f(\partial\mathcal{M}) = 0, f \in \mathcal{H}$ and universality of $\mathcal{H}$ may not satisfy simultaneously. As such, we may consider functions that is (near-)universal with some form of approximation.

For $\epsilon > 0$, we define an ϵ-neighbourhood of the boundary $B_\epsilon(\partial\mathcal{M}) = \{x : x \in \mathcal{M}, \exists y \in \partial\mathcal{M}, d(x,y) < \epsilon\}$. Consider an approximation function g_ϵ, s.t. $g_\epsilon(x) = 1, \forall x \in \mathcal{M}\backslash B_\epsilon(\partial\mathcal{M})$ and "gradually vanishes" to 0 on $B_\epsilon(\partial\mathcal{M})$: one such "gradually vanishing" function can be

$$g_\epsilon(x) = \frac{1}{\epsilon}\min d(x,y), y \in \partial\mathcal{M}, x \in B_\epsilon(\partial\mathcal{M}). \tag{3.16}$$

Consider a compact universal kernel k and an approximation function g_ϵ, we can construct the product kernel $k_\epsilon(x,\cdot) = g_\epsilon(x)k(x,\cdot)$ that is still positive definite. The kernel function k_ϵ is getting closer to compact universal kernel k when $\epsilon \to 0$, which can be useful in practice to construct kernels and mKSD that effectively distinguish two distributions. We illustrate this with an example on disc in Appendix 3.E.

Remarks on mKSD

The mKSDs are build for testing goodness-of-fit of densities on Riemannian manifolds, while each mKSD has its advantages by construction.

Advantage of $\mathrm{mKSD}^{(1)}$ over $\mathrm{mKSD}^{(2)}$ While $\mathrm{mKSD}^{(2)}$ involves optimisation over $\tilde{f} \in \mathcal{H}$, $\mathrm{mKSD}^{(1)}$ optimises over $\mathbf{f} \in \mathcal{H}^d$. Thus, $\mathrm{mKSD}^{(1)}$ is expected to be more flexible and powerful in distinguishing between distributions than $\mathrm{mKSD}^{(2)}$. We will confirm that this is true in Section 3.4.3.

Equivalence of $\mathrm{mKSD}^{(0)}$ and MMD Remind that, for a RKHS $\mathcal{H}$, the maximum mean discrepancy (MMD) [Gretton et al., 2012a] between p and q is defined by

$$\mathrm{MMD}^2(p\|q) = \|\mu_p - \mu_q\|_{\mathcal{H}}^2,$$

where μ_p, μ_q are the kernel mean embedding [Muandet et al., 2017] of p and q, respectively. The following theorem shows that $\mathrm{mKSD}^{(0)}$ is equivalent to MMD.

Theorem 3.6. *If p and q are densities on $\mathcal{M}$ and the reproducing kernel k is compact universal as in Theorem 3.5, then*

$$\mathrm{mKSD}^{(0)}(p\|q) = \mathrm{MMD}(p\|q).$$

Proof. By definition, we have

$$\begin{aligned}\mathrm{mKSD}^{(0)}(p\|q) &= \sup_{\|h\|_{\mathcal{H}}\leq 1} \mathbb{E}_p[\mathcal{A}_q^{(0)}h] \\ &= \sup_{\|h\|_{\mathcal{H}}\leq 1} (\mathbb{E}_p[h] - \mathbb{E}_q[h]).\end{aligned}$$

Hence, taking the supremum in closed form via reproducing property, we obtain

$$\mathrm{mKSD}^{(0)}(p\|q)^2 = \|\mu_p - \mu_q\|_{\mathcal{H}}^2 = \mathrm{MMD}^2(p\|q).$$

All quantities are well-defined due to bounded kernel k. □

The derivation is coordinate invariant as taking the expectation $\mathbb{E}_p$ and $\mathbb{E}_q$ already take the geometry of the manifold into account by integrating over relevant coordinate measures on the underlying space, $\mathcal{M}$. As such, the equivalence relationship is also applicable for KSD in the Euclidean case in Eq.(2.22), $\mathrm{KSD}^{(0)}(p\|q)^2 = \mathrm{MMD}^2(p\|q)$ for densities p and q on $\mathbb{R}^d$.

Moreover, the two quantities has coinciding assumptions on kernels: from KSD perspective with universality based assumptions and from MMD perspective with characteristic kernels. [Carmeli et al., 2010] shows the connections between compact universal kernel and c_0-universal kernel. [Sriperumbudur et al., 2011; Fukumizu et al., 2009] show the equivalence relationship of c_0-universal kernel being characteristic.

The MMD statistic used in two sample problem is also discussed in goodness-of-fit context [Jitkrittum et al., 2017; Yang et al., 2019], where the samples are simulated from the null and the test is conducted by checking whether the observed samples and the simulated are from the same distribution. Theorem 3.6 further uni-

fies such procedure under the KSD framework. While empirical results [Jitkrittum et al., 2017; Yang et al., 2019; Xu and Matsuda, 2020] have shown the empirical test performances between the KSD and MMD test statistics in the context of goodness-of-fit testing, in the next section, we provide additional theoretical analysis regarding the test efficiencies.

3.4.2 Goodness-of-fit Tests with mKSDs

Here, we present a procedure for testing $H_0 : p = q$ with significance level α based on samples $x_1, \ldots, x_n \sim p$.

From Theorem 3.4, an unbiased estimate of mKSD can be obtained in the form of U-statistics [Lee, 1990]:

$$\mathrm{mKSD}_u^{(c)}(p\|q)^2 = \frac{1}{n(n-1)} \sum_{i \neq j} h_q^{(c)}(x_i, x_j). \tag{3.17}$$

Its asymptotic distribution is obtained via U-statistics theory [Lee, 1990; Van der Vaart, 2000] as follows. We denote the convergence in distribution by $\xrightarrow{\mathrm{d}}$.

Theorem 3.7. *For $c = 0, 1, 2$, the following statements hold.*

1. *Under $H_0 : p = q$, the asymptotic distribution of $\mathrm{mKSD}_u^{(c)}(p\|q)^2$ is*

$$n \cdot \mathrm{mKSD}_u^{(c)}(p\|q)^2 = \sum_{j=1}^{\infty} w_j^{(c)} (Z_j^2 - 1), \tag{3.18}$$

   where Z_j are i.i.d. standard Gaussian random variables and $w_j^{(c)}$ are the eigenvalues of the Stein kernel $h_q^{(c)}(x, \tilde{x})$ under $p(\tilde{x})$:

$$\int h_q^{(c)}(x, \tilde{x}) \phi_j(\tilde{x}) p(\tilde{x}) \mathrm{d}\tilde{x} = w_j^{(c)} \phi_j(x),$$

   where $\phi_j(x) \not\equiv 0$.

2. *Under $H_1 : p \neq q$, the asymptotic distribution of $\mathrm{mKSD}_u^{(c)}(p\|q)^2$ is*

$$\sqrt{n} \cdot \left(\mathrm{mKSD}_u^{(c)}(p\|q)^2 - \mathrm{mKSD}^{(c)}(p\|q)^2 \right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, {\sigma_c}^2),$$

   where ${\sigma_c}^2 = \mathrm{Var}_{x \sim p}[\mathbb{E}_{\tilde{x} \sim p}[h_q^{(c)}(x, \tilde{x})]] > 0$.

We employ Theorem 3.7 for goodness-of-fit testing with U-statistic. Namely, we generate bootstrap samples from an approximation of the null distribution Eq. (3.18) of $n \cdot \mathrm{mKSD}_u^{(c)}(p\|q)^2$ and compare their $(1-\alpha)$ quantile with the statistics

Algorithm 1 mKSD test via U-statistics (mKSDu)

Input:

samples $x_1, \ldots, x_n \sim p$, null density q (metric tensor G if required)
kernel function k, test size α, bootstrap sample size B

Objective:

Test $H_0 : p = q$ versus $H_1 : p \neq q$.

Test procedure:

1: Compute the U-statistics $\mathrm{mKSD}_u^{(c)}(p, q)^2$ via (3.17).
2: Compute eigenvalues $\widehat{\omega}_1, \ldots, \widehat{\omega}_n$ of $n \times n$ matrix H, where $H_{ij} = h_q^{(c)}(x_i, x_j)$.
3: **for** $t = 1 : B$ **do**
4: Sample $Z_1, \ldots, Z_n \sim \mathcal{N}(0, 1)$ independently.
5: Compute $S_t = \sum_{j=1}^n \widehat{\omega}_j (Z_j^2 - 1)$.
6: **end for**
7: Determine the $(1 - \alpha)$-quantile $\gamma_{1-\alpha}$ of $S_1, \ldots, S_B$.

Output:

Reject H_0 if $n \cdot \mathrm{mKSD}_u^2(p, q) > \gamma_{1-\alpha}$; otherwise do not reject.

$n \cdot \mathrm{mKSD}_u^{(c)}(p\|q)^2$. To approximate the null distribution in Eq. (3.18), we truncate the infinite sum in Eq. (3.18) following [Gretton et al., 2009a]: $\sum_{j=1}^n \widehat{\omega}_j (Z_j^2 - 1)$, where $\widehat{\omega}_j$ are eigenvalues of the $n \times n$ Stein kernel matrix H with $H_{ij} = h_q^{(c)}(x_i, x_j)$ and $Z_1, \ldots, Z_n$ are independent standard Gaussian random variables. The testing procedure is outlined in Algorithm 1.

For an efficient implementation of the test, we also consider the following V-statistics[7]

$$\mathrm{mKSD}_v^{(c)}(p\|q)^2 = \frac{1}{n^2} \sum_{i,j} h_q^{(c)}(x_i, x_j), \tag{3.19}$$

and adopt the wild bootstrap method [Chwialkowski et al., 2014, 2016] for testing. Instead of simulating the null distribution from the estimated asymptotic limit distribution in Eq. (3.18), the simulated null distribution is generated from weighted re-sampling of the Stein kernel matrix. The wild bootstrap method is not only useful for building a consistent test with non-independent samples [Leucht and Neumann, 2013] and naive permutation or bootstrap procedures fails for kernel-based test statistics [Chwialkowski et al., 2014], it is also useful computationally efficient to implement. Specifically, for each $t = 1, \ldots, B$, we sample uniform i.i.d. variables $U_1, \ldots, U_n \sim \mathrm{U}[0, 1]$, let $W_{0,t} = 1$ and define

$$W_{i,t} = \mathbb{1}_{\{U_i > a_t\}} W_{i-1,t} - \mathbb{1}_{\{U_i < a_t\}} W_{i-1,t}, \tag{3.20}$$

[7] The V-statistic is a biased estimate of the mKSD.

Algorithm 2 mKSD test via wild bootstrap

Input:

samples $x_1, \ldots, x_n \sim p$, null density q, (metric tensor G if required)
kernel function k, test size α, bootstrap size B

Objective:

Test $H_0 : p = q$ versus $H_1 : p \neq q$.

Test procedure:

1: Compute the statistic $\text{mKSD}_v^{(c)}(p\|q)^2$, Eq.(3.19).
2: **for** $t = 1 : B$ **do**
3: Sample $W_{1,t}, \ldots, W_{n,t}$ via Eq.(3.20).
4: Compute S_t by Eq.(3.21).
5: **end for**
6: Determine the $(1-\alpha)$-quantile $\gamma_{1-\alpha}$ of $S_1, \ldots, S_B$.

Output:

Reject H_0 if $\text{mKSD}_v^{(c)}(p\|q)^2 > \gamma_{1-\alpha}$; otherwise do not reject.

for $i = 1, \ldots, n$, where $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function and a_t is the probability of sign change, which is referred to as *wild bootstrap process*. This is particularly useful to incorporate dependencies in the generated samples [Chwialkowski et al., 2014], e.g. sampling from MCMC procedures. When $x_1, \ldots, x_n$ are independent, a_t is set to 0.5, which correspond to independent Radamacher variables. As such, the wild bootstrap samples are given by

$$S_t = \frac{1}{n^2} \sum_{i,j} W_{i,t} W_{j,t} h(x_i, x_j), \quad t = 1, \ldots, n. \tag{3.21}$$

We reject the null if the test statistic $\text{mKSD}_v^{(c)}(p\|q)^2$ in Eq.(3.17) exceeds the $(1-\alpha)$-quantile of $\{S_1, \ldots, S_B\}$. The testing procedure is outlined in Algorithm 2.

Kernel choice The performance of kernel-based testing is sensitive to the choice of kernel parameters. We choose the kernel parameters by maximising an approximation of the test power following [Gretton et al., 2012b; Jitkrittum et al., 2016a; Sutherland et al., 2016]. From Theorem 3.7,

$$D := \sqrt{n} \cdot \frac{\text{mKSD}_u^{(c)}(p\|q)^2 - \text{mKSD}^{(c)}(p\|q)^2}{\sigma_c} \xrightarrow{\text{d}} \mathcal{N}(0, 1)$$

under the alternative $H_1 : p \neq q$. Thus, for sufficiently large n, the test power is approximated as

$$\begin{aligned}\Pr_{H_1}(n \cdot \mathrm{mKSD}_u^{(c)}(p\|q)^2 > r) =& \Pr_{H_1}\left(D > \frac{r}{\sqrt{n}\sigma_c} - \sqrt{n}\frac{\mathrm{mKSD}^{(c)}(p\|q)^2}{\sigma_c}\right) \\ \approx& 1 - \Phi\left(\frac{r}{\sqrt{n}\sigma_c} - \sqrt{n}\frac{\mathrm{mKSD}^{(c)}(p\|q)^2}{\sigma_c}\right) \\ \approx& \Phi\left(\sqrt{n}\frac{\mathrm{mKSD}^{(c)}(p\|q)^2}{\sigma_c}\right),\end{aligned}$$

where Φ denotes the cumulative distribution function of the standard normal distribution and we used the approximation [Sutherland et al., 2016]

$$\frac{r}{\sqrt{n}\sigma_c} - \sqrt{n}\frac{\mathrm{mKSD}^{(c)}(p\|q)^2}{\sigma_c} \approx -\sqrt{n}\frac{\mathrm{mKSD}^{(c)}(p\|q)^2}{\sigma_c} \tag{3.22}$$

for sufficiently large n. Thus, we choose the kernel parameters by maximising an estimate of $\frac{\mathrm{mKSD}^{(c)}(p\|q)^2}{\sigma_c}$ [Jitkrittum et al., 2016a, 2017].

Such kernel choice strategies focus on a given class of kernel, e.g. squared exponential kernel to optimise with the kernel parameters. Such function class can be more flexible, e.g. kernel with Matérn form that utilises the modified Bessel function as well as distance metric in space generalises useful kernels such as Laplace form of kernel and Gaussian form. The extension of kernel choices into these larger class of RKHS kernel are interesting future directions. Despite kernels with Matérn form are still translational invariant, Chapter 6 address flexibility of kernel choice by exploring translational non-invariant kernels parametrised by deep neural networks with applications on two-sample problems.

3.4.3 Comparisons between mKSD Tests

Bahadur Efficiency From Theorem 3.7, mKSD tests are consistent against all alternative distributions q satisfying Theorem 3.5. Thus, to understand which mKSD test is more powerful than others, we investigated their *Bahadur efficiency* [Bahadur et al., 1960], which quantify how fast the p-value goes to zero under alternatives. Here, to focus on the effect of the choice of Stein operator on test performance, we briefly present results for testing of uniformity on the circle $\mathbb{S}^1$ under the von-Mises distribution. The technique of the proof is adapted from Jitkrittum et al. [2017]. More details on Bahadur efficiency are discussed in Appendix 3.B.

Approximate Bahadur Slope (ABS) We first define Bahadur slope for general tests [Gleser, 1966] and its applications in kernel-based tests [Jitkrittum et al., 2017; Garreau et al., 2017]. Consider the test procedure with null hypothesis $H_0 : \omega \in \Omega_0$ and the alternative $H_1 : \omega \in \Omega \backslash \Omega_0$, where Ω and Ω_0 are arbitrary sets. Denote T_n as the test statistic computed from a sample of size n.

Definition 3.2. *For $\omega_0 \in \Omega_0$, let F be the asymptotic null distribution*

$$F(t) = \lim_{n\to\infty} \mathbb{P}_{\omega_0}(T_n < t)$$

which is assumed to be continuous and common $\forall \omega_0 \in \Omega_0$. Assume that there exists a continuous strictly increasing function $\rho : (0, \infty) \to (0, \infty)$ s.t $\lim_{n\to\infty} \rho(n) = \infty$. Denote

$$c(\omega) = -2 \operatorname{plim}_{n\to\infty} \frac{\log(1 - F(T_n))}{\rho(n)}, \tag{3.23}$$

for some bounded non-negative function c such that $c(\omega_0) = 0$ when $\omega_0 \in \Omega_0$. The function $c(\omega)$ is known as approximate Bahadur slope.

Asymptotic Relative Efficiency (ARE) Between Tests with Different $\mathcal{A}_q$ ARE between two statistical testing procedures measures how fast the p-values of one test shrinks to 0, relatively to the other's. If it is faster, for given problem under the alternative, it is more sensitive to pick up the alternative, where we call the test more "statistically efficient". With approximate Bahadur slope (ABS), we are ready to define approximate Bahadur efficiency.

Definition 3.3. *Given two sequences of test statistics, $T_n^{(1)}$ and $T_n^{(2)}$ and their ABS $c^{(1)}$ and $c^{(2)}$, the approximate Bahadur efficiency of $T_n^{(1)}$ relative to $T_n^{(2)}$ is*

$$\mathrm{E}(\omega_A) := \frac{c^{(1)}(\omega_A)}{c^{(2)}(\omega_A)} \tag{3.24}$$

for $\omega_A \in \Omega \backslash \Omega_0$, in the space of alternative models.

If $\mathrm{E}(\omega_\mathrm{A}) > 1$, then $T_n^{(1)}$ is asymptotically more efficient than $T_n^{(2)}$ in the sense of Bahadur, for the particular problem specified by $\omega_A \in \Omega \backslash \Omega_0$.

Theorem 3.8. *(Scaling shift in von-Mises distribution) Let $x \in \mathbb{S}^1$, $q(x) \propto 1$ and $p(x) \propto \exp(\kappa u^\top x)$. Choose the von-Mises kernel of the form $k(x, x') = \exp(x^\top x')$. Denote the approximate Bahadur efficiency between* mKSD *with first*

and second order Stein operators as

$$\mathrm{E}_{1,2}(\kappa) := \frac{c^{(\mathrm{mKSD}^{(1)})}(\kappa)}{c^{(\mathrm{mKSD}^{(2)})}(\kappa)},$$

where $\kappa > 0$. *For non-trivial problems, i.e. small* κ, $0 < \kappa \leq 10$, $\mathrm{E}_{1,2}(\kappa) > 1$.

Adapting [Jitkrittum et al., 2017, Theorem 5], it suffices to show $\mathrm{mKSD}^{(1)}(p\|q) \geq \mathrm{mKSD}^{(2)}(p\|q)$ and

$$\mathbb{E}_{x,\tilde{x}\sim q}[h_q^{(2)}(x,\tilde{x})^2] > \mathbb{E}_{x,\tilde{x}\sim q}[h_q^{(1)}(x,\tilde{x})^2] > 0.$$

Detailed derivations are shown in the supplementary material.

We provide additional discussion on test efficiencies with $\mathrm{mKSD}^{(0)}$ in the supplementary material. In general, since we cannot compute $\mathbb{E}_p$ in closed form, especially with unnormalised density, we need to perform the test with samples, where sampling error makes the $\mathrm{mKSD}^{(0)}$ test less asymptotically efficient [Jitkrittum et al., 2017; Yang et al., 2019; Xu and Matsuda, 2020].

Computational efficiency Since the Stein kernels $h_q^{(1)}$ and $h_q^{(2)}$ depend on q only through the derivative of $\log q$, mKSD tests with the first and second order Stein operators do not require computation of the normalisation constant of q. This is a major computational advantage over existing goodness-of-fit tests on Riemannian manifolds. While the computational cost of $\mathrm{mKSD}_u^{(1)}$ is $O(n^2 d)$, that of $\mathrm{mKSD}^{(2)}$ is $O(n^2 d^3)$ due to the computation of the metric tensor.

On the other hand, mKSD test of zeroth order is equivalent to testing whether two sets of samples are from the same distribution by using MMD [Gretton et al., 2012a][8]. Namely, to test whether $x_1, \ldots, x_n$ is from density q, we draw samples $y_1, \ldots, y_m$ from q and determine whether $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ are from the same distribution. This procedure requires to sample from the null distribution q on Riemannian manifolds, which is computationally intensive in general. Note that the results in Theorem 3.7 with $c = 0$ replicate the asymptotic results for MMD [Gretton et al., 2012a].

Choosing mKSD tests Overall, $\mathrm{mKSD}^{(1)}$ has its advantage in terms of having a large space of test functions with both asymptotic test efficiency and computational efficiency so that it is recommended to use when available. $\mathrm{mKSD}^{(2)}$ can be slightly easier to derive and parametrise in particular scenarios, although it sacrifice test power and computational efficiency. $\mathrm{mKSD}^{(0)}$, or namely MMD, is also applica-

[8] This procedure is sometimes referred to as the pseudo goodness-of-fit test.

ble when it is possible to sample from the given unnormalised density model on Riemannian manifolds.

3.4.4 Model Criticism on Manifold

When the proposed model does not fit the observed data well, understanding which part of the model misfit the data is of practical interest. The model criticism with kernel non-parametric tests has been studied for MMD [Sutherland et al., 2016] and KSD [Jitkrittum et al., 2017]. Here, we propose model criticism methods based on mKSD1.

Let $\mathbf{s}_p(\cdot) = \mathbb{E}_{\tilde{x}\sim p}[\mathcal{A}_q^{(1)} k(\tilde{x}, \cdot)] \in \mathcal{H}^d$. We define the manifold Finite Set Stein Discrepancy (mFSSD) adapted from [Jitkrittum et al., 2017] by

$$\mathrm{mFSSD}^2 = \frac{1}{dJ} \sum_{i=1}^{d} \sum_{j=1}^{J} (\mathbf{s}_p(v_j))_i^2, \tag{3.25}$$

which can be computed in linear time of sample size n. Stein identity of $\mathbf{s}_p(\cdot)$ ensures $\mathrm{mFSSD}^2 = 0$ under H_0 with probability 1 [Jitkrittum et al., 2017, Theorem 1]. To perform model criticism, we extract some test locations that give a higher detection rate (i.e., test power) than others. We choose the test locations $V = \{v_j\}_{j=1}^J$ by maximising the approximate test power:

$$V = \arg\max_{v} \frac{\mathrm{mFSSD}^2}{\tilde{\sigma}_{H_1}},$$

where $\tilde{\sigma}_{H_1}$ is the variance of mFSSD^2 under H_1.

Asymptotics for mFSSD To compute the empirical version of mFSSD, we consider the empirical version $\mathbf{s}_p(\cdot)$ in Eq.(3.25) from samples $x_1, \ldots, x_n \sim p$:

$$\widehat{\mathbf{s}}_p(\cdot) = \frac{1}{n} \sum_{i} [\mathcal{A}_q^{(1)} k(x_i, \cdot)].$$

Then the empirical mFSSD has the form

$$\widehat{\mathrm{mFSSD}^2} = \frac{1}{dJ} \sum_{i=1}^{d} \sum_{j=1}^{J} (\widehat{\mathbf{s}}_p(v_j))_i^2, \tag{3.26}$$

for any set of test locations $\{v_j\}_{j=1}^J$.

Proposition 3.1. *Assume the conditions in Theorem 3.5 hold, and* $\mathbb{E}_{x\sim p}[\|\mathbf{s}_p(x)\|^2] <$

∞. *Under* $H_1 : p \neq q$,

$$\sqrt{n} \cdot \left(\widehat{\mathrm{mFSSD}}^2 - \mathrm{mFSSD}^2 \right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \tilde{\sigma}^2_{H_1}),$$

where $\tilde{\sigma}^2_{H_1}$ *denotes the asymptotic variance for* $\widehat{\mathrm{mFSSD}}^2$.

Proof. With the assumed regularity conditions, Eq.(3.26) is in the form of the non-degenerate U-statistics with $\tilde{\sigma}^2_{H_1} > 0$. The asymptotic normality follows from [Serfling, 2009, Section 5.5.1], similarly described in [Jitkrittum et al., 2017, Proposition 2]. □

The asymptotic normality for $\widehat{\mathrm{mFSSD}}^2$ in Proposition 3.1 enables derivation of the approximate test power, which is similarly to the objective for choosing kernel parameters as described in Section 3.4.2.

Proposition 3.2. [Approximate test power of $n \cdot \widehat{\mathrm{mFSSD}}^2$] *Under* H_1, *for large* n *and fixed* r, *the test power is*

$$\mathbb{P}_{H_1}(n \cdot \widehat{\mathrm{mFSSD}}^2 > r) \approx 1 - \Phi\left(\frac{r}{\sqrt{n}\tilde{\sigma}^2_{H_1}} - \sqrt{n}\frac{\mathrm{mFSSD}^2}{\tilde{\sigma}^2_{H_1}} \right),$$

where Φ *denotes the cumulative distribution function of the standard normal distribution, and* $\tilde{\sigma}^2_{H_1}$ *is defined in Proposition 3.1.*

Due to $\sqrt{n}$ scaling in Proposition 3.1, maximising the approximate test power for $n \cdot \widehat{\mathrm{mFSSD}}^2$ can be approximated by maximising $\frac{\mathrm{mFSSD}^2}{\tilde{\sigma}^2_{H_1}}$ to obtain optimal test locations, $V = \{v_j\}_{j=1}^J$, under the alternative $H_1 : p \neq q$,

3.5 Simulation Results

3.5.1 Goodness-of-fit Tests for Directional Distributions

First, we validate via simulations the proposed first order mKSD tests for directional distributions, i.e. $\mathcal{M} = \mathbb{S}^{d-1}$, focusing on the U-statistic test in Algorithm 1 and V-statistic test in Algorithm 2. We denote the $\mathrm{mKSD}^{(1)}$ in the context of directional distribution as directional kernel Stein discrepancy (dKSD). We refer dKSD_u and dKSD_v as the testing procedure in Algorithm 1 and Algorithm 2 respectively.

We employ the von-Mises kernel of the form $k(x, x') = \exp(\gamma x^\top x')$, which is positive definite [Gneiting et al., 2013], for both the dKSD tests and MMD-based testing procedures which corresponds to $\mathrm{mKSD}^{(0)}$ as shown in the previous section. The von-Mises kernel is compact universal. Rewriting the kernel in the form analogous to the Gaussian kernel: $k(x, y) = \exp(\gamma x^\top x') = C \cdot \exp(-\frac{1}{2}\gamma \cdot \|x - x'\|^2)$,

n	Rayleigh	Kuiper	dKSDu	dKSDv	MMD
30	0.138	0.128	**0.560**	0.338	0.133
50	0.308	0.267	0.750	**0.898**	0.317
100	0.712	0.667	0.820	**1.0**	0.583
200	0.980	0.962	0.900	**1.0**	0.900

Table 3.1: Rejection rates for the circular uniform distribution under the von-Mises distribution with $\kappa = 0.5$ as alternative, with test size $\alpha = 0.01$.

n	Rayleigh	Kuiper	dKSDu	dKSDv	MMD
30	0.757	0.731	0.650	**0.831**	0.600
50	0.957	0.940	0.750	**1.0**	0.833
100	**1.0**	**1.0**	0.833	**1.0**	0.983
200	**1.0**	**1.0**	0.96	**1.0**	**1.0**

Table 3.2: Rejection rates for the circular uniform distribution under the von-Mises distribution with $\kappa = 1$ as alternative, with test size $\alpha = 0.01$.

where C is a constant due to hypersphere structure. Gaussian kernel is universal [Sriperumbudur et al., 2011] and the hypersphere is a compact subset of the space of $\mathbb{R}^d$ matrices, the von-Mises kernel is then also compact-universal from Corollary 3 of Carmeli et al. [2010]. The bootstrap sample size is set to $B = 1000$. The significance level is set to $\alpha = 0.01$. In MMD-based test, we set $m = n$.

Circular uniform distribution

We start by considering the circular ($d = 2$) uniform distribution, for which the goodness-of-fit tests have been proposed such as Rayleigh test and Kuiper test [Mardia and Jupp, 1999]. See supplementary material for details of Rayleigh test and Kuiper test. We compare the proposed dKSD tests with these existing tests as well as MMD-based test. We repeated 600 trials to calculate rejection rates.

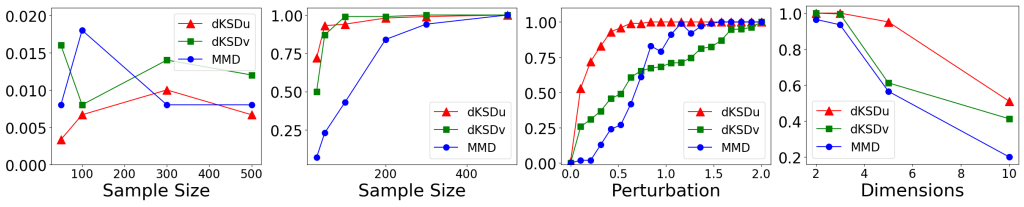
Tables 3.1 and 3.2 present the rejection rate under the von-Mises distribution alternative with concentration parameter $\kappa = 0.5$ and $\kappa = 1$, respectively. The power of all tests increases with increasing n or κ and converges to one. In overall, we observe that the dKSD_v has the highest power, especially in harder problems with small κ and n. Table 3.3 presents the rejection rate under the null and the type-I errors of all tests are well controlled to the test significance level $\alpha = 0.01$.

von-Mises-Fisher distribution on $\mathbb{S}^{d-1}$

Next, we consider testing the von-Mises-Fisher distribution $\mathrm{vMF}(\mu, \kappa)$ in Eq. (3.3). von-Mises-Fisher distribution is commonly view as directional distribution coun-

n	Rayleigh	Kuiper	dKSDu	dKSDv	MMD
30	0.006	0.010	0.011	0.007	0.013
50	0.015	0.011	0.015	0.015	0.016
100	0.010	0.011	0.008	0.011	0.030
200	0.015	0.018	0.010	0.015	0.013

Table 3.3: Type-I error of tests for the circular uniform distribution



(a) $d = 3, \sigma = 0$ (b) $d = 3, \sigma = 1$ (c) $d = 3, n = 200$ (d) $n = 200, \mu = d^{-\frac{1}{2}}\mathbf{1}_d$

Figure 3.2: Rejection rates for von-Mises-Fisher Distributions on $\mathbb{S}^{d-1}$

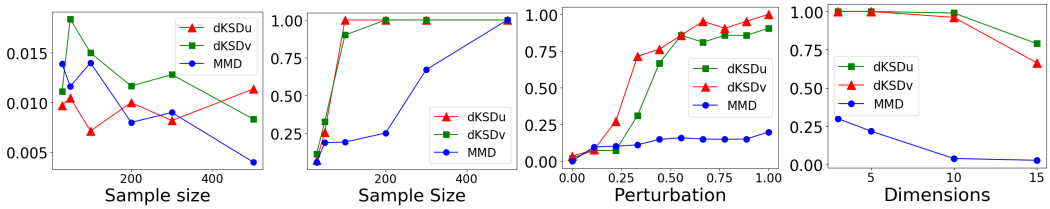
terpart of Gaussian distribution in $\mathbb{R}^d$. It has mean parameter $\mu \in \mathbb{S}^{d-1}$ and concentration parameter κ.

We set the null distributions as $\mathrm{vMF}(\mu_0, 1)$ and the alternative distribution by perturbing the mean location μ and concentration scaling κ, $\mathrm{vMF}(\mu, 1 + \sigma)$ where $\mu_0 = (1, 0, \ldots, 0) \in \mathbb{S}^{d-1}$. $\mu \in \mathbb{S}^{d-1}$ and $\sigma \geq 0$. We generated samples from the von-Mises-Fisher distribution by using the rejection methods[Jakob, 2012; Wood, 1994]. We compare the proposed dKSD tests with MMD-based test in Figure 3.2.

Figure 3.2(a) plots the rejection rate under the null ($\mu = \mu_0$, $\sigma = 0$) with respect to n for $d = 3$. The type-I errors of dKSD tests are well controlled to the significance level $\alpha = 0.01$. Figure 3.2(b) plots the rejection rate with respect to n for $d = 3$, $\mu = \mu_0$ and $\sigma = 1$. Both dKSD_u and dKSD_v have larger power than MMD-based test. Figure 3.2(c) plots the rejection rate with respect to σ for $d = 3$, $n = 200$ and $\mu = \mu_0$. The dKSD achieves maximal test power of 1 as κ increases. Figure 3.2(d) plots the rejection rate with respect to the hypersphere dimension d for $n = 200$. The alternative is set to be $\mu = (1/\sqrt{d})\mathbf{1}_d$ and $\sigma = 0.5$, where $\mathbf{1}_d$ denotes the all one vector. All test powers decrease for problems in higher dimension which is expected. The dKSD tests have larger power than MMD-based test in all dimensions.

Fisher-Bingham distribution on $\mathbb{S}^{d-1}$

In addition, we consider the Fisher-Bingham distribution in Eq. (3.4). Here, we focus on the Fisher-Bingham distribution $\mathrm{FB}(A)$ that only includes second order



(a) $d = 3, \sigma = 0$ (b) $d = 3, \sigma = 1$ (c) $d = 3, n = 200$ (d) $n = 200, \sigma = 1$

Figure 3.3: Rejection rates for Fisher-Bingham Distributions on $\mathbb{S}^{d-1}$

terms:

$$p(x \mid A) \propto \exp(x^\top A x), \quad x \in \mathbb{S}^{d-1},$$

where $A \in \mathbb{R}^{d \times d}$ is symmetric. The normalisation constant does not have closed form in general. We set the null distribution to $\mathrm{FB}(A)$ with

$$A_{ij} = \begin{cases} 2 & (i = j) \\ 1 & (i \neq j) \end{cases},$$

and the alternative distribution to $\mathrm{FB}(A')$ with $A' = A + \sigma \mathbf{1}_{d,d}$, where $\sigma \geq 0$ and $\mathbf{1}_{d,d}$ denotes the $d \times d$ matrix with all entries one. We generated samples from the Fisher-Bingham distribution via rejection sampling with angular central Gaussian proposals [Kent et al., 2013; Fallaize and Kypraios, 2016]. We compare the proposed dKSD tests with MMD-based test in Figure 3.3.

Figure 3.3(a) plots the rejection rate under the null ($\sigma = 0$) with respect to n for $d = 3$. The type-I errors of dKSD tests are controlled to the test significance level $\alpha = 0.01$. Figure 3.3(b) plots the rejection rate with respect to n for $d = 3$ and $\sigma = 1$. The dKSD tests achieves maximal test power as n increases and have higher power than MMD-based tests Figure 3.3(c) plots the rejection rate with respect to σ for $n = 200$ and $d = 3$. Again, the dKSD tests have larger power and capture small perturbation. Figure 3.3(d) plots the rejection rate with respect to d for $n = 200$ and $\sigma = 1$. The dKSD tests attain almost 80% power even when the dimension is as large as 15, whereas the power of the MMD-based test is smaller than 20% for all dimensions.

Computational runtime Due to the MMD-based tests requires generating samples on hypershperes from unnormalised densities which requires Monte Carlo sampling procedure, the computational time for the overall testing procedure becomes much longer. Table 3.4 presents the computational time for the tests based on an example of Fisher-Bingham distribution with $d = 3$. The dKSD tests are more computation-

n	30	50	100	200	300	500
dKSDu	0.005	0.011	0.027	0.096	0.227	0.588
dKSDv	0.009	0.015	0.030	0.105	0.238	0.574
MMD	0.091	0.120	0.180	0.379	0.704	2.614

Table 3.4: Computational time for Fisher-Bingham distributions (in seconds).

ally efficient than MMD-based test.

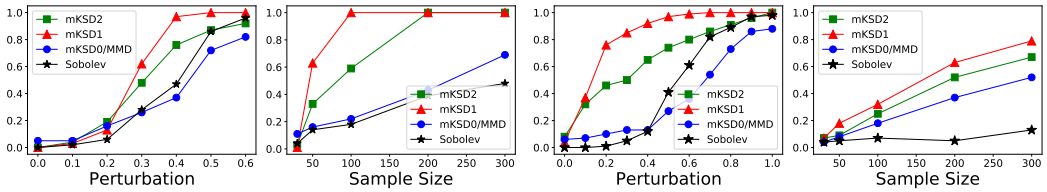
3.5.2 Goodness-of-fit Tests for Rotation Group

Then, we show the validity of the proposed mKSD tests by simulation on the rotation group $\mathrm{SO}(3)$. We use the Euler angle [Chikuse, 2012] as the coordinate system. The bootstrap sample size is set to $B = 1000$. The significance level is set to $\alpha = 0.01$. For the $\mathrm{mKSD}^{(0)}$ test (MMD-based test), the number of samples m draw from the null is set to be equal to the sample size n. We use the kernel for rotation group [Song et al., 2009] $k(X, Y) = \exp(\eta \cdot \mathrm{tr}(X^\top Y))$, where the parameter η was chosen by optimising the approximate test power. To see this, we rewrite the kernel in the form analogous to the Gaussian kernel: $k(X, Y) = \exp(\gamma \cdot \mathrm{tr}(X^\top Y)) = C \cdot \exp(-\frac{1}{2}\gamma \cdot \|X - Y\|_F^2)$, where C is a constant that only depends on d, the dimension of the matrices $X, Y \in \mathrm{SO(d)}$ due to $\mathrm{tr}(X^\top X) = \mathrm{tr}(I_d) = d$ for all $X \in \mathrm{SO(d)}$. Since the Gaussian kernel is universal [Sriperumbudur et al., 2011] and the rotation group $\mathrm{SO(d)}$ is a compact subset of the space of $d \times d$ matrices, the exponential-trace kernel is then also compact-universal from Corollary 3 of Carmeli et al. [2010]. We compare our tests with Sobolev type of tests in [Jupp et al., 2005].

Uniform distribution on SO(3)

First, we consider testing of uniformity on SO(3) and compare the performance of the mKSD tests with the Sobolev test [Jupp et al., 2005]. We generated samples from the exponential trace distribution $p(X \mid \kappa) \propto \exp(\kappa \cdot \mathrm{tr}(X))$ by the rejection sampling [Hoff, 2009]. The uniform distribution corresponds to $\kappa = 0$.

Figure 3.4 (a) plots the rejection rates with respect to κ for $n = 100$. When $\kappa = 0$, the type-I errors of all tests are well controlled to the significance level $\alpha = 0.01$. The power of all tests increases with increasing κ and converges to one. Figure 3.4 (b) plots the rejection rates with respect to n for $\kappa = 0.35$. The power of all tests increases with n and converges to one. When the model becomes increasingly different from the null, the $\mathrm{mKSD}^{(1)}$ is more sensitive to distinguish the difference, with higher power than others.



(a) $n = 100$ (b) $\kappa = 0.35$ (c) $n = 100$ (d) $b = 0.20$

Figure 3.4: Rejection rates at $\alpha = 0.01$: (a)-(b) for uniform density; (c)-(d) for Fisher distribution on SO(3)

Fisher distribution

Next, we consider the Fisher distribution (or matrix-Langevin distribution) of the unnormalised density $p(X \mid F) \propto \exp(\mathrm{tr}(FX))$ [Chikuse, 2003; Sei et al., 2013]. We generated data from $p(X \mid F_0)$ and applied mKSD tests on the null $p(X \mid F_b)$, where

$$F_b = \begin{pmatrix} 1 & b & 0 \\ b & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

We compare the mKSD tests with the extended Sobolev test [Jupp et al., 2005], in which the normalisation constant are computed via Monte Carlo estimation.

Figure 3.4(c) plots the rejection rates with respect to b for $n = 100$. Figure 3.4(d) plots the rejection rates with respect to n for $b = 0.2$. From the plot, we see that all tests achieves the correct test level under the null. When the model becomes increasingly different from the null, the $\mathrm{mKSD}^{(1)}$ is more sensitive to distinguish the difference, with higher power than others. $\mathrm{mKSD}^{(0)}$/MMD test has lower power than $\mathrm{mKSD}^{(1)}$ and $\mathrm{mKSD}^{(0)}$ due to inefficiency from sampling. While the Sobolev test is useful when the null and the alternative are very different, it is not powerful enough for harder problems where the alternative perturbed little from the null.

3.6 Real Data Applications

Finally, we apply the mKSD tests to two real data for testing goodness-of-fit and model criticism.

3.6.1 Vectorcardiogram data

As a real dataset on the rotation group SO(3), we use the vectorcardiogram data previously studied by Jupp et al. [2008]. The data summarises vectorcardiogram from normal children where each data point records 3 perpendicular vectors of directions

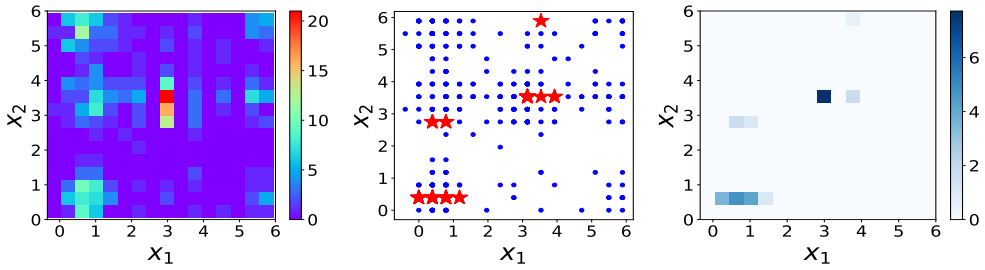


Figure 3.5: Wind direction data. Left: 2D histogram for wind directions; Mid: the 10 optimised locations, without repetition. Right: the objective value for data locations, the higher the darker.

QRS, PRS and T from Frank system for electrical lead placement. Details of this dataset can be found in [Downs, 1972].

Jupp et al. [2005] fitted the Fisher distribution $p(X \mid F) \propto \exp(\mathrm{tr}(F^\top X))$ to 28 data points of children aged between 2 to 10 and obtained the estimate

$$\hat{F} = 5.63 \times \begin{pmatrix} 0.583 & 0.629 & 0.514 \\ 0.660 & -0.736 & 0.151 \\ 0.473 & 0.252 & -0.844 \end{pmatrix}.$$

We use this value as the null model to be tested. We apply the kernel $k(X, Y) = \exp(\eta \cdot \mathrm{tr}(X^\top Y))$, where the parameter η was chosen by optimising the approximate test power from Eq. (3.22). Table 3.5 presents the p-values of each test. All mKSD tests show strong evidence to reject the fitted model at $\alpha = 0.05$; however, the Sobolev test, with p-value=0.126, is not powerful enough to reject the null at the same test level.

Table 3.5: p-values for vectorcardiogram data.

$\mathrm{mKSD}^{(1)}$	$\mathrm{mKSD}^{(2)}$	$\mathrm{mKSD}^{(0)}$/MMD	Sobolev
0.004	0.000	0.010	0.126

3.6.2 Wind direction data

As a real data on torus, we consider wind direction in Tokyo on 00:00 (x_1) and 12:00 (x_2) for each day in 2018[9]. Thus, the sample size is $n = 365$. The data were discretised into 16 directions, such as north-northeast. Figure 3.5 presents a 16×16 histogram of raw data.

Using noise contrastive estimation [Gutmann and Hyvärinen, 2012], Uehara et al. [2020] fitted the bivariate von-Mises distribution to the wind direction data

[9]Data is available on Japan Meteorological Agency website.

and obtained the estimate

$$\widehat{\xi} = (0.7170, 0.3954, 1.1499, 1.1499, -1.1274).$$

By setting this fitted model to the null model, we consider the goodness-of-fit testing of the bivariate von-Mises distribution in Eq.(3.5) via mKSD. We employ the product kernel of the von-Mises kernels as follows, $k((x_1, x_2), (y_1, y_2)) = \exp(\eta_1 \cos(x_1 - y_1) + \eta_2 \cos(x_2 - y_2))$. Similar as before, the parameters η_1 and η_2 were chosen by optimising the approximate test power. The p-value obtained by applying $\mathrm{mKSD}^{(1)}$ test is 0.434, which indicates that the model fits data well.

In addition, we fitted a simpler model with no interactions between x_1 and x_2, i.e. λ_{12} is set to zero in Eq.(3.5) so that the model reduces to the product of two independent von-Mises distribution on each direction. The p-value by $\mathrm{mKSD}^{(1)}$ is 0.002, which finds strong evidence to reject the null model. In other words, there is a significant interaction between wind direction on 00:00 and 12:00. We then carried out model criticism by mFSSD statistic in Eq.(3.25) with optimised test location via maximising approximate test power. Choosing the number of test locations $J = 10$, we plot the optimised locations in Figure 3.5. It provides information about dependence between wind direction at midnight and noon.

Appendices

3.A Proofs and Derivations

Quadratic form of mKSD

Proof of Theorem 3.4

Proof. We show that, the mKSD admits the form of taking expectation over p for bivariate functions $h_q^{(c)}$ which is independent of p. $h_q^{(c)}$ is also referred as the Stein kernel. The proof utilise the reproducing property of relevant RKHS and the fact that $\mathcal{A}_q^{(c)}$ is a linear functional of relevant test function f.

For $c = 1$, the test function is a stack of d-dimensional RKHS functions $\mathbf{f} \in \mathcal{H}^d$. $\mathbb{E}_p[\mathcal{A}_q^{(1)}\mathbf{f}]$ is a linear functional of $\mathbf{f} \in \mathcal{H}^d$. Then, from the Riesz representation theorem, there uniquely exists $\mathbf{r} = (r_1, \ldots, r_d) \in \mathcal{H}^d$ such that $\mathbb{E}_p[\mathcal{A}_q^{(1)}\mathbf{f}] = \langle \mathbf{f}, \mathbf{r} \rangle_{\mathcal{H}^d}$. By using the reproducing property of $\mathcal{H}$ associate with kernel k, we obtain

$$r_i(x) = \mathbb{E}_{\tilde{x}\sim p}\left[k(x,\tilde{x})\frac{\partial}{\partial \tilde{\theta}^i}\log(qJ) + \frac{\partial}{\partial \tilde{\theta}^i}k(x,\tilde{x})\right], \tag{3.27}$$

for $i = 1, \ldots, d$. Thus, the maximisation in $\mathrm{mKSD}^{(1)}(p\|q)$ is attained by $\mathbf{f} = \mathbf{r}/\|\mathbf{r}\|_{\mathcal{H}^d}$ and $\mathrm{mKSD}^{(1)}(p\|q)^2 = \|\mathbf{r}\|_{\mathcal{H}^d}^2$. Therefore, the quadratic form is obtained after straightforward calculations:

$$\begin{aligned}\mathrm{mKSD}^{(1)}(p\|q)^2 &= \left\langle \mathbb{E}_{x\sim p}[\mathcal{A}_q^{(1)}k(x,\cdot)], \mathbb{E}_{\tilde{x}\sim p}[\mathcal{A}_q^{(1)}k(\tilde{x},\cdot)]\right\rangle_{\mathcal{H}^d} \\ &= \mathbb{E}_{x,\tilde{x}\sim p}\Bigg[\underbrace{\left\langle \mathcal{A}_q^{(1)}k(x,\cdot), \mathcal{A}_q^{(1)}k(\tilde{x},\cdot)\right\rangle_{\mathcal{H}^d}}_{h_q^{(1)}(x,\tilde{x})}\Bigg],\end{aligned}$$

and the assertion follows.

For $c = 2$, similar argument applies where the test function is a scalar-valued

RKHS $\tilde{f} \in \mathcal{H}$. Instead of Eq.(3.27), we have $\tilde{r} \in \mathcal{H}$, s.t. $\mathbb{E}_p[\mathcal{A}_q^{(2)}\tilde{f}] = \langle \tilde{f}, \tilde{r} \rangle_{\mathcal{H}}$ and

$$\tilde{r}(x) = \mathbb{E}_{\tilde{x}\sim p}\left[\sum_{ij} g^{ij}\left(\frac{\partial}{\partial\tilde{\theta}^j}k(x,\tilde{x})\frac{\partial}{\partial\tilde{\theta}^i}\log(qJ) + \frac{\partial^2}{\partial\tilde{\theta}^i\partial\tilde{\theta}^j}k(x,\tilde{x})\right)\right], \tag{3.28}$$

and the maximisation in $\mathrm{mKSD}^{(2)}(p\|q)$ is attained by $\tilde{f} = \tilde{r}/\|\tilde{r}\|_{\mathcal{H}}$; thus $\mathrm{mKSD}^{(2)}(p\|q)^2 = \|\tilde{r}\|_{\mathcal{H}}^2$. The assertion then follows from the similar calculations as above.

For $c = 0$, the quadratic form is readily obtained from derivation of maximum-mean-discrepancy (MMD) [Gretton et al., 2012a] form as shown in Theorem 3.6. Alternatively, for scalar test function $h \in \mathcal{H}$, we can write,

$$\mathrm{mKSD}^{(0)}(p\|q) = \sup_{\|h\|_{\mathcal{H}}\leq 1} \mathbb{E}_p[\mathcal{A}_q^{(0)}h] = \sup_{\|h\|_{\mathcal{H}}\leq 1} |\mathbb{E}_p[h] - \mathbb{E}_q[h]|\,,$$

where taking the supremum we get,

$$\begin{aligned}\mathrm{mKSD}^{(0)}(p\|q)^2 &= \Big\langle \mathbb{E}_p\big[k(x,\cdot) - \mathbb{E}_q[k(x,\cdot)]\big], \mathbb{E}_p\big[k(\tilde{x},\cdot) - \mathbb{E}_q[k(\tilde{x},\cdot)]\big]\Big\rangle_{\mathcal{H}} \\ &= \mathbb{E}_{x,\tilde{x}\sim p}\Big\langle \underbrace{k(x,\cdot) - \mathbb{E}_q[k(x,\cdot)]}_{\mathcal{A}_q^{(0)}k(x,\cdot)}, k(\tilde{x},\cdot) - \mathbb{E}_q[k(\tilde{x},\cdot)]\Big\rangle_{\mathcal{H}}.\end{aligned}$$

The assertion follows. □

The quadratic form is useful when computing the empirical estimate for the expectation where only samples from unknown distribution p is observed. We also note that $\mathbb{E}_q[k(\tilde{x},\cdot)]$, in general, is not possible to obtain in analytical form, especially when the density q is only given up to normalisation. Samples from q, if possible to obtain from unnormalised density, can be useful to estimate $\mathcal{A}_q^{(0)}k(x,\cdot)$, where we denote as $\widehat{\mathcal{A}_q^{(0)}}k(x,\cdot)$.

Characterisation of mKSD

Proof of Theorem 3.5

Proof. Denote $\mathbf{s}_p^{(c)}(\cdot) = \mathbb{E}_{\tilde{x}\sim p}[\mathcal{A}_q^{(c)}k(\tilde{x},\cdot)] \in \mathcal{F}$ and we can write $\mathrm{mKSD}^{(c)}(p\|q)^2 = \|\mathbf{s}_p(\cdot)\|_{\mathcal{F}}^2 \geq 0$, where $\mathcal{F}$ can be $\mathcal{H}$ for $c = 0, 2$ or $\mathcal{H}^d$ for $c = 1$. If $p = q$, then $\mathrm{mKSD}^{(c)}(p\|q)^2 = 0$ from the Stein identity.

Conversely, if $\mathrm{mKSD}^{(c)}(p\|q)^2 = 0$, then $\mathbf{s}_p^{(c)}(x) = \mathbf{0}$, a zero vector in $\mathbb{R}^d$ for $c = 1$ and a scalar zero in $\mathbb{R}$ for $c = 0, 2$, $\forall x$, s.t. $p(x) > 0$. Then, from

$\log(q/p) = \log(qJ) - \log(pJ)$, we obtain,

$$\mathbb{E}_{\tilde{x}\sim p}\left[L_i(\tilde{x})k(\tilde{x},x)\right] = (\mathbf{s}_p^{(1)})_i(x) - \mathbb{E}_{\tilde{x}\sim p}\left[\mathcal{A}_p^{(1)}K(\tilde{x},x)\right] = 0,$$

and

$$\mathbb{E}_{\tilde{x}\sim p}\left[L(\tilde{x})k(\tilde{x},x)\right] = (\mathbf{s}_p^{(c)})_i(x) - \mathbb{E}_{\tilde{x}\sim p}\left[\mathcal{A}_p^{(c)}K(\tilde{x},x)\right] = 0,$$

for $c = 0, 2$, for every x with positive densities. Since k is compact-universal, vanishes at $\partial\mathcal{M}$ and $\mathcal{M}$ is smooth and compact, it implies that $L_i^{(1)} = 0, \forall i$ [Carmeli et al., 2010, Theorem 4(b)] (for $c = 1$, $i = [d]$; for $c = 0, 2$, $i = 1$). Therefore, $\log(q/p)$ is constant on $\mathcal{M}$. Since both p and q are both densities on $\mathcal{M}$ that integrate to one, we conclude $p = q$. □

Asymptotics of mKSD

Proof of Theorem 3.7

Proof. To show part 1, it is enough to check the mKSD statistics is degenerate U-statistics under $H_0 : p = q$. By considering test function $f = k(x, \cdot)$ (or its relevant vector-valued form for $c = 1$), Stein identity shows that,

$$\mathbb{E}_{\tilde{x}\sim p}[\mathcal{A}_q^{(c)}k(x,\tilde{x})] = 0, \forall x \in \mathcal{M}, \tag{3.29}$$

so that the variance $\sigma_c^2 = 0$ for $c = 0, 1, 2$. Then the standard results for degenerate U-statistics in [Serfling, 2009, Section 5.5.2] apply and the assertions follow.

In addition, it is interesting to note link the result for $c = 0$ with the asymptotic result in as

$$h_q^{(0)}(x,\tilde{x}) = k(x,\tilde{x}) - \xi(x) - \xi(\tilde{x}) + C,$$

where $C = \mathbb{E}_{x,\tilde{x}\sim q}k(x,\tilde{x})$ is a constant, $\xi(x) = \mathbb{E}_{\tilde{x}\sim q}k(x,\tilde{x})$ is only a function of x and $\xi(\tilde{x}) = \mathbb{E}_{x\sim q}k(x,\tilde{x})$ is only a function of $\tilde{x}$. The formulation is analogous to the asymptotic results for MMD, as shown in [Gretton et al., 2012a, Theorem 8]: $h_q^{(0)}(x,\tilde{x})$ is equivalent to the notion of $\tilde{k}(x,\tilde{x})$ in [Gretton et al., 2012a].

Part 2 follows as $\sigma_c^2 > 0$ under $H_1 : p \neq q$ by Theorem 3.5. Apply asymptotic distribution of non-degenerate U-statistics [Serfling, 2009, Section 5.5.1] and the assertions follow. □

3.B More on Bahadur Efficiency

In this section, we introduce additional relevant concepts to analyse Approximate Relative Efficiency (ARE) between two tests, characterised by *Bahadur slope* [Bahadur et al., 1960] and corresponding *Bahadur efficiency*.

Definition 3.4. *Let $\mathcal{D}(a,t)$ be a class of all continuous cumulative distribution functions (CDF) F such that $-2\log(1-F(x)) = ax^t(1+o(1))$, as $x \to \infty$ for $a > 0$ and $t > 0$.*

Proposition 3.3. *The approximate Bahadur slope (ABS) for the tests with* $\mathrm{mKSD}^{(c)}$, $c = 0, 1, 2$ *is*

$$c^{(\mathrm{mKSD}^{(c)})} := \frac{\mathbb{E}_p[h_q^{(c)}(x,\tilde{x})]}{\mathbb{E}_q[h_q^{(c)}(x,\tilde{x})^2]^{\frac{1}{2}}},$$

where $h_q^{(c)}(x,\tilde{x})$ is the Stein kernel for $\mathrm{mKSD}^{(c)}$*, and $\rho(n) = n$.*

Proof. Using Theorem 9 and Theorem 11 in [Jitkrittum et al., 2017], we know that $n \cdot \mathrm{mKSD}_u^{(c)}(p\|q)^2$ in Eq.(3.17) is in the class of $\mathcal{D}(a = 1/\omega_c, t = 1)$ for ω_c^2 is the variance of the statistic. By Stein identity, $\mathbb{E}_{x,\tilde{x}\sim q}\left[h_q^{(c)}(x,\tilde{x})\right]^2 = 0$. Hence, using second point in Theorem 9 [Jitkrittum et al., 2017] and choosing $\rho = n$, we have $n \cdot \mathrm{mKSD}_u^{(c)}(p\|q)^2 \backslash \rho(n) \to \mathrm{mKSD}^{(c)}(p\|q)^2$ by weak law of large numbers. □

The Case Study on Circular distribution $\mathbb{S}^1$ Proof of Theorem 3.8

Proof. To compute $\mathrm{E}_{1,2}(\kappa)$, we can rewrite the following:

$$\mathrm{E}_{1,2}(\kappa) = \frac{\mathbb{E}_p[h_q^{(1)}(x,\tilde{x})]}{\mathbb{E}_p[h_q^{(2)}(x,\tilde{x})]} \cdot \frac{\mathbb{E}_q[h_q^{(2)}(x,\tilde{x})^2]^{\frac{1}{2}}}{\mathbb{E}_q[h_q^{(1)}(x,\tilde{x})^2]^{\frac{1}{2}}}$$

The second term only involves integrals over $q(x) \propto 1$, which is independent of κ and we can solve it as $\frac{\mathbb{E}_q[h_q^{(2)}(x,\tilde{x})^2]^{\frac{1}{2}}}{\mathbb{E}_q[h_q^{(1)}(x,\tilde{x})^2]^{\frac{1}{2}}} = 1.692 > 1$. For the first term, the ratio is monotonic decreasing w.r.t. $\kappa > 0$ and solving numerically at $\kappa = 10$, we get $\frac{\mathbb{E}_p[h_q^{(1)}(x,\tilde{x})]}{\mathbb{E}_p[h_q^{(2)}(x,\tilde{x})]} = 2.953 > 1$. Hence, for $\kappa \in (0, 10)$, $\mathrm{E}_{1,2} > 1$. □

We can apply similar approach to compare the relative test efficiency $\mathrm{E}_{0,1}(\kappa)$ between $\mathrm{mKSD}^{(0)}$ and $\mathrm{mKSD}^{(1)}$. We plot numerical solutions in Figure 3.6. From Figure 3.6, we see that $\mathbb{E}_{1,2}$ and $\mathbb{E}_{0,1}$ both greater than 1 for $\kappa \in (0, 10)$. For further increase of κ, there is a trend for both relative efficiencies stabilising at some value greater than 1. Theoretical analysis for such limiting behaviour is of an interesting future topic. Although Figure 3.6 shows that $\mathbb{E}_{0,1}(\kappa) > 1$ for

small perturbation from the null, i.e. $\kappa \in (0, 20)$ which suggest the relative efficiency of $\mathrm{mKSD}^{(0)}$ is higher than the first order test $\mathrm{mKSD}^{(1)}$, it is usually not possible to compute MMD analytically and the normalised density is required.

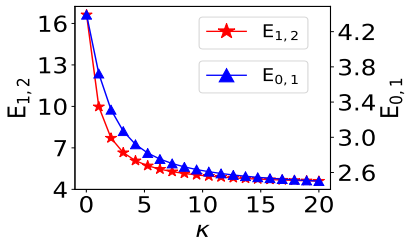


Figure 3.6: Relative Test Efficiency

Intuitively, with sampling error of order $\sqrt{n}$ and $\rho(n) = n$ is chosen to compute Bahadur slope, the MMD computed from samples are less efficient to perform goodness-of-fit test compared to mKSD tests that directly access the unnormalised density, as shown in Figure 3.4. Similar findings are also observed in other settings where MMD is considered to perform goodness-of-fit tests [Liu et al., 2016; Jitkrittum et al., 2017; Yang et al., 2018, 2019; Xu and Matsuda, 2020]. In addition, correctly sampling from Riemannian manifold is non-trivial and can be time-consuming for sample-based tests.

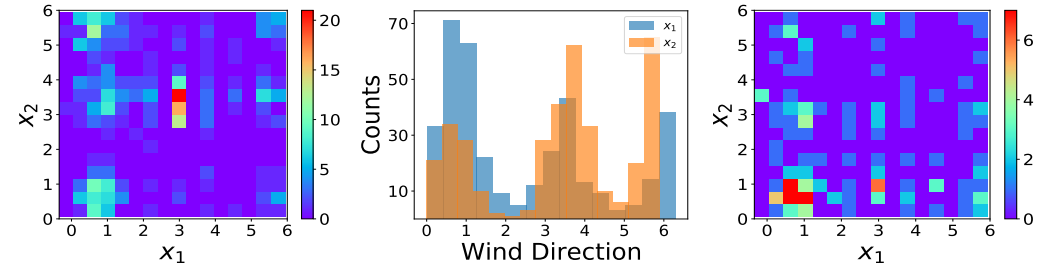
3.C More on Model Criticism

In this section, we provide additional details on model criticism for wind data present in Section 3.6.2. We fitted the model in Eq.(3.5) by using noise contrastive estimation [Uehara et al., 2020] and our test does not find evidence to reject the fitted model, suggesting a good fit for the wind direction data. In addition, we consider the model without interaction term between two directions:

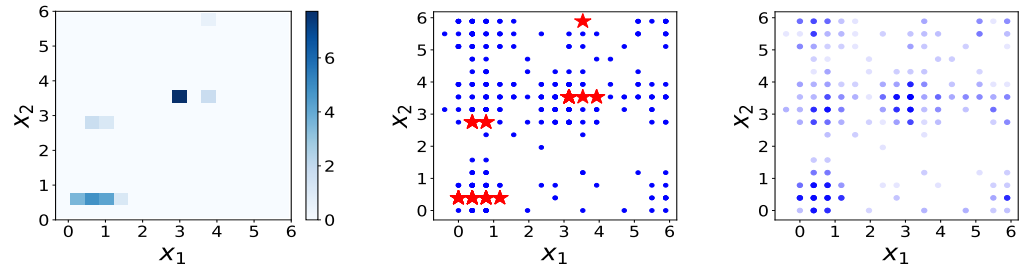
$$\widetilde{q}(x_1, x_2 \mid \widetilde{\xi}) \propto \exp\{\kappa_1 \cos(x_1 - \mu_1) + \kappa_2 \cos(x_2 - \mu_2)\}, \tag{3.30}$$

which is equivalent to model in Eq.(3.5) by imposing $\lambda_{12} = 0$. This model can be viewed as product of marginal distributions of x_1 and x_2 and we refer as factorised model. Our test reject the null at test level $\alpha = 0.05$ suggesting a poor fit of the factorised model.

To further visualise the difference between models in Eq.(3.5) and Eq.(3.30), we plot histogram of each wind direction in Figure 3.7(b) and samples from the factorised model $\widetilde{q}$ in Figure 3.7(c) where no interactions are present between x_1 and x_2. Compare with the wind direction data, shown again in Figure 3.7(a), we can see that Figure 3.7(c) differs the most at the regions of $\tilde{x} = (x_1, x_2) = (2.8, \pi)$ (data model denser) and $\tilde{x}' = (x_1, x_2) = (1, 1)$ ($\tilde{q}$ model denser). Such difference is captured by our optimised test locations from mFSSD in Figure 3.7(e), where $\tilde{x}$ is at the region with 3 stars in a row and $\tilde{x}'$ is around the region with 4-stars in a



(a) Wind direction model (b) Marginals for each direction (c) Samples from factorised model



(d) Objective values for $J = 10$ (e) Optimised locations; $J = 10$ (f) Objective values for a single test location

Figure 3.7: Visualising the fitted model and rejected model for wind direction data.

row. It shows the effectiveness of mFSSD in distinguishing the differences between distributions. As $\tilde{q}$ is referred as imposing data model in Eq.(3.5) to be 0, a negative $\lambda_{12} = -1.1274 < 0$ in the data model implies that positive $\sin(x_1 - \mu_1)\sin(x_2 - \mu_2)$ is less dense. With $\mu_1 = 1.1499 = \mu_2$, $\sin(x_1 - \mu_1)\sin(x_2 - \mu 2)$ is positive around the region the $\tilde{x}'$ making the data model less dense, as shown in Figure 3.7(a) and 3.7(c).

3.D Uniformity Tests for Directional Distributions

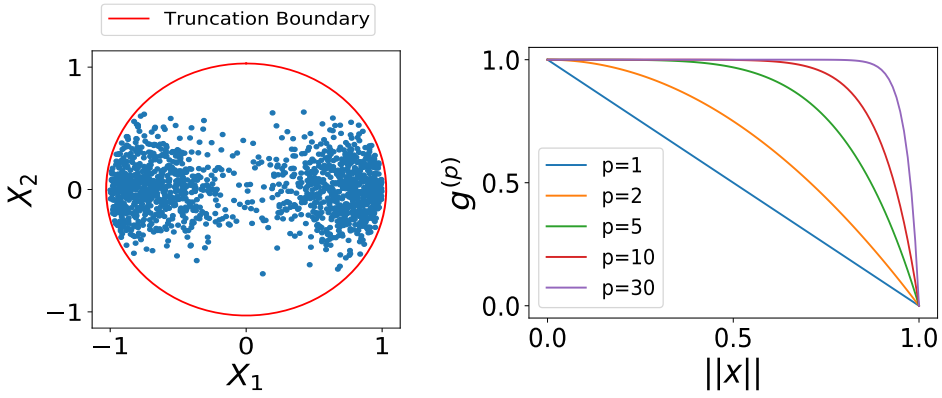
We present Rayleigh test and Kuiper test for uniformity for directional distributions.

Rayleigh Test

The test statistic of Rayleigh test is

$$R_n := \frac{2}{n}\left[\left(\sum_{i=1}^{n}\cos\theta_i\right)^2 + \left(\sum_{i=1}^{n}\sin\theta_i\right)^2\right].$$

Under the null, we have $R_n \sim \chi^2_2$. Therefore, the critical value is given by the quantile of chi-square distribution. For example, if the significance level is set to $\alpha = 0.01$, then the critical value is 9.210.



(a) Samples obtained (b) Approximation function project into 1D

Figure 3.8: Examples on truncated Gaussian mixture

Kuiper Test

Kuiper test for uniformity is based on the cumulative distribution function (CDF). The CDF of the uniform distribution is

$$F(\theta) = \frac{\theta}{2\pi}.$$

We sort the samples to $0 \leq \theta_1 \leq \cdots \leq \theta_n \leq 2\pi$ and compute

$$D_n^+ := \sqrt{n} \sup_{\theta \in [0,2\pi)} \{F_n(\theta) - F(\theta)\} = \sqrt{n} \max_{1 \leq i \leq n} \left(\frac{i}{n} - U_i \right),$$

$$D_n^- := \sqrt{n} \sup_{\theta \in [0,2\pi)} \{F(\theta) - F_n(\theta)\} = \sqrt{n} \max_{1 \leq i \leq n} \left(U_i - \frac{i-1}{n} \right),$$

where $U_i = \theta_i/(2\pi)$. Then, the test statistic is defined as

$$V_n := D_n^+ + D_n^-.$$

The critical value is found in the statistical table. For example, for the significance level $\alpha = 0.01$, the critical value is 2.001.

3.E Additional Discussions on Non-empty Boundary

We consider truncated distribution on a disc, whose boundary is non-empty. The mixture of Gaussian distribution is truncated in a circle $B_1(\mathbb{R}^2)$,

$$\tilde{q}_\nu(x) \propto \frac{1}{2}\mathcal{N}(x|\mu_1, \Sigma_\nu) + \frac{1}{2}\mathcal{N}(x|\mu_2, \Sigma_\nu), \forall x \in B_1(\mathbb{R}^2),$$

	n=100	n=400	n=700	n=1000
mKSD(p=30)	0.106	0.940	1.000	1.000
mKSD(p=10)	0.086	0.938	0.990	1.000
mKSD(p=5)	0.026	0.910	0.930	1.000
mKSD(p=2)	0.018	0.818	0.902	1.000
mKSD(p=1)	0.012	0.406	0.780	1.000
MMD	0.014	0.514	0.818	0.920

Table 3.6: Rejection rate under the alternative $\nu = \frac{1}{2}$; $\alpha = 0.01$; 500 trials.

where $\mu_1 = (-1, 0, 0)$, $\mu_2 = (1, 0, 0)$ and $\Sigma_\nu = \begin{pmatrix} 1 & \nu \\ \nu & 1 \end{pmatrix}$ is shared between two components. The null distribution is set as $\nu = 0$ and null samples are shown in Figure. 3.8(a). The alternative is constructed by perturbing ν.

To understand the "vanishing-at-boundary" effect of the approximation via g, we use $g^{(p)}(x) = 1 - \|x\|^p$, which has value 0 at the red boundary in Figure. 3.8(b). We consider Gaussian kernel with unit bandwidth $k(x, \cdot)$ and the product kernel $k^{(p)} = g^{(p)}(x)k(x, \cdot)$ and call the corresponding test statistics mKSD(p) when $k^{(p)}$ is used as the reproducing kernel. When p is large, $g^{(p)}$ acts as a proxy to g_ϵ stated in Eq. (3.16).

From the simulation result in Table. 3.6, we can see that the power quickly goes to maximal test power when sample size increase. We can also see that as p increases, the test power results is getting very similar and the p does not need to be too large. Such observation implies that with relatively shap decreasing g to vanish at $\partial\mathcal{M}$, the kernel/mKSD is already good enough to distinguish the alternative from the null in the example presented, i.e. $p = 5$ still looks not "too sharp" from Figure. 3.8(b) but test power is not too far from much larger p.

Chapter 4

Goodness-of-fit Tests for Censored Data

Summary We study the non-parametric goodness-of-fit testing procedures for censored data. Censored data occurs when the event time of interest is not accurately observed but, instead, a random interval which the event time belongs to is given. While the testing procedures for uncensored data are unable to adapt for the censored data, we develop a collection of kernelised Stein discrepancy tests to incorporate the presence of censoring. We study each of them theoretically and empirically and discuss the advantages and disadvantages of these tests. Our experimental results show that our proposed methods perform better than existing tests, including a recent test based on the kernelised maximum mean discrepancy.

4.1 Introduction

An important topic of study in statistics is the distribution of times to a critical event, otherwise known as survival times: examples include the infection time from a disease [Andersen et al., 2012; Mirabello et al., 2009]; the death time of a patient in a clinical trial [Collett, 2015; Biswas et al., 2007]; or the possible re-offending times for released criminals [Chung et al., 1991]. Survival data are frequently subject to censoring: the time of interest is not observed, but rather a bound on it. The most common scenario studied that we focus on in this chapter, is the *right censoring*, where a lower bound on the survival time is observed. For instance, a patient might leave a clinical trial before it is completed, meaning that we only obtain a lower bound on the time of death. The definitions and terminologies for the survival analysis setting and censored data will be provided in Section 4.2.

We address the setting where a model of survival times is proposed, and it is desired to test this model against observed data in the presence of censoring, which is in the regime of *goodness-of-fit* testing. When departures from the model follow a known parametric family, a number of classical tests are available, being the most popular in practice the log-rank test [Hollander and Proschan, 1979], and

its generalisation, the weighted log-rank test [Brendel et al., 2014]. For an overview of these and other methods we refer the reader to [Klein and Moeschberger, 2006]

In the event of more general departures from the null, kernel methods may be used to construct a powerful class of non-parametric tests to detect a greater range of alternative scenarios. For the uncensored case, a popular class of kernel goodness-of-fit tests [Liu et al., 2016; Chwialkowski et al., 2016; Jitkrittum et al., 2017] using Stein's method [Barbour and Chen, 2005; Ley et al., 2017; Gorham and Mackey, 2015] has been introduced in Section 2.2, which can be computed even when the model is known only up to normalisation. To tackle the goodness-of-fit problem for censored data, we consider the similar non-parametric hypothesis testing strategy tests that construct the test statistics by taking the supremum over a rich enough RKHS functions with respect to a particular Stein operator of choice. While an alternative strategy would be simply to run a two-sample test using samples from the model, using for instance the MMD [Gretton et al., 2012a], Stein tests are more computationally efficient (with no additional sampling is needed), and can take advantage of model structure to achieve better test power.

In this chapter, we propose to the kernel Stein goodness-of-fit tests to the setting of survival analysis with right-censored data. In Section 4.3, we introduce three separate approaches to constructing a Stein operator in the presence of censoring. The *Survival Stein Operator* is the most direct generalisation of the Stein operator used in the uncensored KSD test. The *Martingale Stein Operator* uses a different construction, based on a classical martingale studied in the survival analysis literature. The *Proportional Stein Operator* is designed for composite null hypotheses where the *hazard function* (that is, the instantaneous probability of an event at a given time, conditioned on survival to that time) is known only up to a constant of proportionality in this case. For instance, we may wish to use a constant hazard as the null hypothesis, without specifying in advance the value of the constant.

The rest of the chapter is structured as follows: in Section 4.4, we construct kernel statistics of goodness-of-fit, based on each of the operators previously introduced. We characterise the asymptotics of each statistic in Section 4.5. We find that in order to guarantee convergence in distribution under the null, the kernel statistic based on the Survival Stein Operator requires more restrictive conditions than the statistic built on the Martingale Stein Operator. In other words, the straightforward extension of the uncensored test is in fact the more restrictive approach of the two. Stronger assumptions again are required in obtaining convergence in distribution for the Proportional Stein Operator statistic, which should come as no surprise, given that the null is now an entire model class. For each statistic, we pro-

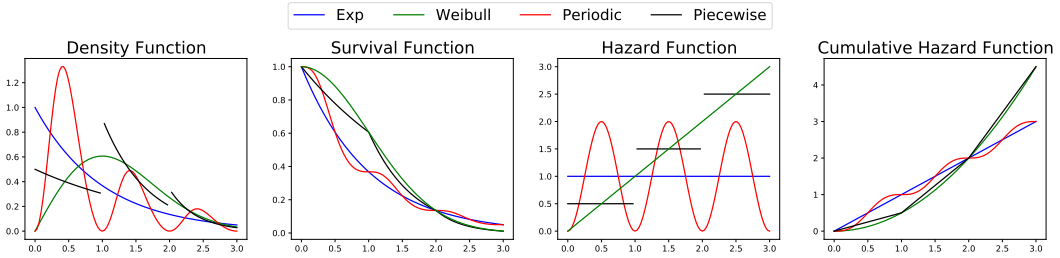


Figure 4.1: Example functions related to survival analysis.

pose a wild bootstrap approach to obtain the test threshold. Empirical studies and results are presented in Section 6.6, where we compare with a recent state-of-the-art non-parametric test for censored data [Fernandez and Gretton, 2019] based on the maximum-mean-discrepancy, which has been shown to outperform classical tests. For challenging cases, our Stein tests surpass the maximum-mean-discrepancy test.

4.2 Survival Analysis Background

4.2.1 Important Functions in Survival Analysis

In survival analysis, we deal with the observed time T, which is associated with the actual survival time of interest X that we do not have direct access to and the censoring time C. We denote by f_T, f_X and f_C, the respective density functions associated with the random variables T, X and C. Similarly, we denote by F_T, F_X and F_C, the respective cumulative distribution functions (c.d.f.); and by $S_T = 1-F_T$, $S_X = 1-F_X$ and $S_C = 1-F_C$, the survival functions. An important element in survival analysis is the hazard function which represents the instantaneous risk of dying[1] at a given time. Given a distribution with density f_X and survival function S_X, the hazard function $\lambda_X(x)$ is given by $f_X(x)/S_X(x)$, which can be seen as the density at x of a random variable X conditioned on the event $\{X \geq x\}$. The corresponding cumulative hazard function is defined as $\Lambda_X(x) = \int_0^x \lambda_X(t)dt$. A useful feature of the hazard function is that there is a one-to-one relation between hazard and density functions through the relation $S_X(x) = e^{-\Lambda_X(x)}$. For the random variables T and C, we denote by λ_T and λ_C their respective hazard functions, and by Λ_T and Λ_C, their cumulative hazards functions. As a remark, every continuous non-negative function $\lambda : \mathbb{R}_+ \to \mathbb{R}$ can be a hazard function, as long as $\int_{\mathbb{R}_+} \lambda(t)dt = \infty$, thus, describing hazards is much easier than describing densities, as we do not need to worry about normalisation constants. Examples of corresponding functions for different models are displayed in Figure 4.1.

[1] Dying refers to the opposite concept of surviving.

4.2.2 Censored Data

Let $(X_1, \ldots, X_n) \overset{\text{i.i.d.}}{\sim} F_X$ be the survival times, which are non-negative real-valued random variables of interest, and let $(C_1, \ldots, C_n) \overset{\text{i.i.d.}}{\sim} F_C$ be the censoring times which is another collection of non-negative random variables. In this chapter, we assume the non-informative censoring setting, where the censoring times are independent of the survival times. The data we observe correspond to (T_i, Δ_i) where $T_i = \min\{X_i, C_i\}$ and $\Delta_i = \mathbb{1}_{\{X_i \leq C_i\}}$. We can imagine that X_i is the time of interest such as the death of a patient and C_i is the time a patient leaves the study for some other reason, thus, for some patients we observe their actual death time, whereas for others we just observe a lower bound: the time they left the study without dying. Δ_i indicates if we are observing X_i or C_i.

As the observations for censored data come as pairs (T_i, Δ_i), it is convenient to consider the joint measure μ on $\mathbb{R}_+ \times \{0, 1\}$ induced by the pair (T, Δ). We write μ_X to denote the measure μ when the survival times of interest X_i are generated according to f_X, and μ_0 if they are generated under f_0 (i.e., under the null). Note that μ_X and μ_0 also depend on f_C, however we don't make this dependence explicit, since for goodness-of-fit we only care about f_0 and f_X. Finally, we know the following identities in survival analysis, which will be useful for later discussions: for any measurable function ϕ,

$$\mathbb{E}_X[\Delta\phi(T)] = \int_0^\infty \phi(s) f_X(s) S_C(s) ds, \tag{4.1}$$

$$\mathbb{E}_X[(1-\Delta)\phi(T)] = \int_0^\infty \phi(s) f_C(s) S_X(s) ds. \tag{4.2}$$

Here $\mathbb{E}_X = \mathbb{E}_{\mu_X}$ means that we are taking expectation w.r.t. $(T, \Delta) \sim \mu_X$. Similarly, we write $\mathbb{E}_0$ to indicate $(T, \Delta) \sim \mu_0$ (under the null hypothesis).

4.3 Stein Operators for Censored Data

In this section, we describe a set of Stein operators for censored data. We denote by Ω the set of functions $\mathbb{R}_+ \times \{0, 1\} \to \mathbb{R}$, and recall that μ_0 is the measure induced by data (T, Δ) under the null hypothesis.

Definition 4.1. *Let $\mathcal{H} \subseteq L^2(f_0)$. We call $\mathcal{T}_0 : L^2(f_0) \to \Omega$ a Stein operator for $\mathcal{H}$ if for each $\omega \in \mathcal{H}$*

$$\mathbb{E}_0\left[(\mathcal{T}_0\omega)(T, \Delta)\right] = 0. \tag{4.3}$$

An interesting technical point is that our operator takes functions $\omega : \mathbb{R}_+ \to \mathbb{R}$ and maps them to Ω. The idea behind having these two spaces is that while our data

of interest is a time (hence the space $\mathcal{H}$ of functions $\mathbb{R}_+ \to \mathbb{R}$), we actually observe pairs (T_i, Δ_i), hence we need functions in Ω.

We choose the general class $\mathcal{H}$ to be an RKHS. We assume that $\mathcal{H}$ contains only differentiable and bounded functions, and that if $\omega \in \mathcal{H}$ then $\omega' \in \mathcal{H}$. These requirements are not restrictive and most of the standard kernels in the literature generate RKHSs with these properties, including the Gaussian kernel. Further properties of $\mathcal{H}$ will be imposed if needed in particular cases.

4.3.1 Survival Stein Operator

Observe that $T_i = X_i$ if and only if $\Delta_i = 1$. One might be tempted to use only the uncensored observations to approximate $\int_0^\infty (\mathcal{T}_0\omega)(x) f_0(x) dx$, where $\mathcal{T}_0$ is the standard Stein operator similar to Eq. (2.19), by computing

$$\frac{1}{n}\sum_{i=1}^{n} \Delta_i (\mathcal{T}_0\omega)(T_i) = \frac{1}{n}\sum_{i=1}^{n} \Delta_i (\mathcal{T}_0\omega)(X_i),$$

however, this sum does not converge to $\int_0^\infty (\mathcal{T}_0\omega)(x) f_0(x) dx$ as the term Δ_i introduces bias due to censoring. Indeed, such an empirical average converges to $\int_0^\infty (\mathcal{T}_0\omega)(x) S_C(x) f_X(x) dx$ by Eq. (4.1). To account for this bias we redefine $\mathcal{T}_0 : \mathcal{H}^{(s)} \to \Omega$ as

$$(\mathcal{T}_0\omega)(x, \delta) = \delta \frac{(\omega(x) S_C(x) f_0(x))'}{S_C(x) f_0(x)} + \omega(0) f_0(0). \tag{4.4}$$

Here we write $\mathcal{H}^{(s)}$ instead of $\mathcal{H}$ whenever we assume that the additional condition is satisfied,

$$\int_{\mathbb{R}^+} \left| (\omega(x) S_C(x) f_0(x))' \right| dx < \infty, \quad \forall \omega \in \mathcal{H}, \tag{4.5}$$

which guarantees that the operator is well-defined. Notice that $\omega(0) f_0(0)$ in equation (4.4) appears since we do not necessarily assume a vanishing boundary at 0.

Under the null hypothesis, $(T_i, \Delta_i) \sim \mu_0$, it holds that

$$\frac{1}{n}\sum_{i=1}^{n} (\mathcal{T}_0\omega)(T_i, \Delta_i) \to \mathbb{E}_0[(\mathcal{T}_0\omega)(T, \Delta)] \tag{4.6}$$

as the number of data points tends to infinity, and $\mathbb{E}_0[(\mathcal{T}_0\omega)(T, \Delta)] = 0$ due to

Eq. (4.1) and the fact that

$$\int_{\mathbb{R}^+} (\omega(x) S_C(x) f_0(x))' dx + \omega(0) f_0(0) = 0, \tag{4.7}$$

which is proved using integration by parts. Notice that in this argument we use that $\mathcal{H}^{(s)}$ only contains bounded functions, allowing us to get rid of the boundary at infinity.

The operator $\mathcal{T}_0$ can be seen as a natural extension of the Stein operator [Gorham and Mackey, 2015] to censored data. Observe that in the uncensored case, $S_C(x) \equiv 1$ recovers the standard Stein operator.

Unfortunately, in the goodness-of-fit setting, we only have access to the null distribution $f_0(x)$ but not to the censoring distribution $f_C(x)$, thus $S_C(x)$ needs to be estimated. The standard estimator for S_C is the Kaplan-Meier estimator [Kaplan and Meier, 1958] which can be very data inefficient, leading to an unsatisfactory testing procedure.

To bypass the approximation of S_C we define the survival Stein operator $\mathcal{T}_0^{(s)}$: $\mathcal{H}^{(s)} \to \Omega$ as

$$\begin{aligned}(\mathcal{T}_0^{(s)}\omega)(x,\delta) = \delta\omega'(x) + \frac{\lambda_0'(x)}{\lambda_0(x)}\delta\omega(x) \\ - \lambda_0(x)\omega(x) + \lambda_0(0)\omega(0)\end{aligned} \tag{4.8}$$

Proposition 4.1. *Consider $\mathcal{T}_0$ and $\mathcal{T}_0^{(s)}$ defined in Eq.* (4.4) *and Eq.* (4.8)*, respectively. Let $(T, \Delta) \sim \mu_0$. Then*

$$\mathbb{E}_0[(\mathcal{T}_0^{(s)}\omega)(T,\Delta)] = \mathbb{E}_0[(\mathcal{T}_0\omega)(T,\Delta)] = 0, \quad \forall \omega \in \mathcal{H}^{(s)}.$$

The previous proposition says that if the data we observed were generated from μ_0 then the expectation of the operators $\mathcal{T}_0$ and $\mathcal{T}_0^{(s)}$ are equal for each function in $\mathcal{H}^{(s)}$. However, the relation between $\mathcal{T}_0$ and $\mathcal{T}_0^{(s)}$ is stronger than merely equality in expectation, indeed, under a slightly stronger condition on the form of the distribution f_0 and f_C we get the following result.

Proposition 4.2. *Assume that*

$$\int_0^\infty (\lambda_C(x) + \lambda_0(x)) f_C(x) f_0(x) < \infty, \tag{4.9}$$

then, under the null hypothesis, i.e. $(T_i, \Delta_i) \sim \mu_0$, we have that, as the number of

data points tends to infinity,

$$\sup_{\omega \in B_1(\mathcal{H})} \frac{1}{n} \sum_{i=1}^{n} (\mathcal{T}_0^{(s)}\omega)(T_i, \Delta_i) - (\mathcal{T}_0\omega)(T_i, \Delta_i) \xrightarrow{p} 0,$$

where $B_1(\mathcal{H})$ denotes the unit ball of RKHS $\mathcal{H}$ and $\xrightarrow{p}$ denotes convergence in probability

The proof utilises a symmetrisation argument followed by law of large numbers and details are shown in Appendix 4.A. To better understand the survival Stein operator, we interpret the proposed Stein operator by making connections to the Stein operator used in the uncensored case.

A careful computation gives the following equivalent expression for the expectation of $(\mathcal{T}_0^{(s)}\omega)(T, \Delta)$ for $(T, \Delta) \sim \mu_X$:

$$\begin{aligned}\mathbb{E}_X[(\mathcal{T}_0^{(s)}\omega)(T, \Delta)] &= \mathbb{E}_X \left[\omega(T)\Delta \left(\log \frac{f_0(T)}{f_X(T)}\right)'\right] \\ &- \mathbb{E}_X \left[\omega(T)(1-\Delta)(\lambda_0 - \lambda_X)(T)\right] + \omega(0)(\lambda_0 - \lambda_X)(0).\end{aligned}$$

Here, we can relate the first expectation to uncensored observations: $\Delta = 1$; the second expectation to censored observations: $\Delta = 0$; and the third term describes a shift due to boundary conditions.

The expectation of the uncensored part is equal to

$$\int_0^\infty \omega(x) \left(\log \frac{f_0(x)}{f_X(x)}\right)' S_C(x) f_X(x) dx,$$

which is analogous to what we obtain in the uncensored case, with an additional S_C weighting. If we have no censoring, then $S_C \equiv 1$, recovering the expression found in [Chwialkowski et al., 2016]. On the other hand, the expectation of the censored part is equal to

$$\int_0^\infty \omega(x) \left(\frac{S_X(x)}{S_0(x)} f_0(x) - f_X(x)\right) f_C(x) dx,$$

which measures the discrepancy between f_0 and f_X through survival weights, under the measure of censoring f_C. In the absence of censoring, $f_C = 0$ a.e., so this term appears due to the censoring variable. Notice that if differences between f_0 and f_X occur at times t where $S_C(t) = 0$, which corresponds to the observations at this time being entirely censored, then no method will detect these differences.

4.3.2 Martingale Stein Operator

While the previous approach mimics the classic diffusion-type Stein operator, it has similar drawbacks. Similarly to what we observe in KSD tests for uncensored case [Chwialkowski et al., 2016; Liu et al., 2016], our Stein operator $\mathcal{T}_0^{(s)}$ requires very strong integrability conditions on the involved distribution functions. In our setting, we find, for example condition c.1 in Section 4.5.1, which involves integrals with respect to hazard functions which are known to satisfy $\int \lambda_0(x)dx = \infty$, leading to a testing procedure with weak theoretical guarantees. While these conditions may hold for some models, it is not hard to find simple examples where they do not hold.

In order to get a more robust test, we exploit a well-known identity in survival analysis, allowing us to deduce a more natural Stein operator. Such an identity is given by

$$\mathbb{E}_0\left[\Delta\phi(T) - \int_0^T \phi(t)\lambda_0(x)dx\right] = 0, \tag{4.10}$$

which holds for any function ϕ such that $\mathbb{E}_0[|\phi(T)|] < \infty$ [Aalen et al., 2008]. This equality is derived by using a martingale identity that appears in the derivation of classical estimators in survival analysis (see Appendix 4.B).

Assuming $\lambda_0(t) > 0$, we replace $\phi = \omega'/\lambda_0$ in Eq. (4.10) to get

$$\mathbb{E}_0\left[\Delta\frac{\omega'(T)}{\lambda_0(T)} - (\omega(T) - \omega(0))\right] = 0.$$

Define the martingale Stein Operator $\mathcal{T}_0^{(m)} : \mathcal{H}^{(m)} \to \Omega$ as

$$(\mathcal{T}_0^{(m)}\omega)(x,\delta) = \delta\frac{\omega'(x)}{\lambda_0(x)} - (\omega(x) - \omega(0)) \tag{4.11}$$

where we write $\mathcal{H}^{(m)}$ instead of $\mathcal{H}$ whenever $\mathcal{H}$ satisfies

$$\int_{\mathbb{R}^+}\left|\frac{\omega'(x)}{\lambda_0(x)}\right| S_C(x)f_0(x)dx < \infty, \quad \forall\omega \in \mathcal{H}. \tag{4.12}$$

From its definition, it is clear that $\mathbb{E}_0[(\mathcal{T}_0^{(m)}\omega)(T,\Delta)] = 0$. Note that, by the definition of the hazard functions, condition in Eq. (4.12) is equivalent to

$$\int_{\mathbb{R}^+} |\omega'(x)|\, S_C(x)S_0(x)dx < \infty, \quad \forall\omega \in \mathcal{H}, \tag{4.13}$$

which holds true if the kernel is bounded (recall that we assume that $\omega' \in \mathcal{H}$).

Therefore, compared to $\mathcal{T}_0^{(s)}$, the testing procedure associated to $\mathcal{T}_0^{(m)}$ has very strong theoretical guarantees. Indeed, we observe that condition c.2 in Section 4.5.1 is much simpler to satisfy because, this time, we consider integrals with respect to the inverse of the hazard function.

Model-Free Implementation Inspired by the test of uniformity through MMD-based test statistic [Fernandez et al., 2019], we transform our data via the null model c.d.f. F_0 to obtain $U_i = F_0(T_i)$ and generate pairs (U_i, Δ_i). Notice that since F_0 is monotone $U_i = F_0(T_i) = \min\{F_0(X_i), F_0(C_i)\}$, thus Δ_i remains consistent. Under this transformation, testing the null hypothesis is equivalent to test whether $F_0(X_i)$ is distributed as a uniform random variable, thus, in this setting, $\lambda_0 = \lambda_{\mathcal{U}} = \frac{1}{1-x}$ and

$$(\mathcal{T}_0^{(m)}\omega)(u, \delta) = \delta\omega'(u)(1-u) - \omega(u) + \omega(0)$$

for $u = F_0(x)$ (notice that $F_0(0) = 0$). It will be shown in the experiments that this transformation is beneficial in terms of power performance. Similarly, we can exploit that $\Lambda_0(X) \sim \text{Exp}(1)$ under the null when the model is transformed via the cumulative hazard function, which is another monotonic function.

4.3.3 Proportional Stein Operator

In some scenarios, we are interested in the shape of the hazard function up to a multiplicative constant, i.e. $\lambda_0(t) = \gamma\lambda(t)$ where we know $\lambda(t)$ but not the constant γ. The family indexed by γ is called a proportional hazards family and it is one of the key objects of study in survival analysis. This object is fundamental because sometimes it is more important to test for qualitative results as "the hazard rate is growing at a constant speed", rather than obtaining precise values of the hazard function. If we only know $\lambda_X(t)$ up to constant and we can ensure that $\omega(0)\lambda(0) = 0$, then we can define a Stein operator based on unnormalised hazard.

In order to define our operator, we assume that

$$\int_{\mathbb{R}_+} |(\omega(x)\lambda_0(x))'|dx < \infty, \quad \text{and}$$
$$\omega(0)\lambda_0(0) = \lim_{x\to\infty} \omega(x)\lambda_0(x) = 0, \quad \forall\omega \in \mathcal{H}. \tag{4.14}$$

As usual, we write $\mathcal{H}^{(p)}$ to indicate that $\mathcal{H}$ satisfies property (4.14). Note that for any function $\omega \in \mathcal{H}^{(p)}$ it holds that

$$\int_0^\infty \frac{(\omega(x)\lambda_0(x))'}{\lambda_0(x)}\lambda_0(x)dx = 0.$$

The integral in the form above can be estimated using the Nelson-Aalen estimator in survival analysis [Nelson, 1972], leading to the statistic

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(\omega(T_i)\lambda_0(T))'}{\lambda_0(T_i)}\frac{\Delta_i}{Y(T_i)/n},$$

where $Y(t)=\sum_{k=1}^{n}\mathbb{1}_{\{T_k\geq t\}}$ is the so-called *risk function*, which counts the number of individuals at risk at time t. This suggests the following operator

$$(\widehat{\mathcal{T}}_0^{(p)}\omega)(x,\delta)=\left(\omega'(x)+\frac{\omega(x)\lambda_0'(x)}{\lambda_0(x)}\right)\frac{\delta}{Y(x)/n}. \tag{4.15}$$

In the definition above we use the notation $\widehat{\mathcal{T}}_0^{(p)}$ to indicate that, the function $Y(t)$ depends on all data points, hence $\widehat{\mathcal{T}}_0^{(p)}$ can be seen as an empirical estimator of a deterministic operator. Indeed, if $(T_i,\Delta_i)\sim\mu_0$, then

$$\frac{Y(x)}{n}\to S_C(x)S_0(x),$$

which indicates that under the null hypothesis, the operator $\widehat{\mathcal{T}}_0^{(p)}$ is similar to $\mathcal{T}_0^{(p)}$, given by

$$(\mathcal{T}_0^{(p)}\omega)(x,\delta)=\left(\omega'(x)+\frac{\omega(x)\lambda_0'(x)}{\lambda_0(x)}\right)\frac{\delta}{S_C(x)S_X(x)}.$$

This operator cannot be directly evaluated since we do not have access to S_C. The following proposition establishes the formal relation between $\widehat{\mathcal{T}}_0^{(p)}$ and $\mathcal{T}_0^{(p)}$.

Proposition 4.3. *Let $(T_i,\Delta_i)\sim\mu_0$, then for every $\omega\in\mathcal{H}^{(p)}$.*

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathcal{T}}_0^{(p)}\omega)(T_i,\Delta_i)\xrightarrow{p}\mathbb{E}_0\left[(\mathcal{T}_0^{(p)}\omega)(T_1,\Delta_1)\right]=0. \tag{4.16}$$

4.4 Censored-Data Kernel Stein Discrepancy

In this section, we derive censored-data Kernel Stein Discrepancies (c-KSD) using each of our three Stein operators defined in the previous section. The idea is to compare the largest discrepancy between two distributions f_X and f_0 over a class of test functions in the RKHS $\mathcal{H}$. Since we have access to censored data, we compare f_X and f_0 through the measures μ_X and μ_0, defined in Section 4.2.

We proceed to define three censored-data kernel Stein discrepancies: the survival Kernel Stein Discrepancy (s-KSD), the martingale Kernel Stein Discrepancy (m-KSD), and the proportional Kernel Stein Discrepancy (p-KSD) based on the

respective Stein operators $\mathcal{T}_0^{(s)}$, $\mathcal{T}_0^{(m)}$ and $\widehat{\mathcal{T}}_0^{(p)}$. In general, for any given Stein operator $\mathcal{T}_0^{(c)} : \mathcal{H}^{(c)} \to \Omega$ we define the c-KSD as

$$\text{c-KSD}(f_X \| f_0) = \sup_{\omega \in B_1(\mathcal{H}^{(c)})} \mathbb{E}_X[(\mathcal{T}_0{}^{(c)}\omega)(T, \Delta)]. \tag{4.17}$$

Denote by $K^{(c)}$ the reproducing kernel of $\mathcal{H}^{(c)}$. For any of the operators $\mathcal{T}_0^{(c)}$, applying $\mathcal{T}_0^{(c)}$ on $K^{(c)}(x, \cdot)$[2] is defined as $(\mathcal{T}_0^{(c)}\omega)(x, \delta)$ but replacing $\omega(x)$ by $K^{(c)}(x, \cdot)$ and $\omega'(x)$ by $\frac{\partial}{\partial x}K^{(c)}(x, \cdot)$. For example, for $c = m$, we get that

$$\left[(\mathcal{T}_0^{(m)}K^{(m)})(x, \delta)\right](\cdot) = \frac{\delta}{\lambda_0(x)}\left(\frac{\partial}{\partial x}K^{(m)}(x, \cdot)\right) - (K^{(m)}(x, \cdot) - K^{(m)}(0, \cdot)), \tag{4.18}$$

which is derived from Eq. (4.11).

Recall that for $c \in \{s, m, p\}$, we assumed that if $\omega \in \mathcal{H}^{(c)}$ then $\omega' \in \mathcal{H}^{(c)}$, and thus $\xi^{(c)}(x, \delta)(\cdot) = \left[(\mathcal{T}_0^{(c)}K^{(c)})(x, \delta)\right](\cdot) \in \mathcal{H}^{(c)}$ since all operators involve ω or ω'. Define the Stein kernel $h^{(c)} : (\mathbb{R}_+ \times \{0, 1\})^2 \to \mathbb{R}$ by

$$h^{(c)}((x, \delta), (x', \delta')) = \left\langle \xi^{(c)}(x, \delta), \xi^{(c)}(x', \delta') \right\rangle_{\mathcal{H}^{(c)}}.$$

By using the reproducing property in the form of Eq. (4.18), we can obtain a closed form expression for c-KSD for taking the supremum given by the following proposition.

Proposition 4.4. *For $c \in \{s, m, p\}$, and let (T, Δ) and (T', Δ') be independent samples from μ_X, and suppose that*

$$\mathbb{E}_X\left[\sqrt{h^{(c)}((T, \Delta), (T, \Delta)}\right] < \infty, \tag{4.19}$$

then

$$\text{c-KSD}(f_X \| f_0)^2 = \mathbb{E}_X\left[h^{(c)}((T, \Delta), (T', \Delta')\right].$$

The derivation is standard for taking supremum over functions in unit ball RKHS and the detailed derivations for the Stein kernels $h^{(c)}((x, \delta), (x', \delta'))$ can be found in Appendix 4.A.

[2]Recall that $K^{(c)}(x, \cdot)$ itself is a function $\mathbb{R}_+ \to \mathbb{R}$.

4.5 Goodness-of-fit Test via c-KSD

In this section, we study goodness-of-fit testing procedures based on c-KSD. We begin by estimating the squared c-KSD using the Stein kernel $h^{(c)}$,

$$\widehat{\text{c-KSD}}^2(f_X \| f_0) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} h^{(c)}((T_i, \Delta_i), (T_j, \Delta_j))$$

where (T_i, Δ_i) are independent samples from μ_X. By construction, under the null hypothesis where $\mu_X = \mu_0$, the estimator above should be close to zero, while under the alternative hypothesis when $\mu_X \neq \mu_0$, we expect it to be separated from zero.

4.5.1 Theoretical Analysis

We state some technical conditions that feature our analysis in order to establish the asymptotic behavior of $\widehat{\text{c-KSD}}^2$.

Technical Conditions

a) *Reproducing kernel conditions:* We assume that K has continuous second-order derivatives, and that $K(x, y)$ and $\frac{\partial^2}{\partial x \partial y} K(x, y)$ are bounded and C_0-universal Definition 3.1(i) [Carmeli et al., 2010, Definition 4.1].

b) *Boundary condition:* $\lim_{x \to 0+} \sqrt{K(x, x)} \lambda_0(x) < \infty$.

c) *Null integrability conditions:* Let $(T, \Delta), (T', \Delta') \overset{i.i.d.}{\sim} \mu_0$, and recall that $\mathbb{E}_0 = \mathbb{E}_{\mu_0}$. Depending on $c \in \{s, m, p\}$, we assume:

1) s-KSD:

   i) $\mathbb{E}_0[\phi(T, \Delta)^2 |K(T, T)|] < \infty$, and

   ii) $\mathbb{E}_0[\phi(T, \Delta)^2 \phi(T', \Delta')^2 K(T, T')^2] < \infty$,

   where $\phi(x, \delta) = \delta \frac{\lambda_0'(x)}{\lambda_0(x)} - \lambda_0(x)$.

2) m-KSD:

   i) $\mathbb{E}_0 \left[\frac{|K^\star(T,T)|\Delta}{\lambda_0(T)^2} \right] < \infty$, and

   ii) $\mathbb{E}_0 \left[\frac{K^\star(T,T')^2 \Delta\Delta'}{\lambda_0(T)^2 \lambda_0(T')^2} \right] < \infty$,

   where $K^\star(x, y) = \frac{\partial^2}{\partial x \partial y} K(x, y)$.

3) p-KSD:

   i) $\mathbb{E}_0 \left[\frac{|K^\star(T,T)|\Delta}{(f_0(T) S_C(T))^2} \right] < \infty$, and

ii) $\mathbb{E}_0\left[\frac{K^\star(T,T')^2\Delta\Delta'}{(f_0(T)f_0(T')S_C(T)S_C(T'))^2}\right] < \infty$,

where $K^\star(x,y) = \left(\frac{\partial^2}{\partial x \partial y}K(x,y)\lambda_0(x)\lambda_0(y)\right)$.

d) *Alternative integrability conditions:* Consider samples $(T,\Delta),(T',\Delta') \overset{i.i.d.}{\sim} \mu_X$. Then, for each $c \in \{s,m,p\}$ we assume:

1) s-KSD:

i) $\mathbb{E}_X[\phi(T,\Delta)^2|K(T,T)|] < \infty$,

where $\phi(x,\delta) = \delta\frac{\lambda_0'(x)}{\lambda_0(x)} - \lambda_0(x)$.

2) m-KSD:

i) $\mathbb{E}_X\left[\frac{|K^\star(T,T)|\Delta}{\lambda_0(T)^2}\right] < \infty$,

where $K^\star(x,y) = \frac{\partial^2}{\partial x\partial y}K(x,y)$.

3) p-KSD:

i) $\mathbb{E}_X\left[\frac{|K^\star(T,T)|\Delta}{S_T(T)^2\lambda_0(T)^2}\right] < \infty$,

where $K^\star(x,y) = \left(\frac{\partial^2}{\partial x\partial y}K(x,y)\lambda_0(x)\lambda_0(y)\right)$.

The following theorem establishes consistency of our empirical kernel Stein discrepancies to their population versions.

Theorem 4.1. *[Asymptotics under the alternative H_1] Let $c \in \{s,m,p\}$, and suppose that f_X satisfies conditions a), b), and the corresponding condition d). Then the following holds:*

$$\left(\widehat{\text{c-KSD}}(f_X\|f_0)\right)^2 \xrightarrow{p} \left(\text{c-KSD}(f_X\|f_0)\right)^2.$$

The previous theorem is not enough to ensure good behavior under the alternative as we need to be sure that the discrepancy of two different distribution functions f_X and f_0 is different from 0, regardless of censoring. We can prove this for c-KSD for $c \in \{s,m\}$. This does not hold true for p-KSD since it is designed to test if the hazard function λ_X is proportional to λ_0, and not for goodness-of-fit testing purposes. Indeed, whenever the hazards are in a proportional relation, p-KSD is 0.

Theorem 4.2. *Let $c \in \{s,m\}$. Assume $S_C(x) = 0$ implies $S_X(x) = 0$ and that K is C_0-universal. Then, under Conditions a), b) and d), $f_0 \neq f_X$ implies* $\text{c-KSD}(f_0\|f_X) > 0$.

Under the null distribution, $f_X = f_0$, we also have that $\widehat{\text{c-KSD}}(f_0 \| f_0) \to 0$, but we can prove an even stronger result that follows from the theory of V-statistics.

Theorem 4.3 (Asymptotics under the null H_0)**.** *Let $c \in \{s, m, p\}$, and suppose that $f_X = f_0$ and that conditions a), b), and the corresponding condition c) are satisfied. Then*

$$n \cdot \left(\widehat{\text{c-KSD}}(f_X \| f_0)\right)^2 \xrightarrow{\mathcal{D}} r_c + \mathcal{Y}_c,$$

where r_c is a constant and $\mathcal{Y}_c$ is an infinite sum of independent χ^2 random variables.

While Theorem 4.3 ensures the existence of a limiting null distribution, which implies that a rejection region for the test is well defined, in practice it is very hard to approximate the limit distribution and the corresponding rejection regions, for which, we rely on a wild bootstrap approach.

We remark that we can obtain concentrations bounds for the test-statistics under the null hypothesis if we assume that the kernels $h^{(s)}$ and $h^{(m)}$ are bounded, by using standard methods. Obtaining concentration bounds for $h^{(p)}$ is harder as it is a random kernel, depending on all data points.

4.5.2 Wild Bootstrap Tests

To re-sample from the null distribution we use the wild bootstrap technique [Dehling and Mikosch, 1994]. This technique is quite generic and it can be applied to any kernel-based testing procedure [Chwialkowski et al., 2014]. The wild bootstrap estimator is given by

$$\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_i W_j h^{(c)}((T_i, \Delta_i), (T_j, \Delta_j)), \tag{4.20}$$

where $W_1, \ldots, W_n$ are independent random variables from a common distribution $\mathcal{W}$ with $\mathbb{E}(W_1) = 0$ and $\mathbb{V}ar(W_1) = 1$. In our experiments we consider W_i sampled from a Rademacher distribution, but any distribution with the properties above is suitable. Dehling and Mikosch [1994] proved that if the limit distribution exists, in the sense of Theorem 4.3, then the wild bootstrap statistic also converges to the same limit distribution.

The testing procedure for goodness-of-fit is performed as follows: **1)** Set a type 1 error $\alpha \in (0, 1)$. **2)** Compute $\widehat{\text{c-KSD}}^2(f_X \| f_0)$ using our n data points. **3)** Compute m-independent copies of the wild bootstrap estimator from Eq. (4.20). **4)** Compute the proportion of wild bootstrap samples that are larger than $\widehat{\text{c-KSD}}^2(f_X \| f_0)$; if such a proportion is smaller than α we reject the null hypothesis, otherwise do not reject it.

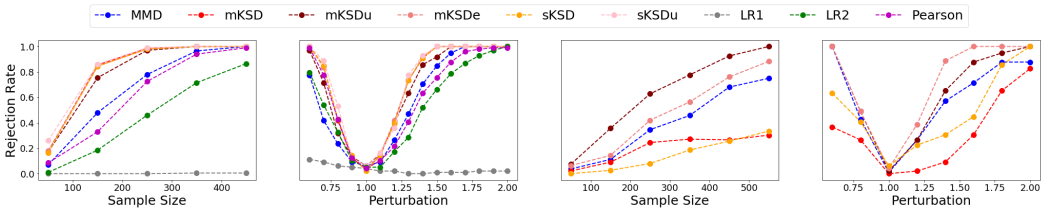


Figure 4.2: Rejection rate w.r.t. sample size and model perturbation. Left two for Weibull Hazard; Right two for Periodic Hazard. $\alpha = 0.01$.

4.6 Experiments

4.6.1 Simulation Results

Proposed approaches: We denote **sKSD**, **mKSD** and **pKSD** as the tests based on the survival, martingale and the proportional kernel Stein discrepancies respectively as described in Section 4.3, where the tests are implemented using the wild bootstrap approach as described in Section 4.5.2. Additionally, we implement the model free transformation described at the end of Section 4.3.2 and denote the c.d.f. transformation F_0 as suffix **u** and the cumulative hazard transformation Λ_0 as suffix **e**; e.g. **mKSDu** uses test statistic given by the **mKSD** test applied to the transformed data $((F_0(T_i), \Delta_i))_{i=1}^n$ to test the null hypothesis $H_0 : F_0(X) \sim \mathcal{U}(0,1)$; **mKSDe** uses test statistic given by the **mKSD** test applied to the transformed data $((\Lambda_0(T_i), \Delta_i))_{i=1}^n$ to test the null hypothesis $H_0 : \Lambda_0(X) \sim \text{Exp}(1)$. Finally, for the experiments, we use an exponentiated quadratic kernel with length-scale parameter chosen by using the median-heuristic, which is the median of all the absolute differences between two different data points. We did not use further optimisation for kernel parameters here to improve the performance of the tests. In this one-dimensional problem setting, median bandwidth is a simple choice that achieves good test power and the test is not super sensitive to kernel bandwidth. Optimise approximate test power [Gretton et al., 2012a; Jitkrittum et al., 2016a, 2017] results into less data for testing and test power can be reduced.

Competing approaches: We denote **MMD** as the maximum-mean-discrepancy approach proposed by Fernandez and Gretton [2019], which provides state-of-the-art results; **Pearson** denotes the Pearson χ^2 goodness-of-fit test proposed by Akritas [1988], which can be competitive in certain cases. **LR1** and **LR2** denote the weighted log-rank tests with respective weights functions $w_1(t) = 1$ and $w_2(t) = \sum_{i=1}^n \mathbb{1}_{\{T_i \geq t\}}$, which are classical tests, but not very competitive except for some very simple settings (e.g. testing $H_0 : \lambda_0(t) = 1$ against $\lambda_X(t) = c$, for constant $c \neq 1$).

Computational costs: As the c.d.f. transformation F_0 and Λ_0 can be computed explicitly and prior to computing the test statistics, **mKSDu** and **mKSDe** has the same order of computational time as **mKSD**. As the log-rank based test can be rewrite into the form of Eq.(4.17) dropping the sup notion and choosing fixed weights, **LR** test can be efficiently computed via wild bootstrap for null simulation and achieve the computational runtime of the same order of **mKSD**. The χ^2-based test only requires computation of statistics without simulating the null and has a computational advantage.

Data Sources: We begin by studying our proposed method via simulated distributions where we consider two data scenarios.

1. Weibull hazard functions

In our first experiment, we consider the Weibull model, which is commonly used in survival analysis [Bradburn et al., 2003]. The Weibull distribution is characterised by the density function of the form $f(x; k, r) = kr\,(rx)^{k-1} \exp\{-(rx)^k\}$, where k and r denote shape and rate parameters, respectively.

2. Periodic hazard functions

A much more interesting scenario is the so-called periodic hazards, which can be used to describe, for example, seasonal diseases such as Influenza. In this example, we consider the hazard function $\lambda_X(x) = 1 - \cos(\theta\pi x)$ studied in Fernandez and Gretton [2019]. The relevant functions f_X, F_X, S_X, Λ_X can be computed if necessary from Section 4.2. Note that when $\theta \to \infty$, then the distribution tends to a exponential of parameter 1. See Figure 4.1 for a comparison between the models.

Testing for Goodness-of-fit

In Figure 4.2, we show the rejection rate for testing the goodness-of-fit, where a particular density model is considered in the null hypothesis. For both models, we investigate the performance of our test in two setting: *increasing sample size* and *perturbations from the null*.

Increasing sample size shown in first and third plot in Figure 4.2 demonstrates how fast the test power converges to 1 w.r.t. the sample size increases. In the Weibull setting we set the null $H_0 : f_0(x) = f(x; 1, 1)$, where the alternative as $f_X(x) = f(x; 1.5, 1)$ and in the periodic setting, we consider the null $H_0 : f_0(x) = e^{-x}$, and generate data from the alternative $\theta = 3$. In both settings we consider 30% of censored data points

In both the Weibull cases, the KSD-based tests outperforms the competitors in terms of test power. In the periodic hazard case which is a harder problem, the KSD-based tests performs better than MMD-based tests after transformation, while

less power without transformation. This is likely due to the difference of the models are not easy to estimate; however, transforming via exact functions (F_0 or Λ_0), will give better estimation of the population statistics over fixed test distribution ($\mathcal{U}$ or $\text{Exp}(1)$), than direct empirical estimation from the data.

Perturbations from the null shown in second and fourth plot in Figure 4.2 demonstrates how sensitive the test are able to capture the alternative when the alternative distribution is getting closer to the null distribution. For the Weibull data, we set $H_0 : f_0(x) = f(x; 1, 1)$ and consider Weibull alternatives $f_X(x) = f(x; k, 1)$ with $k \in (0, \ldots, 2]$. Note that we recover the null hypothesis when $k = 1$. Also, we consider a constant 30% of censored observations and a fixed sample size of $n = 100$. For the periodic experiment we set $H_0 : \lambda_0(x) = 1 - \cos(\pi x)$, which is recovered when we take $\theta = 1$. In this case, we consider alternatives $\theta \in [0.5, 2] \backslash \{1\}$. We consider, again, a constant 30% of censoring, and a fixed sample size of $n = 100$.

For Weibull case, the kernel based methods outperform those non-kernel based methods, under the alternatives; in the sense that they achieve higher power with smaller departure from the null. Moreover, our KSD-based tests perform better than the MMD-based test. In Weibull case which is relatively simple, all KSD-based tests achieves similar results, which is expected. In the more challenging problem of periodic hazard, KSD-based tests with transformations outperforms MMD-based test, the similar trend as in third plot of Figure 4.2.

Testing for Hazard Proportionality

We consider the case of non-normalised models with unnormalised hazard functions. Unlike density functions which integrates to one, the hazard function does not need to. Hence, while testing unnormalised density is asking whether the data comes from a particular density model, to test unnormalised hazards is equivalent to test for a class of models with the same hazard shape, which is referred as testing hazard proportionality. We perform the test with the composite null hypothesis with unnormalised hazards via **pKSD**.

We present the result using the Weibull model with the shape parameter k and scale parameter l: when fixing shape parameter and varying the scale parameter, the hazard function remains the same up to multiplicative constants, i.e. has the same hazard shape; however, when the shape parameter changes, the hazard functions are not the same up to multiplicative constants. From Table 4.1, we can see that, when the shape of the Weibull model is different, the unnormalised methods correctly reject the null. However, the unnormalised method (with **pKSD**) have less test power compared to the goodness-of-fit test methods **mKSDu** and **sKSDu**. A

$\ell = 1.0$	$k = 2.0$			$k = 1.5$			$k = 0.8$		
α	sKSDu	mKSDu	pKSD	sKSDu	mKSDu	pKSD	sKSDu	mKSDu	pKSD
10 %	99.67	98.67	79.33	66.33	57.33	46.33	81.33	82.00	47.00
5 %	97.00	94.00	54.67	59.67	41.33	26.00	68.00	66.33	35.00
1 %	89.00	78.00	10.33	23.00	18.67	7.33	27.33	27.00	19.33

Table 4.1: Rejection Rate (in %) with shape (k) perturbation: the null distribution Weibull($k = 1.2, l = 1.0$), sample size 100, censoring rate: 30%

$k = 1.2$	$\ell = 2.0$			$\ell = 1.5$			$H_0 :$ $\ell = 1.0$		
α	sKSDu	mKSDu	pKSD	sKSDu	mKSDu	pKSD	sKSDu	mKSDu	pKSD
10 %	91.00	100.00	15.00	82.00	94.00	12.00	14.00	7.00	13.00
5 %	79.00	98.00	8.00	71.00	88.00	6.00	5.00	4.00	6.00
1 %	67.00	88.00	3.00	57.00	71.00	3.00	2.00	1.00	2.00

Table 4.2: Rejection Rate (in %) with scale (ℓ) difference: the null distribution Weibull($k = 1.2, \ell = 1.0$), sample size 100, censoring rate: 30%

possible explanation lies in the fact that, since this method tests against the whole model class, it must ignore all differences within this class, which affects the power of the test. In Table 4.2, as changing the scale parameter keeps the same hazard function up to a multiplicative constant, the unnormalised hazard approach does not reject and achieves well-controlled Type-I errors. On the other hand, as the models are different, **sKSDu** and **mKSDu** reject the null hypothesis. This result enables us to efficiently test a class of distribution without first estimating the best-fit model.

4.6.2 Real Data Applications

We perform our tests on the following real datasets to check relevant model assumptions. **aml:** **A**cute **M**yelogenous **L**eukemia survival dataset [Miller Jr, 2011]; **cgd:** **C**hronic **G**ranulotamous **D**isease dataset [Fleming and Harrington, 2011]; **ovarian:** Ovarian Cancer Survival dataset [Edmonson et al., 1979]; **lung:** North Central Cancer Treatment Group (NCCTG) Lung Cancer dataset [Loprinzi et al., 1994]; **stanford:** Stanford Heart Transplant Data [Crowley and Hu, 1977]; **nafld:** Non-alcohol fatty liver disease (NAFLD) [Allen et al., 2018].

Test Results We apply our proposed tests on real dataset for the Testing hazard proportionality and Goodness-of-fit settings. First, we check model class assumption using **pKSD** to test whether the observed data is from a desired family model without fitting model parameters. We check the exponential model class and the Weibull model with shape equals to 2. As the results shown in Table 4.3, our tests does not reject the Exponential model, which is coherent with scientific domain knowledge from the literature.[3]

[3]High-grade serous ovarian carcinoma (HG-SOC) is a major cause of cancer-related death. The growth of HG-SOC acts as an indicator of survival time of ovarian cancer [Gu et al., 2019]. This

p-value	**aml**	**cgd**	**ovarian**
Exponential	0.585	0.460	0.681
Weibull: $k = 2$	0.001	0.002	0.063

Table 4.3: Rejection rate for real data applications on testing hazard proportionality

Dataset	Covarites	p-value
lung	Age	0.167
stanford	T5 mismatch score	0.594
nafld	Weight and Gender	0.108

Table 4.4: Rejection rate on real data applications on testing goodness-of-fit

For the Goodness-of-fit test setting, we fit a cox proportional hazard model from the covariates provided in the datasets. The cox-proportional hazard function has the form $\lambda_X(x_i) = \lambda_b(x_i)\exp(\beta Y_i)$, where $\lambda_b(x)$ is the base hazard and Y_i is the covariate for subject i. The procedure is done via splitting the data into training set and test sets. Fitting the cox proportional-hazard model is applied on the training sets and the test sets are used to perform the goodness-of-fit tests. Results in Table 4.4 shows that all the models does not reject the fitted cox proportional hazard models and validate the proportional hazard assumptions for relevant fitted models, which is coherent with scientific experience stated in the literature.[4]

paper also suggests that that HG-SOC follows exponential expansion, which implies exponentially distributed survival time of ovarian patient.

[4]Chansky et al. [2016] suggests that cox proportional hazard model is a reasonable tool among practitioners for **lung** dataset. [Crowley and Hu, 1977] suggests a fit for cox proportional hazard model for **stanford** dataset. [Allen et al., 2018] states that cox proportional hazards is often used to study the impact of NAFLD on incident metabolic syndrome or death.

Appendices

4.A Proofs and Derivations

Proofs of Section 4.3.1: Survival Stein Operator

Proof of Proposition 4.1

The proof is deliberately presented in an order to reveals the construction procedure for the survival Stein operator $\mathcal{T}_0^{(s)}$. Let $\omega \in \mathcal{H}^{(s)}$. Expanding Eq. (4.4) yields

$$\begin{aligned}(\mathcal{T}_0\omega)(x,\delta) &= \delta\frac{(\omega(x)S_C(x)f_0(x))'}{S_C(x)f_0(x)} + \omega(0)f_0(0)\\ &= \delta\omega'(x) + \delta\omega(x)\frac{(S_C(x)f_0(x))'}{S_C(x)f_0(x)} + \omega(0)f_0(0)\\ &= \delta\omega'(x) + \delta\omega(x)\left(\frac{f_0'(x)}{f_0(x)} - \lambda_C(x)\right) + \omega(0)f_0(0). \end{aligned} \tag{4.21}$$

Recall that $\lambda_C = f_C/S_C$ denotes the hazard function associated to the censoring times and $S_C = 1 - F_C$, so the final line holds.

Notice that, when taking expectation of Eq. (4.21) w.r.t. $\mathbb{E}_0$, the only unknown term is $\mathbb{E}_0\left[\Delta\omega(T)\lambda_C(T)\right]$, since λ_C is not available even under the null hypothesis. Nevertheless, by Eq. (4.1) and Eq. (4.2), a simple inspection shows that

$$\mathbb{E}_0\left[\Delta\omega(T)\lambda_C(T)\right] = \int_0^\infty \omega(x)\frac{f_C(x)}{S_C(x)}S_C(x)f_0(x)dx = \mathbb{E}_0[(1-\Delta)\omega(T)\lambda_0(T)].$$

Therefore, we can replace $\delta\omega(x)\lambda_C(x)$ by $(1-\delta)\omega(x)\lambda_0(x)$ in Eq. (4.21), obtaining

our survival Stein operator:

$$\begin{aligned}(\mathcal{T}_0^{(s)}\omega)(x,\delta) &= \delta\omega'(x) + \delta\omega(x)\frac{f_0'(x)}{f_0(x)} - (1-\delta)\omega(x)\lambda_0(x) + \omega(0)f_0(0) \\ &= \delta\omega'(x) + \delta\omega(x)\left(\frac{\lambda_0'(x)}{\lambda_0(x)} - \lambda_0(x)\right) - (1-\delta)\omega(x)\lambda_0(x) + \omega(0)f_0(0) \\ &= \delta\omega'(x) + \delta\omega(x)\frac{\lambda_0'(x)}{\lambda_0(x)} - \omega(x)\lambda_0(x) + \omega(0)\lambda_0(0).\end{aligned}$$

The second equality holds due to the identity $\frac{f_0'(x)}{f_0(x)} = \frac{\lambda_0'(x)}{\lambda_0(x)} - \lambda_0(x)$, where

$$\frac{\lambda_0'(x)}{\lambda_0(x)} = \frac{f_0'(x)}{S_0(x)\lambda_0(x)} + \frac{f_0(x)^2}{S_0(x)^2\lambda_0(x)} = \frac{f_0'(x)}{f_0(x)} + \lambda_0(x). \tag{4.22}$$

The last line utilises $S_0(0) = 1$. By construction, it holds $\mathbb{E}_0[(\mathcal{T}_0\omega)(T,\Delta)] = \mathbb{E}_0[(\mathcal{T}_0^{(s)}\omega)(T,\Delta)]$ for any $\omega \in \mathcal{H}^{(s)}$.

Proof of Proposition 4.2

By construction, we can write

$$(\mathcal{T}_0^{(s)}\omega)(T,\Delta) - (\mathcal{T}_0\omega)(T,\Delta) = \omega(T)\left[\left(\Delta\frac{\lambda_0'(T)}{\lambda_0(T)} - \lambda_0(T)\right) - \Delta\left(\frac{f_0'(T)}{f_0(T)} - \lambda_C(T)\right)\right].$$

Plug in Eq. (4.22), we have

$$(\mathcal{T}_0^{(s)}\omega)(T,\Delta) - (\mathcal{T}_0\omega)(T,\Delta) = \omega(T)\left(\Delta\lambda_C(T) - (1-\Delta)\lambda_0(T)\right).$$

Then we write the empirical version as

$$\begin{aligned}&\sup_{\omega\in B_1(\mathcal{H})}\frac{1}{n}\sum_{i=1}^n (\mathcal{T}_0^{(s)}\omega)(T_i,\Delta_i) - (\mathcal{T}_0\omega)(T_i,\Delta_i) \\ &= \sup_{\omega\in B_1(\mathcal{H})}\frac{1}{n}\sum_{i=1}^n \omega(T_i)\left(\Delta_i\lambda_C(T_i) - (1-\Delta_i)\lambda_0(T_i)\right) \\ &= \sup_{\omega\in B_1(\mathcal{H})}\left\langle \omega, \frac{1}{n}\sum_{i=1}^n K(T_i,\cdot)\left(\Delta_i\lambda_C(T_i) - (1-\Delta_i)\lambda_0(T_i)\right)\right\rangle_{\mathcal{H}} \\ &= \left\|\frac{1}{n}\sum_{i=1}^n K(T_i,\cdot)\left(\Delta_i\lambda_C(T_i) - (1-\Delta_i)\lambda_0(T_i)\right)\right\|_{\mathcal{H}}\end{aligned}$$

The last line follows from standard trick for taking supremum over unit ball RKHS. We continue by proving that the previous norm converges to zero in probability. Ob-

serve that by the symmetrisation lemma [Vershynin, 2018, Lemma 6.4.2], it holds

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} K(T_i,\cdot)\left(\Delta_i\lambda_C(T_i)-(1-\Delta_i)\lambda_0(T_i)\right)\right\|_{\mathcal{H}}\right] \leq$$
$$2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} W_iK(T_i,\cdot)\left(\Delta_i\lambda_C(T_i)-(1-\Delta_i)\lambda_0(T_i)\right)\right\|_{\mathcal{H}}\right]$$

where $W_1,\ldots,W_n$ are i.i.d. Rademacher random variables, independent of the data $(T_i,\Delta_i)_{i=1}^n$. Then, by Jensen's inequality, and by using that $\mathbb{E}(W_i)=0$, we conclude that the previous expression converges to zero in probability, as

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} W_iK(T_i,\cdot)\left(\Delta_i\lambda_C(T_i)-(1-\Delta_i)\lambda_0(T_i)\right)\right\|^2_{\mathcal{H}}\right] =$$
$$\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n} K(T_i,T_i)\left(\Delta_i\lambda_C(T_i)-(1-\Delta_i)\lambda_0(T_i)\right)^2\right] \to 0$$

a.s., where the limit result holds due the law of large numbers which can be applied under the condition in Eq. (4.9) and by assuming $|K(x,y)|\leq c_1$, as

$$\begin{aligned}
&\mathbb{E}\left[K(T_i,T_i)\left(\Delta_i\lambda_C(T_i)-(1-\Delta_i)\lambda_0(T_i)\right)^2\right]\\
&\leq c_1\mathbb{E}\left[\Delta_i\lambda_C(T_i)^2+(1-\Delta_i)\lambda_0(T_i)^2\right]\\
&= c_1\int_0^\infty \left(\lambda_C(x)^2S_C(x)f_0(x)+\lambda_0(x)^2S_0(x)f_C(x)\right)dx\\
&= c_1\int_0^\infty \left(\lambda_C(x)+\lambda_0(x)\right)f_C(x)f_0(x)dx<\infty.
\end{aligned}$$

Proofs of Section 4.3.3: Proportional Stein Operator

Proof of Proposition 4.3

We start by claiming that the following equation holds true for every $\omega\in\mathcal{H}^{(p)}$:

$$\frac{1}{n}\sum_{i=1}^{n}\left((\widehat{\mathcal{T}}_0^{(p)}\omega)(T_i,\Delta_i)-(\mathcal{T}_0^{(p)}\omega)(T_i,\Delta_i)\right)\xrightarrow{p}0. \tag{4.23}$$

Then, the main result follows from Eq. (4.23) by the law of large numbers and that

$$\begin{aligned}
\mathbb{E}_0\left[(\mathcal{T}_0^{(p)}\omega)(T_1,\Delta_1)\right] &= \int_0^\infty \frac{(\omega(t)\lambda_0(t))'}{\lambda_0(t)}\frac{1}{S_0(t)S_C(t)}S_C(t)f_0(t)dt\\
&= \int_0^\infty \frac{(\omega(t)\lambda_0(t))'}{\lambda_0(t)}\lambda_0(t)dt=0,
\end{aligned}$$

which follows from the definition of our operator (see Eq. (4.14)). We finish the proof by proving our claim in Eq. (4.23). Observe that

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^{n}\left((\widehat{\mathcal{T}}_0^{(p)}\omega)(T_i,\Delta_i)-(\mathcal{T}_0^{(p)}\omega)(T_i,\Delta_i)\right)\right|\\&\leq\frac{1}{n}\sum_{i=1}^{n}\frac{|(\omega(T_i)\lambda_0(T_i))'|}{\lambda_0(T_i)}\left|\frac{\Delta_i}{Y(T_i)/n}-\frac{\Delta_i}{S_T(T_i)}\right|,\end{aligned}\tag{4.24}$$

where $S_T(t) = S_C(t)S_0(t)$ holds under the null hypothesis. We proceed to prove that the previous sum tends to 0 in probability when n grows to infinity. Let $\varepsilon > 0$ and define $t_\varepsilon > 0$ as the infimum of all t such that $\int_t^\infty |(\omega(x)\lambda_0(x))'|\,dx < \varepsilon$. Notice that such t_ε is well-defined since $\int_0^\infty |(\omega(x)\lambda_0(x))'|\,dx < \infty$. We continue by splitting the sum in Eq. (4.24) into two regions, $\{T_i \leq t_\epsilon\}$ and $\{T_i < t_\epsilon\}$, obtaining that Eq. (4.24) equals

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\frac{|(\omega(T_i)\lambda_0(T_i))'|}{\lambda_0(T_i)}\left|\frac{\Delta_i}{Y(T_i)/n}-\frac{\Delta_i}{S_T(T_i)}\right|\mathbb{1}_{\{T_i\leq t_\varepsilon\}}\\&+\frac{1}{n}\sum_{i=1}^{n}\frac{|(\omega(T_i)\lambda_0(T_i))'|}{\lambda_0(T_i)}\left|\frac{\Delta_i}{Y(T_i)/n}-\frac{\Delta_i}{S_T(T_i)}\right|\mathbb{1}_{\{T_i>t_\varepsilon\}},\end{aligned}\tag{4.25}$$

and we prove that both sums tend to 0 in probability when n grows to infinity. We start with the first term. Observe that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\frac{|(\omega(T_i)\lambda_0(T_i))'|}{\lambda_0(T_i)}\left|\frac{\Delta_i}{Y(T_i)/n}-\frac{\Delta_i}{S_T(T_i)}\right|\mathbb{1}_{\{T_i\leq t_\varepsilon\}}\\&\leq\sup_{t\leq t_\epsilon}\left|\frac{1}{Y(t)/n}-\frac{1}{S_T(t)}\right|\frac{1}{n}\sum_{i=1}^{n}\frac{|(\omega(T_i)\lambda_0(T_i))'|}{\lambda_0(T_i)}\Delta_i\mathbb{1}_{\{T_i\leq t_\varepsilon\}}\\&=o_p(1),\end{aligned}$$

where the previous result holds since $\sup_{t\leq t_\epsilon}\left|\frac{1}{Y(t)/n}-\frac{1}{S_T(t)}\right| \to 0$ almost surely by the Glivenko-Cantelli Theorem, and since

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\frac{|(\omega(T_i)\lambda_0(T_i))'|}{\lambda_0(T_i)}\Delta_i\mathbb{1}_{\{T_i\leq t_\epsilon\}}\to\mathbb{E}\left[\frac{|(\omega(T_1)\lambda_0(T_1))'|}{\lambda_0(T_1)}\Delta_1\mathbb{1}_{\{T_1\leq t_\varepsilon\}}\right]\\&=\int_0^{t_\epsilon}\frac{|(\omega(t)\lambda_0(t))'|}{\lambda_0(t)}S_C(t)f_0(t)dt=\int_0^{t_\epsilon}|(\omega(t)\lambda_0(t))'|\,dt<\infty,\end{aligned}$$

where the last expression is finite due to Eq. (4.14).

For the second term in Eq. (4.25), Theorem 3.2.1. of [Gill, 1980] yields

$$\sup_{t\leq\tau_n}\left|1-\frac{Y(T_i)/n}{S_T(T_i)}\right| = O_p(1),$$

where $\tau_n = \max\{T_1,\ldots,T_n\}$, and, Lemma 2.7 of [Gill, 1983] yields $\sup_{t\leq\tau_n} nS_T(t)/Y(t) = O_p(1)$. Recall that $S_T(t) = S_0(t)S_C(t)$. From the previous results, we get

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\Delta_i\frac{|(\omega(T_i)\lambda_0(T_i))'|}{\lambda_0(T_i)}\left|\frac{1}{Y(T_i)/n}-\frac{1}{S_T(T_i)}\right|\mathbb{1}_{\{T_i>t_\varepsilon\}}\\
&=\frac{1}{n}\sum_{i=1}^{n}\Delta_i\frac{|(\omega(T_i)\lambda_0(T_i))'|}{\lambda_0(T_i)}\frac{1}{Y(T_i)/n}\left|1-\frac{Y(T_i)/n}{S_T(T_i)}\right|\mathbb{1}_{\{T_i>t_\varepsilon\}}\\
&=O_p(1)\frac{1}{n}\sum_{i=1}^{n}\Delta_i\frac{|(\omega(T_i)\lambda_0(T_i))'|}{\lambda_0(T_i)}\frac{1}{Y(T_i)/n}\mathbb{1}_{\{T_i>t_\varepsilon\}}\\
&=O_p(1)\frac{1}{n}\sum_{i=1}^{n}\Delta_i\frac{|(\omega(T_i)\lambda_0(T_i))'|}{\lambda_0(T_i)}\frac{1}{S_0(T_i)S_C(T_i)}\mathbb{1}_{\{T_i>t_\varepsilon\}}.
\end{aligned}$$

Now, notice that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\Delta_i\frac{|(\omega(T_i)\lambda_0(T_i))'|}{\lambda_0(T_i)}\frac{1}{S_0(T_i)S_C(T_i)}\mathbb{1}_{\{T_i>t_\varepsilon\}}\\
&\overset{a.s.}{\to}\mathbb{E}_0\left[\Delta_1\frac{|(\omega(T_1)\lambda_0(T_1))'|}{\lambda_0(T_1)}\frac{1}{S_0(T_1)S_C(T_1)}\mathbb{1}_{\{T_1>t_\varepsilon\}}\right]\\
&=\int_{t_\varepsilon}^{\infty}\frac{|(\omega(x)\lambda_0(x))'|}{\lambda_0(x)}\frac{f_0(x)S_C(x)}{S_0(x)S_C(x)}dx\\
&=\int_{t_\varepsilon}^{\infty}|(\omega(x)\lambda_0(x))'|dx<\varepsilon,
\end{aligned}$$

where the first equality holds by Eq. (4.1), and the last inequality comes from the definition of t_ε. Since we can choose $\varepsilon > 0$ as small as desired, we conclude the result.

Proofs Section 4.4: Censored-Data Kernel Stein Discrepancy

Proof of Proposition 4.4

Notice that, by the definition of the random function $\xi^{(c)}(\Delta, T)$, we have that $(\mathcal{T}^{(c)}\omega)(T,\Delta) = \langle\omega, \xi^{(c)}(T,\Delta)\rangle_{\mathcal{H}^{(c)}}$. Also notice that, $\xi^{(c)}(x,\delta) \in \mathcal{H}^{(c)}$ for each fixed (x,δ), and that the expectation, $\mathbb{E}_X\left[\xi^{(c)}(T,\Delta)\right] \in \mathcal{H}^{(c)}$ if and only if equation (4.19) is satisfied (the previous expectation has to be understood in the Bochner

sense, as we are taking expectation of a random function). Then,

$$\begin{aligned}
\text{c-KSD}(f_X \| f_0)^2 &= \sup_{\omega \in B_1(\mathcal{H}^{(c)})} \mathbb{E}_X \left[(\mathcal{T}_0^{(c)} \omega)(T, \Delta) \right]^2 \\
&= \sup_{\omega \in B_1(\mathcal{H}^{(c)})} \mathbb{E}_X \left[\left\langle \omega, \xi^{(c)}(T, \Delta) \right\rangle_{\mathcal{H}^{(c)}} \right]^2 \\
&= \sup_{\omega \in B_1(\mathcal{H}^{(c)})} \left\langle \omega, \mathbb{E}_X \left[\xi^{(c)}(T, \Delta) \right] \right\rangle_{\mathcal{H}^{(c)}}^2 \\
&= \left\| \mathbb{E}_X \left[\xi^{(c)}(T, \Delta) \right] \right\|_{\mathcal{H}^{(c)}}^2 \\
&= \left\langle \mathbb{E}_X \left[\xi^{(c)}(T, \Delta) \right], \mathbb{E}_X \left[\xi^{(c)}(T', \Delta') \right] \right\rangle_{\mathcal{H}^{(c)}} \\
&= \mathbb{E}_X \left[\left\langle \xi^{(c)}(T, \Delta), \xi^{(c)}(T', \Delta') \right\rangle_{\mathcal{H}^{(c)}} \right] \\
&= \mathbb{E}_X \left[h^{(c)}((T, \Delta), (T', \Delta')) \right],
\end{aligned}$$

where the third equality is due to the linearity of expectation and the inner product, the fourth equality follows from the definition of norm (and since we are taking supremum in the unit ball), and the second to last equality is, again, due to the linearity of the expectation and inner product.

Explicit computation of $h^{(c)}$

Denote $\phi(x, \delta) = \delta \frac{\lambda_0'(x)}{\lambda_0(x)} - \lambda_0(x)$, and $L_1(x, y) = \frac{\partial}{\partial x} K^{(c)}(x, y)$, $L_2(x, y) = \frac{\partial}{\partial y} K^{(c)}(x, y)$ and $L = \frac{\partial^2}{\partial x \partial y} K^{(c)}(x, y)$. For simplicity of exposition, we will drop the superscript (c) in all cases.

Survival Stein operator $(c = s)$: For this case, we have

$$\begin{aligned}
\xi(x, \delta) &= (\mathcal{T}_0 K)((x, \delta), \cdot) \\
&= \delta \frac{\partial}{\partial x} K(x, \cdot) + \left(\delta \frac{\lambda_0'(x)}{\lambda_0(x)} - \lambda_0(x) \right) K(x, \cdot) + \lambda_0(0) K(0, \cdot) \\
&= \delta L_1(x, \cdot) + \phi(x, \delta) K(x, \cdot) + \lambda_0(0) K(0, \cdot).
\end{aligned}$$

Notice that a simple computation shows that $L(x, y) = \langle L_1(x, \cdot), L_1(y, \cdot) \rangle_{\mathcal{H}}$, then

$$\begin{aligned}
h^{(s)}((x, \delta), (x', \delta')) &= \delta \delta' L(x, x') + \delta \phi(x', \delta') L_1(x, x') + \delta \lambda_0(0) L_1(x, 0) \\
&+ \phi(x, \delta) \delta' L_2(x, x') + \phi(x, \delta) \phi(x', \delta') K(x, x') + \phi(x, \delta) \lambda_0(0) K(x, 0) \\
&+ \lambda_0(0) \delta' L_2(0, x') + \lambda_0(0) \phi(x', \delta') K(0, x') + \lambda_0(0)^2 K(0, 0).
\end{aligned}$$

Martingale Stein operator $(c = m)$: Observe that in this case

$$\xi(x,\delta) = (\mathcal{T}_0 K)((s,\delta),\cdot) = \frac{\delta}{\lambda_0(x)} L_1(x,\cdot) - K(x,\cdot) + K(0,\cdot).$$

Then, by the reproducing kernel property

$$\begin{aligned} h^{(m)}(x,\delta),(x',\delta')) = &\frac{\delta}{\lambda_0(x)}\frac{\delta'}{\lambda_0(x')} L(x,x') - \frac{\delta}{\lambda_0(x)} L_1(x,x') + \frac{\delta}{\lambda_0(x)} L_1(x,0) \\ &- \frac{\delta'}{\lambda_0(x')} L_2(x,x') + K(x,x') - K(x,0) \\ &+ \frac{\delta'}{\lambda_0(x')} L_2(0,x') - K(0,x') + K(0,0). \end{aligned}$$

Proportional Stein operator $(c = p)$: Notice that, in this case, we use $\widehat{\mathcal{T}}_0^{(p)}$, given in Eq. (4.15), to compute $\widehat{\xi}^{(p)}(x,\delta) = (\widehat{\mathcal{T}}_0^{(p)} K^{(p)})((x,\delta),\cdot)$ since $\mathcal{T}_0^{(p)}$ is not available, as it depends on S_C, which is unknown even under the null hypothesis. Then,

$$\widehat{\xi}(x,\delta) = (\widehat{\mathcal{T}}_0 K)((x,\delta),\cdot) = \left(L_1(x,\cdot) + \frac{\lambda_0'(x)}{\lambda_0(x)} K(x,\cdot) \right) \frac{\delta}{Y(x)/n}.$$

Define $K^\star(x,y) = \left(\frac{\partial^2}{\partial x \partial y} \lambda_0(x)\lambda_0(y) K(x,y) \right)$. Then, by the reproducing property,

$$\widehat{h}^{(p)}((x,\delta),(x',\delta')) = n^2 \frac{\delta\delta'}{Y(x)Y(x')} K^\star(x,x').$$

Recall that $Y(t) = \sum_{k=1}^n \mathbb{1}_{\{T_k \geq t\}}$ denotes the risk function, which depends on all the data points, hence we write $\widehat{h}^{(p)}$ to remind that this kernel is a random one.

Proofs of Section 4.5: Goodness-of-fit via c-KSD

The following lemmas show that, under Conditions c) and d) in Section 4.5.1, depending in the case, the kernels $h^{(c)}$ have finite first and second moment. These moment conditions on the kernel are important to deduce asymptotic results.

Lemma 4.1. *Let (T',Δ') and (T,Δ) be independent samples from μ_X, and assume that Condition d) holds. Then,*

$$\mathbb{E}_X\left[|h^{(c)}((T,\Delta),(T,\Delta))|\right] < \infty, \quad \textit{and} \quad \mathbb{E}_X\left[|h^{(c)}((T,\Delta),(T',\Delta'))|\right] < \infty$$

for $c \in \{s,m,p\}$, under the alternative hypothesis.

Lemma 4.2. *Let (T',Δ') and (T,Δ) be independent samples from μ_0, and assume*

that Condition c) holds. Then

$$\mathbb{E}_0\left[|h^{(c)}((T,\Delta),(T,\Delta))|\right] < \infty, \quad \textit{and} \quad \mathbb{E}_0\left[h^{(c)}((T,\Delta),(T',\Delta'))^2\right] < \infty$$

for $c \in \{s, m, p\}$, under the null hypothesis.

We just proof Lemma 4.1 since the proof of Lemma 4.2 is essentially the same.

Proof of Lemma 4.1. First of all, note that for any kernel (positive-definite function), it holds

$$h^{(c)}((x,\delta),(x',\delta')) \leq \frac{1}{2}h^{(c)}((x,\delta),(x,\delta)) + \frac{1}{2}h^{(c)}((x',\delta'),(x',\delta')),$$

hence, it is enough to only prove the first part of the lemma.

Survival Stein operator $(c = s)$:
Recall $\xi^{(s)}(x,\delta) = \delta L_1(x,\cdot) + \phi(x,\delta)K(x,\cdot) + \lambda_0(0)K(0,\cdot)$, where $L_1(x,y) = \frac{\partial}{\partial x}K(x,y)$ $\phi(x,\delta) = \delta\frac{\lambda_0'(x)}{\lambda_0(x)} - \lambda_0(x)$, then

$$\begin{aligned}
&\mathbb{E}_X\left[|h^{(s)}((T,\Delta),(T,\Delta))|\right] = \mathbb{E}_X\left[\left\|\xi^{(s)}(T,\Delta)\right\|^2_{\mathcal{H}^{(s)}}\right] \\
&\leq 4\mathbb{E}_X\left[\|\Delta L_1(T,\cdot)\|^2_{\mathcal{H}^{(s)}} + \|\phi(T,\Delta)K(T,\cdot)\|^2_{\mathcal{H}^{(s)}}\right] + 4\left\|\lambda_0(0)K(0,\cdot)\right\|^2_{\mathcal{H}^{(s)}} \\
&\leq 4\mathbb{E}_X\left[\|\Delta L_1(T,\cdot)\|^2_{\mathcal{H}^{(s)}}\right] + 4\mathbb{E}_X\left[\|\phi(T,\Delta)K(T,\cdot)\|^2_{\mathcal{H}^{(s)}}\right] + 4\lambda_0(0)^2K(0,0).
\end{aligned}$$

The first and third term in the previous equation are finite under the technical Conditions a) and b). Thus, we only need to check

$$\mathbb{E}_X\left[\|\phi(T,\Delta)K(T,\cdot)\|^2_{\mathcal{H}^{(s)}}\right] = \mathbb{E}_X\left[\phi(T,\Delta)^2|K(T,T)|\right] < \infty,$$

which is guaranteed by Condition d).

Martingale Stein operator $(c = m)$:
Recall that $\xi^{(m)}(x,\delta) = \phi(x,\delta)L_1(x,\cdot) - K(x,\cdot) + K(0,\cdot)$, where $L_1(x,y) = \frac{\partial}{\partial x}K(x,y)$ and $\phi(x,\delta) = \frac{\delta}{\lambda_0(x)}$. Then

$$\begin{aligned}
&\mathbb{E}_X\left[|h^{(m)}((T,\Delta),(T,\Delta))|\right] = \mathbb{E}_X\left[\left\|\xi^{(m)}(T,\Delta)\right\|^2_{\mathcal{H}^{(m)}}\right] \\
&\leq 4\mathbb{E}_X\left[\|\phi(T,\Delta)L_1(T,\cdot)\|^2_{\mathcal{H}^{(s)}}\right] + 4\mathbb{E}\left[\|K(T,\cdot)\|^2_{\mathcal{H}^{(s)}}\right] + 4\left\|K(0,\cdot)\right\|^2_{\mathcal{H}^{(s)}}.
\end{aligned}$$

Observe that the second and third term are finite under Condition a). Additionally,

define $L(x,y) = \frac{\partial^2}{\partial x \partial y} K(x,y)$ and notice that

$$\mathbb{E}_X\left[\|\phi(T,\Delta)L_1(T,\cdot)\|^2_{\mathcal{H}^{(s)}}\right] = \mathbb{E}_X\left[\phi(T,\Delta)^2 L(T,T)\right] = \mathbb{E}_X\left[\frac{\Delta}{\lambda_0(T)^2}L(T,T)\right] < \infty$$

holds under Condition d) (Notice that $L = K^\star$ in Condition d.2)).

Proportional Stein operator ($c = p$).
This case follows directly from Condition d.3). □

Proof of Theorem 4.1

We distinguish between two cases: first, when $h^{(c)}$ is a deterministic kernel (that is $c \in \{s, m\}$), and second, when $\widehat{h}^{(c)}$ is a random kernel, meaning $c = p$.

Deterministic kernel ($c \in \{s, m\}$) :
For the first case, we have

$$\widehat{\text{c-KSD}}^2(f_X||f_0) = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n h^{(c)}((T_i,\Delta_i),(T_j,\Delta_j)),$$

which is a V-statistic of order 2. Thus, by using the law of large numbers for V-statistics, we deduce

$$\widehat{\text{c-KSD}}^2(f_X||f_0) \overset{a.s.}{\to} \mathbb{E}_X\left(h^{(c)}((T,\Delta),(T',\Delta'))\right) = \text{c-KSD}^2(f_X||f_0),$$

as n grows to infinity. Notice that the previous limit result requires the following conditions: $\mathbb{E}_X\left(|h^{(c)}((T,\Delta),(T,\Delta))|\right) < \infty$ and $\mathbb{E}_X\left(|h^{(c)}((T,\Delta),(T',\Delta'))|\right) < \infty$, which are satisfied under Condition d) by Lemma 4.1.

Random kernel ($c = p$) :
For the second case, recall that

$$\widehat{\text{p-KSD}}^2(f_X||f_0) = \sum_{i=1}^n\sum_{j=1}^n \widehat{h}^{(p)}((T_i,\Delta_i),(T_j,\Delta_j)), \tag{4.26}$$

where $\widehat{h}^{(p)}$ is a random kernel. Our first step will be to assume that we can replace the random kernel $\widehat{h}^{(p)}$, given by $\widehat{h}^{(p)}((x,\delta),(x',\delta')) = n^2\frac{\delta\delta' K^\star(x,x')}{Y(x)Y(x')}$, by its limit $h^{(p)}((x,\delta),(x',\delta')) = \frac{\delta\delta' K^\star(x,x')}{S_T(x)S_T(x')}$, where $K^\star(x,y) = \left(\frac{\partial^2}{\partial x\partial y}K(x,y)\lambda_0(x)\lambda_0(y)\right)$.

We claim that

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\widehat{h}^{(p)}((T_i,\Delta_i),(T_j,\Delta_j)) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}h^{(p)}((T_i,\Delta_i),(T_j,\Delta_j)) + o_p(1), \tag{4.27}$$

and then we have that

$$\begin{aligned}
\widehat{\text{p-KSD}}^2(f_X||f_0) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\widehat{h}^{(p)}((T_i,\Delta_i),(T_j,\Delta_j)) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}h^{(p)}((T_i,\Delta_i),(T_j,\Delta_j)) + o_p(1) \\
&= \mathbb{E}_X(h^{(p)}((T,\Delta),(T',\Delta'))) + o_p(1) \\
&= \text{p-KSD}^2(f_X||f_0) + o_p(1),
\end{aligned}$$

where the third equality is due to the standard law of large numbers for V statistics, and by Condition d.3) and Lemma 4.1.

We finish the proof by proving the claim made in Eq. (4.27). Recall that

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\widehat{h}^{(p)}((T_i,\Delta_i),(T_j,\Delta_j)) = \left\|\frac{1}{n}\sum_{i=1}^{n}\widehat{\xi}^{(p)}(T_i,\Delta_i)\right\|^2_{\mathcal{H}^{(p)}},$$

and

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}h^{(p)}((T_i,\Delta_i),(T_j,\Delta_j)) = \left\|\frac{1}{n}\sum_{i=1}^{n}\xi^{(p)}(T_i,\Delta_i)\right\|^2_{\mathcal{H}^{(p)}}, \tag{4.28}$$

where $\widehat{\xi}^{(p)}(x,\delta) = n\frac{(K(x,\cdot)\lambda_0(x))'}{\lambda_0(x)}\frac{\delta}{Y(x)}$ and $\xi^{(p)}(x,\delta) = \frac{(K(x,\cdot)\lambda_0(x))'}{\lambda_0(x)}\frac{\delta}{S_T(x)}$. Then, by the triangular inequality, and by taking square[5], the claim in Eq. (4.27) follows from proving:

i) $\left\|\frac{1}{n}\sum_{i=1}^{n}\widehat{\xi}^{(p)}(T_i,\Delta_i) - \xi^{(p)}(T_i,\Delta_i)\right\|_{\mathcal{H}^{(p)}} = o_p(1)$, and

ii) $\left\|\frac{1}{n}\sum_{i=1}^{n}\xi^{(p)}(T_i,\Delta_i)\right\|_{\mathcal{H}^{(p)}} = O_p(1)$.

Notice that item ii) holds trivially by Eq. (4.28), and by the law of large numbers for V-statistics, which can be applied due to Lemma 4.1, under Condition d). We finish by proving the result in item i). Following the same steps used in Eq. (4.24),

[5]notice that $\|b\| - \|a-b\| \leq \|a\| \leq \|b\| + \|a-b\|$

we have that

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \widehat{\xi}^{(p)}(T_i,\Delta_i) - \xi^{(p)}(T_i,\Delta_i) \right\|_{\mathcal{H}^{(p)}} \tag{4.29}$$

$$= \left\| \frac{1}{n}\sum_{i=1}^{n} \frac{(K(T_i,\cdot)\lambda_0(T_i))'}{\lambda_0(T_i)} \left(\frac{\Delta_i}{Y(T_i)/n} - \frac{\Delta_i}{S_T(T_i)} \right) \right\|_{\mathcal{H}^{(p)}}$$

$$= \sup_{\omega \in B_1(\mathcal{H}^{(p)})} \frac{1}{n}\sum_{i=1}^{n} \frac{(\omega(T_i)\lambda_0(T_i))'}{\lambda_0(T_i)} \left(\frac{\Delta_i}{Y(T_i)/n} - \frac{\Delta_i}{S_T(T_i)} \right)$$

$$\leq \sup_{\omega \in B_1(\mathcal{H}^{(p)})} \frac{1}{n}\sum_{i=1}^{n} \frac{(\omega(T_i)\lambda_0(T_i))'}{\lambda_0(T_i)} \left(\frac{\Delta_i}{Y(T_i)/n} - \frac{\Delta_i}{S_T(T_i)} \right) \mathbb{1}_{\{T_i \leq t_\varepsilon\}} \tag{4.30}$$

$$+ \sup_{\omega \in B_1(\mathcal{H}^{(p)})} \frac{1}{n}\sum_{i=1}^{n} \frac{(\omega(T_i)\lambda_0(T_i))'}{\lambda_0(T_i)} \left(\frac{\Delta_i}{Y(T_i)/n} - \frac{\Delta_i}{S_T(T_i)} \right) \mathbb{1}_{\{T_i > t_\varepsilon\}}, \tag{4.31}$$

where $\varepsilon > 0$ and $t_\varepsilon > 0$, and t_ϵ is the infimum over all $t > 0$ such that

$$\int_t^\infty \int_t^\infty \frac{|K^\star(t,s)|}{\lambda_0(t)\lambda_0(s)S_T(t)S_T(s)} S_C(t)S_C(s)f_X(t)f_X(s)dtds \leq \varepsilon.$$

Notice that such a t_ϵ is well-defined by Lemma 4.1 and Condition d.3). For the term in Eq. (4.30), observe that

$$\left(\sup_{\omega \in B_1(\mathcal{H}^{(p)})} \frac{1}{n}\sum_{i=1}^{n} \frac{(\omega(T_i)\lambda_0(T_i))'}{\lambda_0(T_i)} \left(\frac{\Delta_i}{Y(T_i)/n} - \frac{\Delta_i}{S_T(T_i)} \right) \mathbb{1}_{\{T_i \leq t_\varepsilon\}} \right)^2 \tag{4.32}$$

$$\leq \sup_{t \leq t_\varepsilon} \left(\frac{1}{Y(t)/n} - \frac{1}{S_T(t)} \right)^2 \frac{1}{n^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_i\Delta_j \frac{K^\star(T_i,T_j)}{\lambda_0(T_i)\lambda_0(T_j)} \mathbb{1}_{\{T_i \leq t_\varepsilon\}} \mathbb{1}_{\{T_j \leq t_\varepsilon\}}$$

$$= o_p(1) \tag{4.33}$$

where the last line holds since $\sup_{t \leq t_\epsilon} \left| \frac{1}{Y(t)/n} - \frac{1}{S_T(t)} \right| = o_p(1)$ a.s., by an application of Glivenko-Cantelli, and since the double sum converges to

$$\mathbb{E}\left(\Delta_1\Delta_2 \frac{K^\star(T_1,T_2)}{\lambda_0(T_1)\lambda_0(T_2)} \mathbb{1}_{\{T_1 \leq t_\varepsilon\}} \mathbb{1}_{\{T_2 \leq t_\varepsilon\}} \right),$$

which is finite by Lemma 4.1 and Condition d.3).

Finally, we prove that the term in Eq. (4.31) is $o_p(1)$. Define $R(t) = \left| \frac{S_T(t)}{Y(t)/n} - 1 \right|$. Gill [1983] proved that $\sup_{t \leq \tau_n} R(t) = O_p(1)$ where $\tau_n = \max\{T_1, \ldots, T_n\}$. By using this result, the term in Eq. (4.31) satisfies

$$
\begin{aligned}
&\left(\sup_{\omega \in B_1(\mathcal{H}^{(p)})} \frac{1}{n} \sum_{i=1}^{n} \frac{(\omega(T_i)\lambda_0(T_i))'}{\lambda_0(T_i)} \left(\frac{\Delta_i}{Y(T_i)/n} - \frac{\Delta_i}{S_T(T_i)} \right) \mathbb{1}_{\{T_i > t_\varepsilon\}} \right)^2 \\
&\leq \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\Delta_i \Delta_j |K^\star(T_i, T_j)|}{\lambda_0(T_i)\lambda_0(T_j) S_T(T_i) S_T(T_j)} R(T_i) R(T_j) \mathbb{1}_{\{T_i > t_\varepsilon\}} \mathbb{1}_{\{T_j > t_\varepsilon\}} \\
&= O_p(1) \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\Delta_i \Delta_j |K^\star(T_i, T_j)|}{\lambda_0(T_i)\lambda_0(T_j) S_T(T_i) S_T(T_j)} \mathbb{1}_{\{T_i > t_\varepsilon\}} \mathbb{1}_{\{T_j > t_\varepsilon\}} \\
&= O_p(1) \int_{t_\varepsilon}^{\infty} \int_{t_\varepsilon}^{\infty} \frac{|K^\star(t, s)|}{\lambda_0(t)\lambda_0(s) S_T(t) S_T(s)} S_C(t) S_C(s) f_X(t) f_X(s) dt ds \\
&= O_p(1)\varepsilon,
\end{aligned}
$$

where, in the second line, we use that $\sup_{t \leq \tau_n} R(t) = O_p(1)$, and in the fourth line we use the law of large numbers, and the definition of t_ε. Since ε is arbitrary, we conclude that equation (4.31) tends to 0 in probability.

Proof of Theorem 4.2

Survival Stein operator (c=s):
We proceed by contradiction. Assume that $f_X \neq f_0$ but

$$
\text{c-KSD}(f_X \| f_0) = \sup_{\omega \in B_1(\mathcal{H}^{(s)})} \mathbb{E}_X((\mathcal{T}_0^{(s)}\omega)(T, \Delta)) = 0.
$$

Recall that

$$
\begin{aligned}
&\mathbb{E}_X((\mathcal{T}_0^{(s)}\omega)(T, \Delta)) = \mathbb{E}_X((\mathcal{T}_0\omega)(T, \Delta)) \\
&= \mathbb{E}_X \left[\Delta\omega'(T) + \Delta\omega(T) \frac{f_0'(T)}{f_0(T)} - \Delta\omega(T)\lambda_C(T) \right] + \omega(0) f_0(0).
\end{aligned}
$$

Similarly, define

$$
(\mathcal{T}_X\omega)(x, \delta) = \delta\omega'(x) + \delta\omega(x) \frac{f_X'(x)}{f_X(x)} - \delta\omega(x)\lambda_C(x) + \omega(0) f_X(0),
$$

and notice that $\mathbb{E}_X((\mathcal{T}_X\omega)(T,\Delta)) = 0$ by the Stein's identity. Then

$$\begin{aligned}\mathbb{E}_X\left((\mathcal{T}_0^{(s)}\omega)(T,\Delta)\right) &= \mathbb{E}_X\left((\mathcal{T}_0\omega)(T,\Delta)\right)\\ &= \mathbb{E}_X\left((\mathcal{T}_0\omega)(T,\Delta) - (\mathcal{T}_X\omega)(T,\Delta)\right)\\ &= \mathbb{E}_X\left(\Delta\omega(T)\left(\frac{f_0'(T)}{f_0(T)} - \frac{f_X'(T)}{f_X(T)}\right) + \omega(0)(f_0(0) - f_X(0))\right)\\ &= \mathbb{E}_X\left(\Delta\omega(T)\left(\log\frac{f_0(T)}{f_X(T)}\right)'\right) + \omega(0)(f_0(0) - f_X(0)),\end{aligned}$$

and thus

$$\begin{aligned}\text{s-KSD}(f_X\|f_0) &= \sup_{\omega\in B_1(\mathcal{H}^{(s)})} \mathbb{E}_X((\mathcal{T}_0^{(s)}\omega)(T,\Delta))\\ &= \sup_{\omega\in B_1(\mathcal{H}^{(s)})} \mathbb{E}_X\left(\Delta\omega(T)\left(\log\frac{f_0(T)}{f_X(T)}\right)'\right) + \omega(0)(f_0(0) - f_X(0))\\ &= \sup_{\omega\in B_1(\mathcal{H}^{(s)})} \left\langle \omega, \int_0^\infty K(x,\cdot)d\nu(x)\right\rangle\\ &= \left\|\int_0^\infty K(x,\cdot)d\nu(x)\right\|_{\mathcal{H}^{(s)}} = 0,\end{aligned}$$

where $d\nu(x) = \left(\log\frac{f_0(x)}{f_X(x)}\right)' S_C(x)f_X(x)dx + (f_0(x) - f_X(x))\delta_0(x)$, and where we identify $\int_0^\infty K(x,\cdot)d\nu(x)$ as the mean kernel embedding of the measure ν. We shall assume that the above embedding is well-defined, otherwise we have $\text{s-KSD}(f_X\|f_0) \neq 0$. Since the kernel is C_0-universal, the previous set of equations implies ν is the zero measure, which implies that $f_0(0) = f_X(0)$, and

$$\left(\log\frac{f_0(x)}{f_X(x)}\right)' = 0, \tag{4.34}$$

as long as $f_X(x) > 0$ implies $S_C(x)f_X(x) > 0$ (which does, since we assume $S_C(x) = 0$ implies $S_X(x) = \int_x^\infty f_X(x)dx = 0$). Eq. (4.34) yields $f_0 \propto f_X$ and $f_X = f_0$ since both, f_0 and f_X, are probability density functions. This finalises our proof.

Martingale Stein operator (c=m):
Write the martingale Stein operator w.r.t. μ_X:

$$(\mathcal{T}_X^{(m)}\omega)(x,\delta) = \omega'(x)\frac{\delta}{\lambda_X(x)} - (\omega(x) - \omega(0)),$$

and notice that $\mathbb{E}_X[(\mathcal{T}_X^{(m)}\omega)(T,\Delta)] = 0$ from the martingale identity. Observe that

$$\begin{aligned}
\text{m-KSD}(f_X\|f_0) &= \sup_{\omega\in B_1(\mathcal{H}^{(m)})} \mathbb{E}_X((\mathcal{T}_0^{(m)}\omega)(T,\Delta)) \\
&= \sup_{\omega\in B_1(\mathcal{H}^{(m)})} \mathbb{E}_X((\mathcal{T}_0^{(m)}\omega)(T,\Delta)) - \mathbb{E}_X((\mathcal{T}_X^{(m)}\omega)(T,\Delta)) \\
&= \sup_{\omega\in B_1(\mathcal{H}^{(m)})} \mathbb{E}_X\left(\omega'(T)\Delta\left(\frac{1}{\lambda_0(T)} - \frac{1}{\lambda_X(T)}\right)\right) \\
&= \sup_{\omega\in B_1(\mathcal{H}^{(m)})} \int_0^\infty \omega'(x)\left(\frac{1}{\lambda_0(x)} - \frac{1}{\lambda_X(x)}\right) f_X(x)S_C(x)dx.
\end{aligned}$$

Denote $\alpha(x) = \left(\frac{1}{\lambda_0(x)} - \frac{1}{\lambda_X(x)}\right) f_X(x)S_C(x)$, and, as usual, $K^\star(x,y) = \frac{\partial^2}{\partial x\partial y}K(x,y)$. Then,

$$\text{m-KSD}(f_X\|f_0)^2 = \int_0^\infty \int_0^\infty \alpha(x)K^\star(x,y)\alpha(y)dxdy.$$

Since $K^\star$ is C_0-universal by Condition a), the previous term is equal to 0 if and only if $\alpha(x) = 0$ for all $x > 0$. Now, $\alpha(x) = 0$ if and only if $\frac{1}{\lambda_0(x)} - \frac{1}{\lambda_X(x)} = 0$, which holds if and only if $f_0(x) = f_X(x)$ for all $x > 0$.

Proof of Theorem 4.3

Deterministic kernels $(c \in \{s, m\})$:
For $c \in \{s, m\}$ which are associated to a deterministic Stein kernel function $h^{(c)}((T,\Delta),(T',\Delta'))$, the result follows from the classical theory of V-statistics since $h^{(c)}$ are degenerate kernels, and under the following moment conditions:

i) $\mathbb{E}_0\left[|h^{(c)}((T,\Delta),(T,\Delta))|\right] < \infty$,

ii) $\mathbb{E}_0\left[h^{(c)}((T,\Delta),(T',\Delta'))^2\right] < \infty$,

which are satisfied due to Lemma 4.2.

Random kernel $(c \in \{p\})$:
Observe that

$$\begin{aligned}
\sqrt{n}\cdot\widehat{\text{c-KSD}}(f_X\|f_0) &= \sup_{\omega\in B_1(\mathcal{H}^{(p)})} \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{(\omega(T_i)\lambda_0(T_i))'}{\lambda_0(T_i)}\frac{\Delta_i}{Y(T_i)/n} \\
&= \sup_{\omega\in B_1(\mathcal{H}^{(p)})} \frac{1}{\sqrt{n}}\int_0^{\tau_n} \frac{(\omega(x)\lambda_0(x))'}{\lambda_0(x)}\frac{1}{Y(x)/n}dN(x),
\end{aligned}$$

where $dN(x) = \sum_{i=1}^{n} \Delta_i \delta_{T_i}(x)$. By hypothesis, $\int_0^\infty (\omega(x)\lambda_0(x))' dx = 0$ for all $\omega \in \mathcal{H}^{(p)}$, then

$$\begin{aligned}
&\sqrt{n} \cdot \widehat{\text{c-KSD}}(f_X \| f_0) \\
&= \sup_{\omega \in B_1(\mathcal{H}^{(p)})} \frac{1}{\sqrt{n}} \int_0^{\tau_n} \frac{(\omega(x)\lambda_0(x))'}{\lambda_0(x)} \frac{1}{Y(x)/n} dN(x) - \sqrt{n} \int_0^\infty (\omega(x)\lambda_0(x))' dx \\
&= \sup_{\omega \in B_1(\mathcal{H}^{(p)})} \frac{1}{\sqrt{n}} \int_0^{\tau_n} \frac{(\omega(x)\lambda_0(x))'}{\lambda_0(x)} \frac{1}{Y(x)/n} dM(x) - \sqrt{n} \int_{\tau_n}^\infty (\omega(x)\lambda_0(x))' dx
\end{aligned}$$

where $dM(x) = dN(x) - Y(x)\lambda_0(x)dx$. Therefore we conclude that $\sqrt{n} \cdot \widehat{\text{c-KSD}}(f_X \| f_0) \in [a - b, a + b]$, where

$$\begin{aligned}
a &= \sup_{\omega \in B_1(\mathcal{H}^{(p)})} \frac{1}{\sqrt{n}} \int_0^{\tau_n} \frac{(\omega(x)\lambda_0(x))'}{\lambda_0(x)} \frac{1}{Y(x)/n} dM(x), \text{ and} \\
b &= \sup_{\omega \in B_1(\mathcal{H}^{(p)})} \sqrt{n} \int_{\tau_n}^\infty (\omega(x)\lambda_0(x))' dx
\end{aligned}$$

We will prove that $b = o_p(1)$. Let $K^\star(x, y) = \left(\frac{\partial^2}{\partial x \partial y} \lambda_0(x)\lambda_0(y)K(x, y)\right)$, then

$$\begin{aligned}
&\left(\sup_{\omega \in B_1(\mathcal{H}^{(p)})} \sqrt{n} \cdot \int_{\tau_n}^\infty (\omega(x)\lambda_0(x))' dx \right)^2 = n \int_{\tau_n}^\infty \int_{\tau_n}^\infty \frac{K^\star(x, y)}{f_T(x) f_T(y)} f_T(x) f_T(y) dx dy \\
&\leq n S_T(\tau_n)^{1/2} \left(\int_{\tau_n}^\infty \left(\int_{\tau_n}^\infty \frac{K^\star(x, y)}{f_T(x) f_T(y)} f_T(x) dx \right)^2 f_T(y) dy \right)^{1/2} \\
&\leq n S_T(\tau_n) \left(\int_{\tau_n}^\infty \int_{\tau_n}^\infty \frac{K^\star(x, y)^2}{f_T(x)^2 f_T(y)^2} f_T(x) f_T(y) dx dy \right)^{1/2},
\end{aligned}$$

where the two inequalities above follow from the Cauchy-Schwarz inequality, by the fact that $nS_T(\tau_n) = O_p(1)$ [Yang, 1994], and the previous double integral converges to 0 by Condition c.3), since $\tau_n = \max\{T_1, \ldots, T_n\} \to \infty$. From the previous result, we deduce

$$\sqrt{n} \cdot \widehat{\text{c-KSD}}(f_X \| f_0) = \sup_{\omega \in B_1(\mathcal{H}^{(p)})} \frac{1}{\sqrt{n}} \int_0^{\tau_n} \frac{(\omega(x)\lambda_0(x))'}{\lambda_0(x)} \frac{1}{Y(x)/n} dM(x) + o_p(1).$$

The previous step is important the analysis as it allows to write $\sqrt{n} \cdot \widehat{\text{c-KSD}}(f_X \| f_0)$ in terms of $M(x)$. Our next step is to prove that we can replace the term $Y(x)/n$,

in the previous equation, by $S_T(x)$. Observe

$$\begin{aligned}
&\sqrt{n}\cdot\widehat{\text{c-KSD}}(f_X\|f_0)\\
&=\sup_{\omega\in B_1(\mathcal{H}^{(p)})}\frac{1}{\sqrt{n}}\int_0^{\tau_n}\frac{(\omega(x)\lambda_0(x))'}{\lambda_0(x)}\left(\frac{1}{Y(x)/n}-\frac{1}{S_T(x)}+\frac{1}{S_T(x)}\right)dM(x)+o_p(1)\\
&=\sup_{\omega\in B_1(\mathcal{H}^{(p)})}\frac{1}{\sqrt{n}}\int_0^{\tau_n}\frac{(\omega(x)\lambda_0(x))'}{\lambda_0(x)}\frac{1}{S_T(x)}dM(x)\\
&\quad\pm\sup_{\omega\in B_1(\mathcal{H}^{(p)})}\frac{1}{\sqrt{n}}\int_0^{\tau_n}\frac{(\omega(x)\lambda_0(x))'}{\lambda_0(x)}\left(\frac{1}{Y(x)/n}-\frac{1}{S_T(x)}\right)dM(x)+o_p(1).
\end{aligned}$$

The $\pm$ notation above denotes lower, given by $-$, and upper, given by $+$, bounds for $\sqrt{n}\cdot\widehat{\text{c-KSD}}(f_X\|f_0)$. Finally, by taking square, the result is deduced by proving

$$\sup_{\omega\in B_1(\mathcal{H}^{(p)})}\frac{1}{\sqrt{n}}\int_0^{\tau_n}\frac{(\omega(x)\lambda_0(x))'}{\lambda_0(x)}\left(\frac{1}{Y(x)/n}-\frac{1}{S_T(x)}\right)dM(x)=o_p(1),$$

and

$$\sup_{\omega\in B_1(\mathcal{H}^{(p)})}\frac{1}{\sqrt{n}}\int_0^{\tau_n}\frac{(\omega(x)\lambda_0(x))'}{\lambda_0(x)}\frac{1}{S_T(x)}dM(x)=O_p(1).$$

The second equation won't be verified as, at the end of this proof, we will show that such a quantity converges in distribution to some random variable, thus it will be bounded in probability. For the first equation, notice that

$$\begin{aligned}
&\left(\sup_{\omega\in B_1(\mathcal{H}^{(p)})}\frac{1}{\sqrt{n}}\int_0^{\tau_n}\frac{(\omega(x)\lambda_0(x))'}{\lambda_0(x)}\left(\frac{1}{Y(x)/n}-\frac{1}{S_T(x)}\right)dM(x)\right)^2\\
&=\frac{1}{n}\int_0^{\tau_n}\int_0^{\tau_n}\frac{K^\star(x,y)}{\lambda_0(x)\lambda_0(y)}\left(\frac{1}{Y(x)/n}-\frac{1}{S_T(x)}\right)\left(\frac{1}{Y(y)/n}-\frac{1}{S_T(y)}\right)dM(x)dM(y),
\end{aligned}$$

is a double integral with respect to the $M(x)$. Then, by [Fernandez and Rivera, 2019, Theorem 17], it is enough to verify

$$\frac{1}{n}\int_0^{\tau_n}\frac{K^\star(x,x)}{\lambda_0(x)^2}\left(\frac{1}{Y(x)/n}-\frac{1}{S_T(x)}\right)^2Y(x)\lambda_0(x)dx=o_p(1).$$

Observe that

$$\begin{aligned}
&\frac{1}{n}\int_0^{\tau_n} \frac{K^\star(x,x)}{\lambda_0(x)^2}\left(\frac{1}{Y(x)/n} - \frac{1}{S_T(x)}\right)^2 Y(x)\lambda_0(x)dx \\
&= \int_0^{\tau_n} \frac{K^\star(x,x)}{\lambda_0(x)^2}\left(1 - \frac{Y(x)/n}{S_T(x)}\right)^2 \frac{1}{Y(x)/n}\lambda_0(x)dx \\
&= O_p(1)\int_0^{\tau} \frac{K^\star(x,x)}{\lambda_0(x)^2}\left(1 - \frac{Y(x)/n}{S_T(x)}\right)^2 \frac{1}{S_T(x)}\lambda_0(x)dx \\
&= o_p(1),
\end{aligned}$$

where the second equality follows from $n/Y(x) = O_p(1)1/S_T(x)$ uniformly for all $x \leq \tau_n$ [Gill, 1983], and the last equality is due to dominated convergence in sets of probability as high as desired, as $\left(1 - \frac{Y(x)/n}{S_T(x)}\right) \to 0$ for all $x < \infty$ from the Glivenko-Cantelli Theorem, and

$$\frac{K^\star(x,x)}{\lambda_0(x)^2}\left(1 - \frac{Y(x)/n}{S_T(x)}\right)^2 \frac{1}{S_T(x)}\lambda_0(x) = O_p(1)\frac{K^\star(x,x)}{f_0(x)^2 S_C(x)}f_0(x),$$

which is integrable by Condition c.3).

Putting everything together, we have shown that

$$\begin{aligned}
&\sqrt{n}\cdot \widehat{\text{c-KSD}}^2(f_X\|f_0) \\
&= \left(\sup_{\omega\in B_1(\mathcal{H}^{(p)})} \frac{1}{\sqrt{n}}\int_0^{\tau_n} \frac{(\omega(x)\lambda_0(x))'}{\lambda_0(x)}\frac{1}{S_T(x)}dM(x)\right)^2 + o_p(1) \\
&= \frac{1}{n}\int_0^{\tau_n}\int_0^{\tau_n} \frac{K^\star(x,y)}{f_0(x)f_0(y)S_C(x)S_C(y)}dM(x)dM(y) + o_p(1) \\
&= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n \int_0^{X_i}\int_0^{X_j} \frac{K^\star(x,y)}{f_0(x)f_0(y)S_C(x)S_C(y)}dM_j(x)dM_i(y) + o_p(1) \\
&= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n J((T_i,\Delta_i),(T_j,\Delta_j)) + o_p(1),
\end{aligned}$$

where $M_i(x) = N_i(x) - \int_0^x \mathbb{1}_{\{T_i\geq y\}}\lambda_0(y)dy = \Delta_i\mathbb{1}_{\{T_i\leq x\}} - \int_0^x \mathbb{1}_{\{T_i\geq y\}}\lambda_0(y)dy$. Notice that the process $M_i(x)$ only depends on the i-th observation (T_i,Δ_i). Notice that the previous expression is approximately a V-statistic with kernel given by $J((T_i,\Delta_i),(T_j,\Delta_j)) = \int_0^{T_i}\int_0^{T_j} \frac{K^\star(x,y)}{f_0(x)f_0(y)S_C(x)S_C(y)}dM_j(x)dM_i(y)$. By [Fernandez and Rivera, 2019, Proposition 23], we have that $\mathbb{E}[J((T_i,\Delta_i),(T_j,\Delta_j))|T_i,\Delta_i] = 0$,

thus J is a degenerate V-statistic kernel. By the classical theory of V-statistics,

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} J((T_i,\Delta_i),(T_j,\Delta_j)) \xrightarrow{\mathcal{D}} r_p + \mathcal{Y}_p,$$

where r_p is a constant and $\mathcal{Y}_p$ is a (potentially) infinite sum of independent χ^2 random variables, as long as the following moment conditions are satisfied:

$$\text{i) } \mathbb{E}_0(|J((T_1,\Delta_1),(T_1,\Delta_1))|) < \infty, \quad \text{and} \quad \text{ii) } \mathbb{E}_0(J((T_1,\Delta_1),(T_2,\Delta_2))^2) < \infty.$$

Again, by Proposition 23 of [Fernandez and Rivera, 2019], checking those moment conditions is equivalent to verify:

$$i)\mathbb{E}_0\left[\frac{K^\star(T,T)\Delta}{(f_0(T)S_C(T))^2}\right] < \infty \quad \text{and} \quad ii)\mathbb{E}_0\left[\frac{K^\star(T,T')^2\Delta\Delta'}{(f_0(T)f_0(T')S_C(T)S_C(T'))^2}\right] < \infty,$$

which are exactly the conditions assumed in Condition c.3).

4.B Known Identities

Martingales in Survival Analysis

In Section 4.3.2, we use the following identity to derive the martingale Stein operator

$$\mathbb{E}_0\left[\Delta\phi(T) - \int_0^T \phi(t)\lambda_0(t)dt\right] = 0,$$

which holds under the null hypothesis, where λ_0 is the hazard function under the null. Let $N_i(x)$ and $Y_i(x)$ be the individual counting and risk processes, defined by by $N_i(x) = \Delta_i \mathbb{1}_{\{T_i \le x\}}$ and $Y_i(x) = \mathbb{1}_{\{T_i \ge x\}}$, respectively. Then, the individual zero-mean martingale for the i-th individual corresponds to $M_i(x) = N_i(x) - \int_0^x Y_i(y)\lambda_0(y)dy$, where $\mathbb{E}_0(M_i(x)) = 0$ for all x.

Additionally, let $\phi : \mathbb{R}_+ \to \mathbb{R}$ such that $\mathbb{E}_0\left|\int_0^x \phi(y)dM_i(y)\right| < \infty$ for all x, then $\int_0^x \phi(y)dM_i(y)$ is a zero-mean $(\mathcal{F}_x)$-martingale (see Chapter 2 of [Aalen et al., 2008]). Then, taking expectation, we have

$$\begin{aligned}\mathbb{E}_0\left[\int_0^\infty \phi(x)dM_i(x)\right] &= \mathbb{E}_0\left[\int_0^\infty \phi(x)(dN_i(x) - Y_i(x)\lambda_0(x)dx)\right] \\ &= \mathbb{E}_0\left[\Delta\phi(T) - \int_0^T \phi(x)\lambda_0(x)dx\right] = 0,\end{aligned}$$

as stated above.

Chapter 5

A Kernel Test for Quasi-independence

Summary We consider settings in which the data of interest correspond to pairs of ordered times, e.g, the entry and survival times of patients in a clinical trial. In these settings, the two times are not independent (the second occurs after the first), yet it is still of interest to determine whether there exists significant dependence *beyond* their ordering in time. We refer to this notion as "quasi-(in)dependence". In this chapter, we propose a non-parametric statistical test of quasi-independence. The tests apply in the right-censored setting: an essential feature in clinical trials, where patients can withdraw from the study. We provide an asymptotic analysis of our test-statistic, and demonstrate in experiments that our test obtains better power than existing approaches, while being more computationally efficient.

5.1 Introduction

Many practical scientific problems require the study of events which occur consecutively in time. We focus here on the setting where event-times, X and Y, are only observed if they are in the ordered relationship $X \leq Y$. This type of data is commonly known as *truncated data*, and, in particular, we say that X is right-truncated by Y, or Y is left-truncated by X. In clinical trails, for example, only patients still alive at the beginning of the study can be recruited, hence the recruitment times X and the survival times Y are ordered. In the field of insurance, a liability claim may be placed at a time Y as a consequence of an incident at a time X. In e-commerce, the time Y of first purchase by a new user may only happen after the time X when the user registers with the website.

Our goal is to determine whether there exists an association between X and Y in the truncated data setting. Given that $X \leq Y$, the times X and Y will clearly not be independent (with the exception of trivial cases in which, for instance, X and Y have disjoint support). Thus, while it is not meaningful to test for statistical independence in the truncated setting, we can nevertheless still test for whether X

and Y are uncoupled apart from the fact that $X \leq Y$, using the notion of *quasi-independence*. We will make this notion formal in Section 5.2.

Testing for an association between ordered X and Y may be important in making business/medical decisions. In the setting of clinical trials, it is important to ensure that survival times are as "independent" as possible from recruitment times, in order to avoid bias in the recruitment process. In e-commerce, it may be of interest to test whether the purchase time for an item, such as a swimsuit, depends on the registration time, to determine seasonal effects on consumer behaviour and refine advertising strategies. In statistical modelling, a common working assumption is that X and Y are independent, but can only be observed when $X \leq Y$ holds [Hyde, 1977; Tsai, 1988; Woodroofe, 1985]. The independence assumption can be weakened to quasi-independence, which is testable, and under which typical methods are still valid [Klein and Moeschberger, 2006; Lagakos et al., 1988; Tsai et al., 1987; Turnbull, 1976; Wang, 1991; Woodroofe, 1985].

Our tests apply in the setting where Y is right-censored. This is a very common scenario in real-world applications, particularly in clinical trials, where patients may withdraw from the study before their event of interest is observed. In the e-commerce example, there may be registered users that have not yet made a purchase when the study ends. Formally, the data corresponds to the triple (X, T, Δ), where $T = \min\{C, Y\}$ is the minimum between the survival time Y of a given patient, and the time C at which said patient leaves the study (or the study ends), and $\Delta = \mathbb{1}_{\{T=Y\}}$, similarly defined in Chapter 4. Given the truncated data setting, we have further that $X \leq \min\{Y, C\}$. We emphasise that quasi-independence and right-censoring are very different data properties. Quasi-independence is a deterministic hard constraint $(X \leq Y)$, while right-censoring is a stochastic property of the data (incomplete observations). Quasi-independence has been widely studied in the statistics community, including for right-censored data [Chiou et al., 2018; Emura and Wang, 2010; Tsai, 1990]. We provide a brief review below and more detailed descriptions of relevant concepts and methods in subsequent Section 5.2.

In this chapter, we propose a non-parametric statistical test for quasi-independence, which can be applicable under right censoring. Our test statistic is a non-parametric generalisation of the log-rank test [Emura and Wang, 2010], where the departure from the null is characterised by functions in a RKHS. Consequently, we are able to straightforwardly detect a very rich family of alternatives, including non-monotone alternatives [Chiou et al., 2018]. Our test generalises the non-parametric statistical tests of independence based on the Hilbert-Schmidt Independence Criterion [Gretton et al., 2008]; which were adapted to the right-censoring

setting [Fernandez et al., 2019; Rindt et al., 2019]. Due to the additional correlations present in the test statistic under quasi-independence, however, we require the new approaches in our analysis of the consistency and asymptotic behaviour of our test statistic.

The rest of the chapter is organised as follows. In Section 5.2, we introduce the notion of quasi-independence. We next propose an RKHS statistic to detect this quasi-independence, and its finite sample estimate from data. We contrast the statistic for quasi-independence with the analogous RKHS statistic for independence, noting the additional sample dependencies on account of the left-truncation. In Section 5.2.2, we generalise the quasi-independence statistics to account for the presence of right-censored observations. In Section 5.3, we provide our main theoretical results: an asymptotic analysis for our test statistic, and a guarantee of consistency under the alternative. In order to determine the test threshold in practice, we introduce a wild bootstrap procedure to approximate the test threshold. In Section 5.4 we give a detailed empirical evaluation of our method. We begin with challenging synthetic datasets exhibiting periodic quasi-dependence, as would be expected for example from seasonal or daily variations, where our approach strongly outperforms the alternatives. Additionally, we show our test is consistently the best test in data-scenarios in which the censoring percentage is relatively high, see Figure 5.6. Next, we apply our test statistic to three real-data scenarios, shown in Figure 5.1: a survival analysis study for residents in the Channing House retirement community in Palo Alto, California [Hyde, 1977]; a study of transfusion-related AIDS [Lagakos et al., 1988]; and a study on spontaneous abortion [Meister and Schaefer, 2008]. For this last dataset, our general-purpose test is able to detect a mode of quasi-dependence discovered by a model that exploits domain-specific knowledge, but not found by alternative general-purpose testing approaches. This was a particular challenge due to the large percentage of censored observations in the abortion dataset; see censored marking in Figure 5.1. More details regarding censoring level are shown in Figure 5.6. Proofs of all results are given in the Appendices.

5.2 Quasi-independence

Our goal is to infer the null hypothesis of quasi-independence between X and Y. Formally, this null hypothesis is characterised as

$$H_0 : \pi(x, y) = \widetilde{F}_X(x)\widetilde{S}_Y(y), \quad \text{for all } x \leq y, \tag{5.1}$$

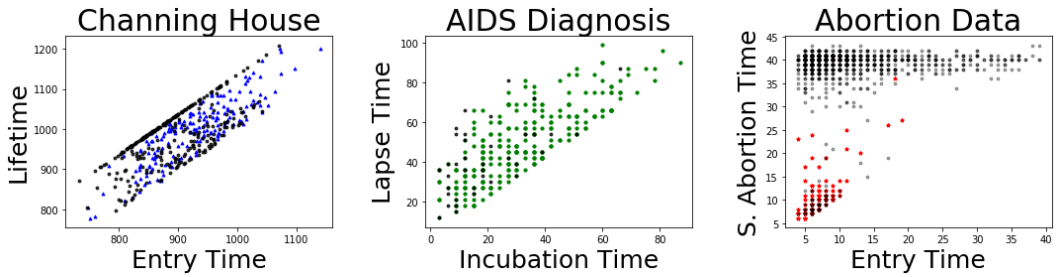


Figure 5.1: Channing House dataset: the x-axis shows the entry time to the retirement center; and the y-axis shows the right-censored lifetimes . Events are censored by withdrawal from the center or study finishes at July 1, 1975. AIDS dataset: the x-axis shows the incubation time X; and the y-axis shows the censored lapse time Y, measured from infection to recruitment time. Events are censored by death or left the study. Infected patients were recruited in the study only if they developed AIDS within the study period, therefore, in this dataset, the incubation time X does not exceed the lapse time Y. Abortion dataset: the x-axis shows the time to enter the study; and the y-axis shows the right-censored time for spontaneous abortion. Events are censored due to life birth and induced abortions. All censored times are marked in dark.

where $\pi(x, y) = \mathbb{P}(X \leq x, Y \geq y)$, and $\tilde{F}_X(x)$ and $\tilde{S}_Y(y)$ are functions that only depend on x and y, respectively. In case of independent X and Y, $\widetilde{F}_X(x)$ and $\widetilde{S}_Y(y)$ coincide with $F_X(x) = \mathbb{P}(X \leq x)$ and $S_Y(y) = \mathbb{P}(Y \geq y)$, but in general they may differ. For simplicity, X and Y are assumed continuously distributed on $\mathbb{R}_+$, and f_{XY}, f_X and f_Y denote the joint density and the corresponding marginals, and $f_{Y|X=x}$ denotes the conditional density of Y given $X = x$.

To simplify the notation, we suppose throughout that $X \leq Y$ always holds, and thus write $\pi(x, y) = \mathbb{P}(X \leq x, Y \geq y)$ instead of $\pi(x, y) = \mathbb{P}(X \leq x, Y \geq y|X \leq Y)$, as $\mathbb{P}(X \leq Y) = 1$. We remark, however, that the ordering $X \leq Y$ can be ensured by considering a conditional probability space given $X \leq Y$, and restricting calculations of probabilities, expectation etc. to this space [Chiou et al., 2018; Emura and Wang, 2010; Tsai, 1990].

The notion of quasi-independence must not be confused with the notion of independent increments, i.e., $X \perp (Y - X)$. For instance, generate X and Y such that $X \leq Y$ by sampling i.i.d. uniform random variables, say (U_1, U_2), in the interval $(0, 1)$, and make $X = U_1$ and $Y = U_2$ for the first pair (U_1, U_2) such that $U_1 \leq U_2$. It can be verified that this construction leads to quasi-independent random variables (X, Y), but X and $Y - X$ are not independent as the distribution of $Y - X$ is constrained by how large the original value of X was. The larger X is, the smaller is the value of $Y - X$.

In Emura and Wang [2010], the authors propose to measure quasi-

independence by using a log-rank-type test-statistic and pre-defined (fixed) weight function ω, which estimates

$$\int_{x \leq y} \omega(x,y)\rho(x,y)dxdy, \tag{5.2}$$

where

$$\rho(x,y) = -\pi(x,y)\frac{\partial^2 \pi(x,y)}{\partial x \partial y} + \frac{\partial \pi(x,y)}{\partial x}\frac{\partial \pi(x,y)}{\partial y},\ x \leq y. \tag{5.3}$$

The function ρ is originally inspired by the odds ratio [Chaieb et al., 2006], notwithstanding that ρ here is a difference measure, rather than a ratio. Under the assumption of quasi-independence, $\rho \equiv 0$, and thus $\int_{x \leq y} \omega(x,y)\rho(x,y)dxdy = 0$.

5.2.1 Kernel Quasi-independence Criterion (KQIC)

Nevertheless, it may be that $\int_{x \leq y} \omega(x,y)\rho(x,y)dxdy = 0$ even if the quasi-independence assumption is not satisfied, since the quantity depends on the function ω; for instance, it is trivially zero when $\omega = 0$. To avoid choosing a specific weight function ω, we optimise over a class of weight functions, taking an RKHS approach,

$$\Psi = \sup_{\omega \in B_1(\mathcal{H})} \int_{x \leq y} \omega(x,y)\rho(x,y)dxdy, \tag{5.4}$$

where $B_1(\mathcal{H})$ is the unit ball of a RKHS $\mathcal{H}$ with bounded measurable kernel given by $\mathfrak{K} : \mathbb{R}^2_+ \times \mathbb{R}^2_+ \to \mathbb{R}$. We refer to the measure Ψ^2 as *Kernel Quasi-Independent Criterion (KQIC)*. It can easily be verified that $\Psi \geq 0$; and, if X and Y are quasi-independent, then $\Psi = 0$. For c_0-universal kernels [Sriperumbudur et al., 2011], we have that $\Psi = 0$ if and only if X and Y are quasi-independent: see Theorem 5.2. Given the i.i.d. samples $((X_i, Y_i))_{i \in [n]}$, we can estimate Ψ via Ψ_n, defined as

$$\Psi_n = \sup_{\omega \in B_1(\mathcal{H})} \left(\frac{1}{n}\sum_{i=1}^n \omega(X_i,Y_i)\widehat{\pi}(X_i,Y_i) - \frac{1}{n^2}\sum_{i=1}^n \sum_{k=1}^n \omega(X_i,Y_k)\mathbb{1}_{\{X_k \leq X_i < Y_k \leq Y_i\}} \right) \tag{5.5}$$

where $\widehat{\pi}(x,y) = \frac{1}{n}\sum_{m=1}^n \mathbb{1}_{\{X_m \leq x, Y_m \geq y\}}$, and notice that $\widehat{\pi}(x,y)$ estimates $\pi(x,y)$. Using a reproducing kernel $\mathfrak{K}$ that factorises, we obtain a simple expression for Ψ_n^2:

Proposition 5.1. *Consider* $\mathfrak{K}((x,y),(x',y')) = K(x,x')L(y,y')$. *Then*

$$\Psi_n^2 = \frac{1}{n^2}\mathrm{tr}(\boldsymbol{K}\widehat{\boldsymbol{\pi}}\boldsymbol{L}\widehat{\boldsymbol{\pi}} - 2\boldsymbol{K}\widehat{\boldsymbol{\pi}}\boldsymbol{L}\boldsymbol{A}^\top + \boldsymbol{K}\boldsymbol{A}\boldsymbol{L}\boldsymbol{A}^\top)$$

where $\boldsymbol{K}$, $\boldsymbol{L}$, *and* $\boldsymbol{A}$ *are* $n \times n$*-matrices with entries given by* $\boldsymbol{K}_{ik} = K(X_i, X_k)$, $\boldsymbol{L}_{ik} = L(Y_i, Y_k)$ *and* $\boldsymbol{A}_{ik} = \mathbb{1}_{\{X_k \leq X_i < Y_k \leq Y_i\}}/n$, *and* $\widehat{\boldsymbol{\pi}}$ *is a diagonal matrix with entries* $\widehat{\boldsymbol{\pi}}_{ii} = \widehat{\pi}(X_i, Y_i)$.

We remark that the previous expression is similar in form to the Hilbert Schmidt Independence Criterion [Gretton et al., 2005b]. In particular, for empirical distributions, $\mathrm{HSIC}(\widehat{F}_{XY}, \widehat{F}_X\widehat{F}_Y) = \frac{1}{n^2}\mathrm{tr}(\boldsymbol{K}\boldsymbol{H}^\top\boldsymbol{L}\boldsymbol{H})$ with $\boldsymbol{H} = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$, whereas our test-statistic can be rewritten as $\Psi_n^2 = \frac{1}{n^2}\mathrm{tr}(\boldsymbol{K}\tilde{\boldsymbol{H}}^\top\boldsymbol{L}\tilde{\boldsymbol{H}})$ with $\tilde{\boldsymbol{H}} = (\widehat{\boldsymbol{\pi}} - \boldsymbol{A}^\top)$. Note that $\tilde{\boldsymbol{H}}$ is much more complex than $\boldsymbol{H}$, being a random matrix where each entry depends on all the data points. As we will see, this issue makes the asymptotic analysis in our case much more challenging; by contrast, the asymptotic distribution for HSIC can be readily obtained using standard results on U-statistics [Gretton et al., 2008; Chwialkowski and Gretton, 2014].

Our test can be understood as a generalisation of the log-rank test proposed by [Emura and Wang, 2010], where instead of considering a single log-rank test with a specific weight function, we consider the supremum over a collection of log-rank tests with weight functions in $B_1(\mathcal{H})$. By choosing a sufficiently rich RKHS, for example the RKHS induced by the exponentiated quadratic kernels, we are able to ensure test power against a broad family of alternatives. Conversely, simple kernels can recover classical parametric tests such as the aforementioned log-rank tests. As explained by Equation 7 in Emura and Wang [2010], the simplest possible (constant) function space recovers the well-known conditional Kendall's tau statistic.

Proposition 5.2 (Recovering conditional Kendall's tau)**.** *Consider* $\mathfrak{K} = 1$, *then* $\Psi_n^2 = K_a^2/n^2$, *where* $K_a = \sum_{i<k} \mathbb{1}_{\{X_i \vee X_k \leq Y_i \wedge Y_k\}} \mathrm{sign}\left((X_i - X_k)(Y_i - Y_k)\right)$ *is an empirical estimator of the conditional Kendall's tau.*

5.2.2 KQIC with Right-censoring

In clinical trails, for example, patients might withdraw from the study before observing the time Y of interest leading to so-called right-censored data. To model this kind of data, we introduce additionally the random censoring time C. The data correspond now to i.i.d. samples $((X_i, T_i, \Delta_i))_{i\in[n]}$, where $T_i = \min\{Y_i, C_i\}$ is the observation time, and $\Delta_i = \mathbb{1}_{\{T_i = Y_i\}}$ is the corresponding censoring status. In particular, if $\Delta_i = 0$, we only observe the censoring time $T_i = C_i$, and not the time of interest Y_i. Throughout, we assume that $X_i \leq T_i$ always holds, to reflect the

natural ordering of the times, i.e. first recruitment and second the event of interest or the withdrawal from the study. As for the censored setting, X, Y and C are supposed to be continuously distributed on $\mathbb{R}_+$. Our results are valid under the standard non-informative censoring assumption:

Assumption 5.1. *The censoring times are independent of the survival times given the entry times, i.e., $C_i \perp Y_i | X_i$.*

Standard notation for marginal, joint and conditional densities will be used: for instance, f_C, f_{XT} and $f_{Y|X=x}$, are the marginal density of C, the joint density of X and T, and the conditional density of Y given $X = x$, respectively. Moreover, S_Y denotes the survival function of Y, defined as $S_Y(y) = \mathbb{P}(Y \geq y)$ and $S_{C|X=x}(y) = \mathbb{P}(Y \geq y | X = x)$ is the conditional survival function of Y given $X = x$. Under Assumption 5.1 we have $S_{T|X=x}(y) = S_{Y|X=x}(y) S_{C|X=x}(y)$.

The null hypothesis of *quasi-independence* is formulated, for the right-censored setting, as

$$H_0 : f_{XY}(x, y) = \tilde{f}_X(x)\tilde{f}_Y(y), \quad \text{for all } x \leq y, \text{ s.t. } S_{T|X=x}(y) > 0. \tag{5.6}$$

As with the uncensored case, $\tilde{f}_X$ and $\tilde{f}_Y$ are not necessarily equal to the marginal densities f_X and f_Y. The additional condition $S_{T|X=x}(y) > 0$ ensures that the pair (x, y) is actually observable despite the censoring. The statistic Ψ from Eq. (5.4) is then extended to the censored setting,

$$\Psi_c = \sup_{\omega \in B_1(\mathcal{H})} \int_{x \leq y} \omega(x, y) \rho^c(x, y) dx dy \geq 0,$$

where

$$\rho^c(x, y) = -\pi^c(x, y) \frac{\partial^2}{\partial x \partial y} \pi_1^c(x, y) + \frac{\partial \pi^c(x, y)}{\partial x} \frac{\partial \pi_1^c(x, y)}{\partial y},$$

$\pi_1^c(x, y) = \mathbb{P}(X \leq x, T \geq y, \Delta = 1)$ and $\pi^c(x, y) = \mathbb{P}(X \leq x, T \geq y)$ for $x \leq y$.

Proposition 5.3. *We have $\Psi_c = 0$ if the null hypothesis H_0 of quasi-independence is fulfilled.*

The updated estimator for KQIC that incorporates censoring, Ψ_c, is defined by

$$\Psi_{c,n} = \sup_{\omega \in B_1(\mathcal{H})} \Big(\frac{1}{n} \sum_{i=1}^{n} \Delta_i \omega(X_i, T_i) \widehat{\pi}^c(X_i, T_i) - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{k=1}^{n} \Delta_k \omega(X_i, T_k) \mathbb{1}_{\{X_k \leq X_i < T_k \leq T_i\}} \Big) \tag{5.7}$$

where $\widehat{\pi}^c(x,y) = \frac{1}{n}\sum_{m=1}^{n} \mathbb{1}_{\{X_m \leq x, T_m \geq y\}}$ is the natural estimator for $\pi^c(x,y)$. In the uncensored case, i.e. $\Delta = 1$ with probability 1, the updated KQIC with censoring Ψ_c and its estimator $\Psi_{c,n}$ collapse to the respective quantities Ψ and Ψ_n from Section 5.2. Moreover, the estimator $\Psi_{c,n}$ can be simplified for factorising kernels.

Proposition 5.4. *Consider* $\mathfrak{K}((x,y),(x',y')) = K(x,x')L(y,y')$, *then*

$$\Psi^2_{c,n} = \frac{1}{n^2}\mathrm{tr}(\boldsymbol{K}\widehat{\boldsymbol{\pi}}^c\tilde{\boldsymbol{L}}\widehat{\boldsymbol{\pi}}^c - 2\boldsymbol{K}\widehat{\boldsymbol{\pi}}^c\tilde{\boldsymbol{L}}\boldsymbol{B}^\top + \boldsymbol{K}\boldsymbol{B}\tilde{\boldsymbol{L}}\boldsymbol{B}^\top) \tag{5.8}$$

where $\boldsymbol{K}_{ik} = K(X_i, X_k)$, $\tilde{\boldsymbol{L}}_{ik} = \Delta_i\Delta_k L(T_i, T_k)$, $\boldsymbol{B}_{ik} = \mathbb{1}_{\{X_k \leq X_i < T_k \leq T_i\}}/n$, *and* $\boldsymbol{\pi}^c$ *is a diagonal matrix where* $\widehat{\boldsymbol{\pi}}^c_{ii} = \widehat{\pi}(X_i, T_i)$.

5.3 Asymptotic Analysis and Wild Bootstrap Test

We now present our main two theoretical results. First, we establish the asymptotic null distribution of our statistic $n\Psi^2_{c,n}$.

Theorem 5.1. *Assume* $\mathfrak{K}$ *is bounded. Then, under the null hypothesis,* $n\Psi^2_{c,n} \xrightarrow{\mathrm{d}} \mu + \mathcal{Y}$, *where* μ *is a positive constant,* $\mathcal{Y} = \sum_{i=1}^{\infty}\lambda_i(\xi_i^2 - 1)$, $\xi_1, \xi_2, \ldots$ *are independent standard normal random variables, and* $\lambda_1, \lambda_2, \ldots$ *are non-negative constants depending on the distribution of the random variables* (X, Y, C) *and the kernel* $\mathfrak{K}$.

To verify Theorem 5.1, we show that the scaled version of our statistic, $n\Psi^2_{c,n}$, can be expressed under the null hypothesis as the sum of a certain V-statistic and an asymptotically vanishing term. To find this representation, we write our test-statistic as a double integral with respect to a martingale, and use martingale techniques, and the results introduced in [Fernández and Rivera, 2020], to show that the error incurred by replacing certain quantities by their population versions vanishes as the number of data points grows to infinity. The full proof is provided in Appendix 5.A. We next establish conditions for consistency of the test under the alternative.

Theorem 5.2. *Let* $\mathfrak{K}$ *be a bounded,* c_0*-universal kernel [Sriperumbudur et al., 2011]. Then* $\Psi^2_{c,n} \to \Psi^2_c$ *in probability. Moreover, whenever the null hypothesis is violated,* Ψ^2_c *is positive, implying that* $n\Psi^2_{c,n} \to \infty$ *in probability,.*

We remark that the factorised kernel $\mathfrak{K}((x,y),(x',y')) = K(x,x')L(y,y')$ is required to be c_0-universal in the product space, which is true for instance when K and L are exponentiated quadratic kernels [Fukumizu et al., 2007b]. In the case of independence testing, a simpler condition on the kernel can be used, where kernels are required to be individually characteristic to their respective domains [Gretton,

2015]. Whether this simple condition can be generalised to the quasi-independence setting remains a topic for future work.

The consistency result in Theorem 5.2 relies on the interpretation of the test statistic $\Psi_{c,n}$ and the KQIC Ψ_c, as the Hilbert space distances of the embeddings of certain positive measures. These distances measure the degree of (quasi)-dependence. Under the c_0-universality assumption, the embedding of finite signed measures are injective [Sriperumbudur et al., 2011], which, in our case, implies $\rho^c(x, y) = 0$ for almost all $x \leq y$. It remains to prove that quasi-independence holds. To show this, we first note that $\rho^c(x, y) = 0$ implies

$$\frac{\partial^2 \pi_1^c(x, y)}{\partial x \partial y} = \frac{1}{\pi^c(x, y)} \frac{\partial \pi^c(x, y)}{\partial x} \frac{\partial \pi_1^c(x, y)}{\partial y}, \tag{5.9}$$

and that $\frac{\partial^2 \pi_1^c(x,y)}{\partial x \partial y} = S_{C|X=x}(y) f_{XY}(x, y)$. By carefully analysing Eq. (5.9) we find an explicit decomposition of $f_{XY}(x, y)$ into the product of two functions only depending on x and y, respectively, from which quasi-independence follows. A detailed proof is provided in Appendix 5.A.

As noted above, the eigenvalues λ_i in Theorem 5.1 — and thus, the limit distribution of our test statistic under the null hypothesis — depend on the unknown distribution of (X, Y, C). For this reason, we propose to approximate the limit null distribution and its $(1 - \alpha)$-quantile q_α of $\mu + \mathcal{Y}$ using a wild bootstrap approach. This strategy is well-established for V- and U-statistics [Dehling and Mikosch, 1994], and has successfully been applied in scenarios, similar to the present one, where the test statistic behaves asymptotically as a V-statistic [Fernandez et al., 2019; Fernandez and Rivera, 2019].

Wild Bootstrap Testing Procedure We introduce the wild bootstrap counterpart $\Psi_{c,n}^{\mathrm{WB}}$ of our statistic $\Psi_{c,n}$. Let $W_1, \ldots, W_n$ be independent and identically distributed Rademacher random variables, and define the $n \times n$ matrix $\boldsymbol{K}^W$ with entries $\boldsymbol{K}_{ik}^W = W_i W_k K(X_i, X_k)$. Then,

$$(\Psi_{c,n}^{\mathrm{WB}})^2 = \frac{1}{n^2} \mathrm{tr}(\boldsymbol{K}^W \widehat{\boldsymbol{\pi}}^c \tilde{\boldsymbol{L}} \widehat{\boldsymbol{\pi}}^c - 2\boldsymbol{K}^W \widehat{\boldsymbol{\pi}}^c \tilde{\boldsymbol{L}} \boldsymbol{B}^\top + \boldsymbol{K}^W \boldsymbol{B} \tilde{\boldsymbol{L}} \boldsymbol{B}^\top).$$

We propose the test $\varphi_n^{\mathrm{WB}} = \mathbb{1}\{\Psi_{c,n}^2 > q_\alpha^{\mathrm{WB}}\}$ to infer H_0, where q_α^{WB} denotes the $(1 - \alpha)$-quantile of the simulated null from wild bootstrap $(\Psi_{c,n}^{\mathrm{WB}})^2$ given the observations $((X_i, \Delta_i, T_i))_{i \in [n]}$.

5.4 Experiments

We perform synthetic experiments followed by real data applications. In the first set of synthetic examples, we replicate the settings studied in [Chiou et al., 2018], where Gaussian copula models were used to create dependencies between X and Y. In the second synthetic experiment, we investigate distribution functions $f_{Y|X=x}$ that have a periodic dependence on x. We then apply our tests to real-data scenarios such as those studied in [Emura and Wang, 2010] and [Meister and Schaefer, 2008].

Quasi-independence Methods

We implement the proposed quasi-independence test based on the test-statistic **KQIC** given in Eq. (5.8). The kernels are chosen to be Gaussian with bandwidth optimised by using approximate test power [Gretton et al., 2009a; Jitkrittum et al., 2017]. See Appendix 5.D for details. Competing approaches include: **WLR**, the weighted log-rank test proposed in [Emura and Wang, 2010], with weight function chosen equal to $n\widehat{\pi}^c(x,y)$;[1] **WLR_SC**, the weighted log-rank test proposed in [Emura and Wang, 2010], with weight function chosen as suggested by the authors, i.e, $W(x,y) = \int_0^x \widehat{S}_{C_R}((y-u)-)^{-1}\widehat{\pi}^c(du,y)$, where $\widehat{S}_{C_R}$ is the Kaplan-Meier estimator associated to the data $((C_i - X_i, 1-\Delta_i))_{i=1}^n$; **M&B**, the conditional Kendall's tau statistic modified to incorporate censoring as proposed in [Martin and Betensky, 2005]; and **MinP1** and **MinP2**, the "minimal p-value selection" tests proposed in [Chiou et al., 2018], which rely on permutations of the observed pairs. A review of these approaches can be found in Appendix 5.B. For the synthetic experiments, we recorded the rejection rate over 200 trials. The wild bootstrap size for **KQIC** and the permutation size for **MinP1**, **MinP2** are set to be 500.

5.4.1 Simulation Results

Monotonic Dependency The first synthetic example from [Chiou et al., 2018] is generated as follows: $X \sim \text{Exp}(5)$ and $Y \sim \text{Weibull}(3, 8.5)$; (X,Y) are then coupled via a 2-dimensional Gaussian copula model with correlation parameter ρ. The censoring variable is set to be exponentially distributed and truncation applies. With the copula construction, the magnitude of the correlation parameter ρ is a fair indicator of the degree of dependence, with $\rho = 0$ denoting independence. Rejection rates are reported in Table 5.1. At $\rho = 0$, the null hypothesis holds, and the rejection rates refer to the Type-I error. All the tests achieve a correct Type-I error around a test level $\alpha = 0.05$. For $\rho \neq 0$, the alternative holds, and the

[1] Our test-statistic recovers, as a particular case, the squared of this log-rank test by choosing $\mathfrak{K} = 1$

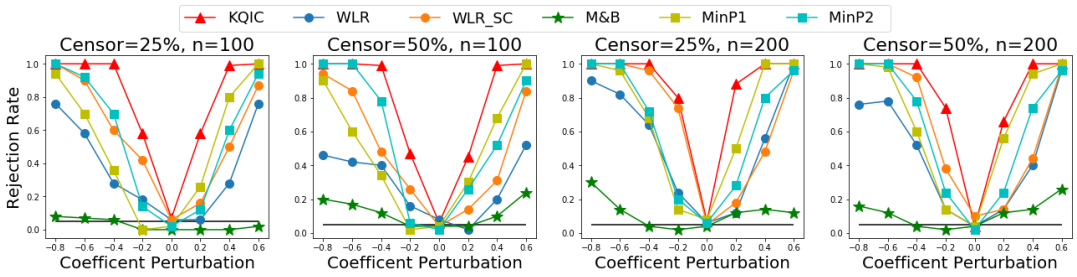


Figure 5.2: Rejection rate for V-shape Gaussian copula model

rejection rates correspond to test power (the higher the better). The highest value is in bold. Test results w.r.t. different censoring rates can be found in the Appendix. Overall, our method outperforms all competing approaches.

ρ	-0.4	-0.2	0.0	0.2	0.4	-0.4	-0.2	0.0	0.2	0.4
KQIC	**0.93**	**0.46**	0.06	**0.42**	**0.86**	**0.99**	**0.67**	0.05	**0.63**	**1.00**
WLR	0.80	0.33	0.10	0.18	0.66	0.94	0.52	0.06	0.32	0.94
WLR_SC	0.85	0.42	0.03	0.24	0.74	0.93	0.53	0.06	0.43	0.99
M&B	0.64	0.22	0.02	0.16	0.74	0.94	0.28	0.03	0.42	0.92
MinP1	0.58	0.12	0.03	0.17	0.62	0.84	0.12	0.10	0.34	0.84
MinP2	0.33	0.04	0.06	0.10	0.28	0.56	0.08	0.08	0.28	0.52

Table 5.1: Rejection rates for monotonic dependency models based on Gaussian copula, with $n = 100$ on the left; $n = 200$ on the right; $\alpha = 0.05$; censoring rate: 50%.

V-shaped Dependency Another synthetic example [Chiou et al., 2018], in which the authors compare the behaviour of their tests against the conditional Kendall's tau test of [Martin and Betensky, 2005] is detecting non-monotonic dependencies. The following V-shaped dependency structure applies: $X \sim \text{Weibull}(0.5, 4)$; $Y \sim \text{Uniform}[0, 1]$; $(X, |Y - 0.5|)$ is coupled via the 2-dimensional Gaussian copula with correlation coefficient ρ as above. Exponential censoring and truncation apply. Rejection rates are plotted against the perturbation of correlation coefficient ρ in Figure 5.2, where KQIC outperforms competing methods.

Periodic Dependency Apart from the V-shaped dependencies studied in Chiou et al. [2018], we investigate more complicated non-monotonic dependencies structures. The data are generated with a periodic dependency structure, $X \sim \text{Exp}(1)$; $Y|X \sim \text{Exp}(e^{\cos(2\pi\beta X)})$. The coefficient β controls the frequency of the dependence. A set of examples with different parameters β is shown in Figure 5.3, with $\beta = 0$ implying independence. Further details are discussed in Appendix 5.D.1.

Examining the results in Figure 5.4, we see that our method outperforms competing approaches. Unlike the correlation coefficient ρ in Gaussian copula models,

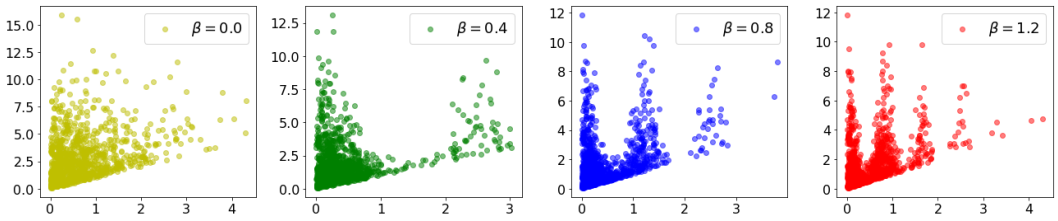


Figure 5.3: Samples from Periodic Dependency Model w.r.t. Frequency Coefficient β

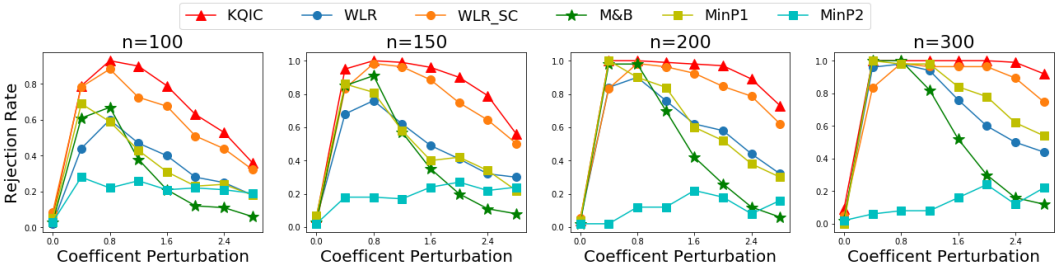


Figure 5.4: Rejection Rate for Periodic Dependency Model with 25% data censored.

the coefficient β does not directly imply the "amount" of dependence; rather, a higher β indicates a more "difficult" problem. Thus, as anticipated, power drops for large values of β, and the effect is more apparent at low sample sizes. Note in particular that the permutation based tests [Chiou et al., 2018] are more affected by an increase in frequency at which dependence occurs, while our test shows a more robust behaviour.

High Frequency Dependency In the period dependency problem above , the parameter β controls the frequency of sinusoidal dependence. At a given sample size, the dependence becomes harder to detect as the frequency β increases. We visually show this in Appendix 5.D.1. For problems with high frequency dependence, a larger sample size is required.

When the sample size increases, KQIC is able to successfully reject the null at relatively high frequencies (large β), as shown in Figure 5.5. At lower frequencies $\beta = 3.0$, WLR_SC has similar test power as KQIC. As the problem gets harder with larger β, KQIC outperforms WLR_SC. The IMQ kernel has similar test power as the Gaussian kernel on this example. We report the Type-I error that is well controlled in Appendix 5.D.1 Table 5.5.

Increasing Censoring Level We investigate how our test is affected by the censoring level, in particular when the censoring percentage increases. We analyse performance under both the null and alternative hypotheses. The Type-I error is well controlled for KQIC and details are reported in Appendix 5.D.3.

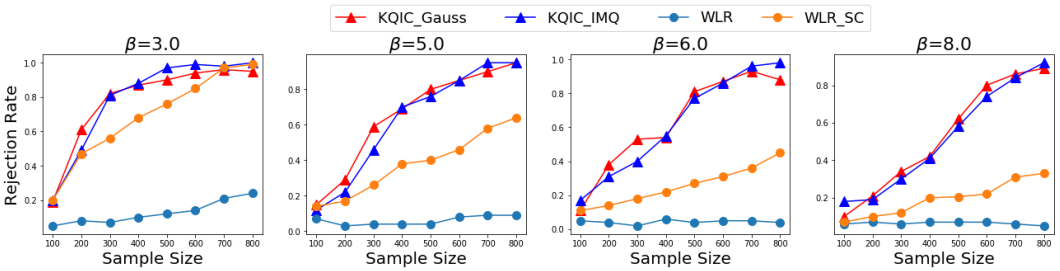


Figure 5.5: Rejection rate for high frequency dependency, with $\alpha = 0.05$, 40% data censored

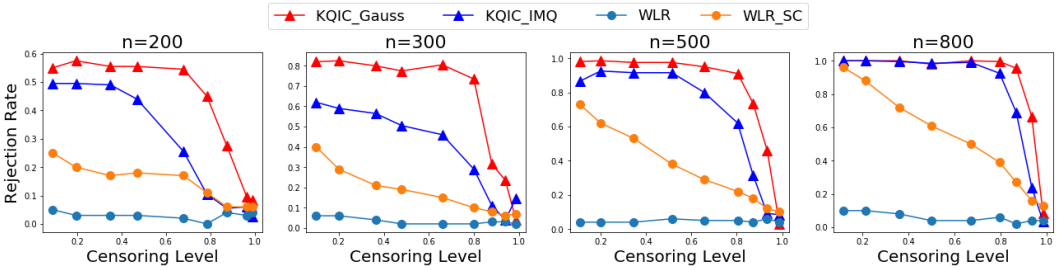


Figure 5.6: Rejection rate for periodic dependencies ($\beta = 5.0$), with $\alpha = 0.05$ and 200 trials.

Under the alternative hypothesis, in Figure 5.6, we show the rejection rate w.r.t. different censoring percentages and fixed sample size. This is done in our periodic dependency setting. From the plot, we see that KQIC with Gaussian and IMQ kernels is more robust to censoring, with test power starting to drop at 85% of censoring for sample size = 800. WLR_SC is strongly affected by censoring. WLR is not capable of detecting H_1 in this hard problem with high frequency.

In addition, we study the test behaviour with dependent censoring, since in Assumption 5.1, only conditional independence $Y \perp C|X$ is required [Emura and Wang, 2010]. Detailed results are reported in Appendix 5.D.2.

Computational runtime

As shown in Table 5.2, our proposed test, implemented as described in Appendix 5.C, has a significantly lower runtime when compared with the permutation approaches which require much longer run-time. M&B implements the conditional Kendall's tau statistic which has a closed-form expression for the null distribution, therefore the runtime is much lower again.

Our proposed test, implemented as described in Appendix 5.C, has a significantly lower runtime when compared with the competing permutation approaches. M&B implements the conditional Kendall's tau statistic, which has a closed-form expression for the null distribution, therefore its runtime is lowest of all.

n	100	200	300	400	500	600	700	800	900
KQIC	0.012	0.019	0.031	0.041	0.063	0.085	0.130	0.152	0.200
MinP1	15.77	41.62	56.61	90.52	113.7	154.4	254.4	299.2	389.1
MinP2	20.33	35.08	59.09	101.4	123.7	174.3	242.4	300.9	354.2
M&B	0.002	0.002	0.002	0.003	0.004	0.006	0.006	0.009	0.021

Table 5.2: The runtime, in seconds, for a single trial using 500 wild bootstrap samples for KQIC and 500 permutations for MinP1 and MinP2. M&B does not require to approximate the null distribution.

5.4.2 Real Data Applications

We consider three real data scenarios: **Channing House** [Hyde, 1977]: contains the recorded entry times and lifetimes of 461 patients (97 men and 364 women). Among them, 268 subjects withdrew from the retirement center, yielding to a censoring proportion of 0.62. The data are naturally left truncated, as only patients who entered the center are observed; **AIDS** [Lagakos et al., 1988]: the data contain the incubation time and lapse time, measured from infection to recruitment time, for 295 subjects. A censoring of proportion of 0.125 occurs due to death or withdrawal from the study. Left truncation applies since only patients that developed AIDS within the study period were recruited, thus only patients with incubation time not exceeding the lapse time were observed; and **Abortion** [Meister and Schaefer, 2008]: contains the entry time and the spontaneous abortion time for 1186 women (197 control group and 989 treatment group exposed to Coumarin derivatives). A censoring proportion of 0.906 occurs due to live birth or induced abortions. Delayed entry to the study is substantial in this dataset: 50% of the control cohort entered the study in week 9 or later, while in the treatment group this occurs for 25% of the cohort.

Implementations For our test we used both Gaussian kernels KQIC_Gauss and IMQ kernels KQIC_IMQ. For competing approaches, the implementation is as discussed at the beginning of this section.

Results For the Channing house dataset, in Table 5.3, we observe that all tests agree in not rejecting the null hypothesis for the combined and female groups at a level $\alpha = 0.05$. For the male group, all tests but MinP2 and M&B reject the null hypothesis at $\alpha = 0,05$. Our results agree with [Emura and Wang, 2010]. For the AIDS dataset, all tests reach a consensus of rejecting the null, which is consistent with [Emura and Wang, 2010], except for MinP2 marked in blue. For the abortion dataset, our test rejects the null hypothesis, suggesting dependency between the en-

(p-value)	Channing House			AIDS	Abortion Times		
	Combined	Male	Female		Combined	Control	Treatment
KQIC_Gauss	0.072	0.012	0.566	0.030	0.014	0.440	0.028
KQIC_IMQ	0.078	0.022	0.414	0.010	0.032	0.158	0.048
WLR	0.058	0.016	0.444	0.035	0.408	0.868	0.748
WLR_SC	0.086	0.020	0.422	0.030	0.511	0.674	0.450
MinP1	0.084	0.036	0.396	0.012	0.584	0.584	0.452
MinP2	0.198	0.426	0.118	0.406	0.694	0.572	0.346
M&B	0.178	0.199	0.495	0.010	0.712	0.693	0.752
% Events	0.379	0.474	0.354	0.875	0.094	0.069	0.098

Table 5.3: Real data test p-value, with marked results contradicting and supporting the scientific literature.

try time X and the spontaneous abortion time Y in both the treatment group and the combined case (in red). This finding is in accordance with domain knowledge [Meister and Schaefer, 2008], where the presence of this dependence was indicated to be due to the study design. The competing tests were unable to detect the dependence; however, did not reject the null hypothesis.

Appendices

5.A Proofs

The following Proposition is an intermediate result, which is needed to prove Lemmas 5.2 and 5.4.

Proposition 5.5. *Define*

$$A_n = \frac{1}{n}\sum_{i=1}^{n} \Delta_i \mathbb{1}_{\{X_i<T_i\}}(\widehat{\pi}^c(X_i,T_i) - \pi^c(X_i,T_i))^2. \tag{5.10}$$

Then, the following results hold: i) $A_n \to 0$ *almost surely as* n *grows to infinity and ii)* $|A_n| \leq c$*, for some constant* c*, for all large* n*.*

Proof. Since $\widehat{\pi}^c$ and π^c are both bounded by 1, we have $|A_n| = A_n \leq 4$ for all n and, thus, ii) is proven.

Let us consider the statement i). It is easy to see that $\mathbb{E}(\mathbb{1}_{\{X_m\leq x, T_m\geq t\}}) = \pi^c(x,t)$. In particular, we have $\mathbb{E}(g(m,i)|X_i,T_i,\Delta_i) = 0$ for $i \neq m$, where $g(m,i) = \mathbb{1}_{\{X_m\leq X_i, T_m\geq T_i\}} - \pi^c(X_i,T_i)$. Now, notice that we can rewrite A_n as V-statistic of order 3:

$$\begin{aligned} A_n &= \frac{1}{n}\sum_{i=1}^{n} \Delta_i \mathbb{1}_{\{X_i\leq T_i\}} \left(\frac{1}{n}\sum_{m=1}^{n}(\mathbb{1}_{\{X_m\leq X_i, T_m\geq T_i\}} - \pi^c(X_i,T_i))\right)^2 \\ &= \frac{1}{n^3}\sum_{i=1}^{n}\sum_{m=1}^{n}\sum_{k=1}^{n} \Delta_i \mathbb{1}_{\{X_i\leq T_i\}} g(m,i)g(k,i). \end{aligned}$$

Combining this and the law of large numbers for V-statistics yields

$$\begin{aligned} A_n \xrightarrow{a.s.} \mathbb{E}(\Delta_1 g(1,2)g(1,3)) &= \mathbb{E}(\Delta_1 \mathbb{E}(g(2,1)g(3,1)|X_1,T_1,\Delta_1)) \\ \text{(independence)} &= \mathbb{E}\left(\Delta_1 \mathbb{E}(g(2,1)|X_1,T_1,\Delta_1)\mathbb{E}(g(3,1)|X_1,T_1,\Delta_1)\right) \\ &= 0. \end{aligned}$$

□

Proofs of Sections 5.2 and 5.2.2

Proof of Proposition 5.1

Proof. From Eq. (5.5), we have

$$\Psi_n = \sup_{\omega \in B(\mathcal{H})} \frac{1}{n} \sum_{i=1}^{n} \left(\omega(X_i, Y_i)\widehat{\pi}(X_i, Y_i) - \sum_{k=1}^{n} \omega(X_i, Y_k) \boldsymbol{A}_{ik} \right),$$

where $\boldsymbol{A}_{ik} = \mathbb{1}_{\{X_k \leq X_i < Y_k \leq Y_i\}}/n$.

The previous result and the reproducing kernel property yield

$$\begin{aligned}
\Psi_n^2 &= \sup_{\omega \in B_1(\mathcal{H})} \left(\frac{1}{n} \sum_{i=1}^{n} \left(\omega(X_i, Y_i)\widehat{\pi}(X_i, Y_i) - \sum_{k=1}^{n} \omega(X_i, Y_k) \boldsymbol{A}_{ik} \right) \right)^2 \\
&= \sup_{\omega \in B_1(\mathcal{H})} \left\langle \omega(\cdot), \frac{1}{n} \sum_{i=1}^{n} K(X_i, \cdot) \left(L(Y_i, \cdot)\widehat{\boldsymbol{\pi}}_{ii} - \sum_{k=1}^{n} L(Y_k, \cdot) \boldsymbol{A}_{ik} \right) \right\rangle^2 \\
&= \left\| \frac{1}{n} \sum_{i=1}^{n} K(X_i, \cdot) \left(L(Y_i, \cdot)\widehat{\boldsymbol{\pi}}_{ii} - \sum_{k=1}^{n} L(Y_k, \cdot) \boldsymbol{A}_{ik} \right) \right\|_{\mathcal{H}}^2 \\
&= \frac{1}{n^2} \sum_{i,j=1}^{n} \boldsymbol{K}_{ij} \boldsymbol{L}_{ij} \widehat{\boldsymbol{\pi}}_{ii} \widehat{\boldsymbol{\pi}}_{jj} - \frac{2}{n^2} \sum_{i,j,l=1}^{n} \boldsymbol{K}_{ij} \boldsymbol{L}_{il} \widehat{\boldsymbol{\pi}}_{ii} \boldsymbol{A}_{jl} + \frac{1}{n^2} \sum_{i,j,k,l=1}^{n} \boldsymbol{K}_{ij} \boldsymbol{L}_{kl} \boldsymbol{A}_{ik} \boldsymbol{A}_{jl} \\
&= \frac{1}{n^2} \mathrm{tr}(\boldsymbol{K}\widehat{\boldsymbol{\pi}}\boldsymbol{L}\widehat{\boldsymbol{\pi}} - 2\boldsymbol{K}\widehat{\boldsymbol{\pi}}\boldsymbol{L}\boldsymbol{A}^\top + \boldsymbol{K}\boldsymbol{A}\boldsymbol{L}\boldsymbol{A}^\top),
\end{aligned}$$

where the second to last equality follows from

$$\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{K}_{ij} \boldsymbol{L}_{ij} \widehat{\boldsymbol{\pi}}_{ii} \widehat{\boldsymbol{\pi}}_{jj} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{K}_{ij} (\widehat{\boldsymbol{\pi}}\boldsymbol{L}\widehat{\boldsymbol{\pi}})_{ij} = \frac{1}{n^2} \mathrm{tr}(\boldsymbol{K}\widehat{\boldsymbol{\pi}}\boldsymbol{L}\widehat{\boldsymbol{\pi}}),$$

$$\begin{aligned}
\frac{2}{n^2} \sum_{i,j,l=1}^{n} \boldsymbol{K}_{ij} \boldsymbol{L}_{il} \widehat{\boldsymbol{\pi}}_{ii} \boldsymbol{A}_{jl} &= \frac{2}{n^2} \sum_{j=1}^{n} \sum_{l=1}^{n} \left(\sum_{i=1}^{n} \boldsymbol{K}_{ij} (\widehat{\boldsymbol{\pi}}\boldsymbol{L})_{il} \right) \boldsymbol{A}_{jl} \\
&= \frac{2}{n^2} \sum_{j=1}^{n} \sum_{l=1}^{n} \left(\boldsymbol{K}\widehat{\boldsymbol{\pi}}\boldsymbol{L} \right)_{jl} \boldsymbol{A}_{lj}^\top \\
&= \frac{2}{n^2} \mathrm{tr}(\boldsymbol{K}\widehat{\boldsymbol{\pi}}\boldsymbol{L}\boldsymbol{A}^\top),
\end{aligned}$$

and

$$\begin{aligned}\frac{1}{n^2}\sum_{i,j,k,l=1}^{n}\boldsymbol{K}_{ij}\boldsymbol{L}_{kl}\boldsymbol{A}_{ik}\boldsymbol{A}_{jl} &= \frac{1}{n^2}\sum_{k=1}^{n}\sum_{j=1}^{n}\left(\sum_{i=1}^{n}\boldsymbol{K}_{ij}\boldsymbol{A}_{ik}\right)\left(\sum_{l=1}^{n}\boldsymbol{L}_{kl}\boldsymbol{A}_{jl}\right)\\ &= \frac{1}{n^2}\sum_{k=1}^{n}\sum_{j=1}^{n}\left(\boldsymbol{K}\boldsymbol{A}\right)_{jk}\left(\boldsymbol{L}\boldsymbol{A}^\top\right)_{kj}\\ &= \frac{1}{n^2}\mathrm{tr}(\boldsymbol{K}\boldsymbol{A}\boldsymbol{L}\boldsymbol{A}^\top).\end{aligned}$$

□

Proof of Proposition 5.4

Proof. Eq. (5.7) yields

$$\begin{aligned}\Psi_{c,n} &= \sup_{\omega\in B_1(\mathcal{H})}\frac{1}{n}\sum_{i=1}^{n}\left(\Delta_i\omega(X_i,T_i)\widehat{\pi}^c(X_i,T_i) - \frac{1}{n}\sum_{k=1}^{n}\Delta_k\omega(X_i,T_k)\mathbb{1}_{\{X_k\leq X_i\leq T_k\leq T_i\}}\right)\\ &= \sup_{\omega\in B_1(\mathcal{H})}\frac{1}{n}\sum_{i=1}^{n}\left(\Delta_i\omega(X_i,T_i)\widehat{\boldsymbol{\pi}}^c_{ii} - \sum_{k=1}^{n}\Delta_k\omega(X_i,T_k)\boldsymbol{B}_{ik}\right),\end{aligned}$$

where $\boldsymbol{B}_{ik} = \mathbb{1}_{\{X_k\leq X_i<T_k\leq T_i\}}/n$ and $\widehat{\boldsymbol{\pi}}^c$ is a diagonal matrix with entries $\widehat{\boldsymbol{\pi}}^c_{ii} = \widehat{\pi}^c(X_i,T_i)$.

Then, by following the exact same computations of the proof of Proposition 5.1, we deduce

$$\Psi^2_{c,n} = \frac{1}{n^2}\mathrm{tr}(\boldsymbol{K}\widehat{\boldsymbol{\pi}}\tilde{\boldsymbol{L}}\widehat{\boldsymbol{\pi}} - 2\boldsymbol{K}\widehat{\boldsymbol{\pi}}\tilde{\boldsymbol{L}}\boldsymbol{B}^\top + \boldsymbol{K}\boldsymbol{B}\tilde{\boldsymbol{L}}\boldsymbol{B}^\top),$$

where $\tilde{\boldsymbol{L}}_{ik} = \Delta_i\Delta_k L(T_i,T_k)$. □

Proof of Proposition 5.3

Proof. Under Assumption 5.1, we have that for all $x \leq y$,

$$\begin{aligned}\pi^c_1(x,y) = \mathbb{P}(X\leq x, T\geq y, \Delta = 1) &= \mathbb{E}\left(\mathbb{1}_{\{X\leq x, Y\geq y\}}\mathbb{E}\left(\mathbb{1}_{\{C\geq Y\}}|X,Y\right)\right)\\ &= \mathbb{E}\left(\mathbb{1}_{\{X\leq x, Y\geq y\}}S_{C|X}(Y)\right)\\ &= \int_0^x\int_y^\infty S_{C|X=x'}(y')f_{XY}(x',y')dx',dy',\end{aligned}$$

and

$$\pi^c(x,y) = \mathbb{P}(X \le x, T \ge y) = \mathbb{E}\left(\mathbb{1}_{\{X\le x\}} S_{C|X}(y) S_{Y|X}(y)\right) = \int_0^x S_{C|X=x'}(y) S_{Y|X=x'}(y) f_X(x') dx'.$$

The null hypothesis states $f_{XY}(x,y) = \tilde{f}_X(x)\tilde{f}_Y(y)$ for all $x \le y$ such that $S_{T|X=x}(y) > 0$. Thus

$$\pi_1^c(x,y) = \int_0^x \int_y^\infty S_{C|X=x'}(y') \tilde{f}_X(x') \tilde{f}_Y(y') dx' dy',$$
$$\pi^c(x,y) = \tilde{S}_Y(y) \int_0^x S_{C|X=x'}(y) \tilde{f}_X(x') dx'.$$

By using the previous result, it is easy to see that, under the null,

$$-\pi^c(x,y) \frac{\partial^2}{\partial x \partial y} \pi_1^c(x,y) = \left(\tilde{S}_Y(y) \int_0^x S_{C|X=x'}(y) \tilde{f}_X(x') dx'\right) S_{C|X=x}(y) \tilde{f}_X(x) \tilde{f}_Y(y),$$

and

$$\frac{\partial \pi^c(x,y)}{\partial x} \frac{\partial \pi_1^c(x,y)}{\partial y} = -\left(\tilde{S}_Y(y) S_{C|X=x}(y) \tilde{f}_X(x)\right) \int_0^x S_{C|X=x'}(y) \tilde{f}_X(x') dx' \tilde{f}_Y(y) = -\left(\tilde{S}_Y(y) \int_0^x S_{C|X=x'}(y) \tilde{f}_X(x') dx'\right) S_{C|X=x}(y) \tilde{f}_X(x) \tilde{f}_Y(y),$$

from which it follows that $\rho^c = 0$, and thus $\Psi = 0$. □

Proof of Theorem 5.1

Before proving Theorem 5.1 we give some essential definitions which will be used by our proofs. We will first introduce Lemma 5.1, which is an essential step in the proof of Theorem 5.1. A full proof for Lemma 5.1 is given later in this section.

Our data are considered to live in a common filtrated probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\ge 0}, \mathbb{P})$, where $\mathcal{F}$ is the natural σ-algebra, and $\mathcal{F}_t$ is the filtration generated by

$$\left\{\mathbb{1}_{\{T_i \le s, \Delta_i = 1\}}, \mathbb{1}_{\{T_i \le s, \Delta_i = 0\}}, X_i : 0 \le s \le t, i \in [n]\right\},$$

and the $\mathbb{P}$-null sets of $\mathcal{F}$.

We define $\tau_n = \max\{T_1, \ldots, T_n\}$. For each $i \in [n]$, we define the i-th individual counting and risk processes, $N_i(t)$ and $Y_i(t)$, by $N_i(t) = \Delta_i \mathbb{1}_{\{T_i \leq t\}}$ and $Y_i(t) = \mathbb{1}_{\{T_i \geq t\}}$, respectively. For each individual i, we define the process $(M_i(t))_{t\geq 0}$ by

$$M_i(t) = N_i(t) - \int_{(0,t]} \mathbb{1}_{\{X_i \leq s\}} Y_i(s) \tilde{\lambda}_Y(s) ds.$$

It is standard to verify that $M_i(t)$ is an $(\mathcal{F}_t)$-martingale under the null hypothesis, and that, for any bounded predictable process $(H_i(t))_{t\geq 0}$, $\int_{(0,t]} H_i(s) dM_i(s)$ is also an $(\mathcal{F}_t)$-martingale under the null hypothesis.

Let (T_1', Δ_1', X_1') and (T_2', Δ_2', X_2') be independent copies of our data $((T_i, \Delta_i, X_i))_{i=1}^n$. Sometimes our results are written in terms of $\tilde{\mathbb{E}}$ which is defined by $\tilde{\mathbb{E}}(\cdot) = \mathbb{E}\left(\cdot | ((T_i, \Delta_i, X_i))_{i=1}^n\right)$. Additionally, we denote by Y_1' and Y_2', the individual risk functions associated to T_1' and T_2', which are defined by $Y_1'(t) = \mathbb{1}_{\{T_1' \geq t\}}$ and $Y_2'(t) = \mathbb{1}_{\{T_2' \geq t\}}$, respectively. Finally, we define $Z_i(t) = \omega(X_i, t)\mathbb{1}_{\{X_i \leq t\}}$ for all $i \in [n]$, and, based on (T_1', Δ_1', X_1') and (T_2', Δ_2', X_2'), we define $Z_1'(t) = \omega(X_1', t)\mathbb{1}_{\{X_1' \leq t\}}$ and $Z_2'(t) = \omega(X_2', t)\mathbb{1}_{\{X_2' \leq t\}}$.

Lemma 5.1. *Assume that $\mathfrak{K}$ is bounded. Then, under the null hypothesis*

$$\sqrt{n}\Psi_{n,c} = \sup_{\omega \in B_1(\mathcal{H})} \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^{\tau_n} \left(Z_i(t)\pi^c(X_i, t) - \tilde{\mathbb{E}}\left(Z_1'(t) Y_1'(t) \mathbb{1}_{\{X_i \leq X_1'\}} \right) \right) dM_i(t) + o_p(1). \tag{5.11}$$

Proof of Theorem 5.1

By the reproducing property, we have $Z_i(t) = \langle \omega, \mathfrak{K}((X_i, t), \cdot)\rangle_{\mathcal{H}} \mathbb{1}_{\{X_i \leq t\}}$ and $Z_1'(t) = \langle \omega, \mathfrak{K}((X_1', t), \cdot)\rangle_{\mathcal{H}} \mathbb{1}_{\{X_1' \leq t\}}$; together with Eq.(5.11) in Lemma 5.1, write

$$\begin{aligned}
&\left(Z_i(t)\pi^c(X_i, t) - \tilde{\mathbb{E}}\left(Z_1'(t) Y_1'(t) \mathbb{1}_{\{X_i \leq X_1'\}} \right) \right) \\
&= \left(\langle \omega, \mathfrak{K}((X_i, t), \cdot)\rangle_{\mathcal{H}} \mathbb{1}_{\{X_i \leq t\}} \pi^c(X_i, t) - \tilde{\mathbb{E}}\left(\langle \omega, \mathfrak{K}((X_1', t), \cdot)\rangle_{\mathcal{H}} \mathbb{1}_{\{X_1' \leq t\}} Y_1'(t) \mathbb{1}_{\{X_i \leq X_1'\}} \right) \right) \\
&= \left\langle \omega, \mathfrak{K}((X_i, t), \cdot) \mathbb{1}_{\{X_i \leq t\}} \pi^c(X_i, t) - \tilde{\mathbb{E}}\left(\mathfrak{K}((X_1', t), \cdot) \mathbb{1}_{\{X_1' \leq t\}} Y_1'(t) \mathbb{1}_{\{X_i \leq X_1'\}} \right) \right\rangle_{\mathcal{H}},
\end{aligned}$$

where the second equality follows from the linearity of expectation, assuming Bochner integrability of the feature map (true for bounded $\mathfrak{K}$). To ease notation, we define the functions $a : \mathbb{R}^2 \to \mathbb{R}$ and $b : \mathbb{R}^3 \to \mathbb{R}$ by $a(X_i, t) = \mathbb{1}_{\{X_i \leq t\}} \pi^c(X_i, t)$

and $b(X_1', X_i, t) = Y_1'(t)\mathbb{1}_{\{X_i \leq X_1' \leq t\}}$, respectively, and write

$$\begin{aligned}
&\left(Z_i(t)\pi^c(X_i,t) - \tilde{\mathbb{E}}\left(Z_1'(t)Y_1'(t)\mathbb{1}_{\{X_i \leq X_1'\}}\right)\right) \\
&= \left\langle \omega, \mathfrak{K}((X_i,t),\cdot)a(X_i,t) - \tilde{\mathbb{E}}\left(\mathfrak{K}((X_1',t),\cdot)b(X_1',X_i,t)\right)\right\rangle_{\mathcal{H}}. \qquad (5.12)
\end{aligned}$$

From the previous result, we can see that

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau_n}\left(Z_i(t)\pi^c(X_i,t) - \tilde{\mathbb{E}}\left(Z_1'(t)Y_1'(t)\mathbb{1}_{\{X_i \leq X_1'\}}\right)\right)dM_i(t) = \\
&\left\langle \omega, \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau_n}(\mathfrak{K}((X_i,t),\cdot)a(X_i,t) - \tilde{\mathbb{E}}\left(\mathfrak{K}((X_1',t),\cdot)b(X_1',X_i,t)\right))dM_i(t)\right\rangle_{\mathcal{H}}
\end{aligned}$$

and thus

$$\begin{aligned}
&\sup_{\omega \in B_1(\mathcal{H})}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau_n}\left(Z_i(t)\pi^c(X_i,t) - \tilde{\mathbb{E}}\left(Z_1'(t)Y_1'(t)\mathbb{1}_{\{X_i \leq X_1'\}}\right)\right)dM_i(t)\right)^2 \\
&\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau_n}\left(\mathfrak{K}((X_i,t),\cdot)a(X_i,t) - \tilde{\mathbb{E}}\left(\mathfrak{K}((X_1',t),\cdot)b(X_1',X_i,t)\right)\right)dM_i(t)\right\|_{\mathcal{H}}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}J((T_i,\Delta_i,X_i),(T_j,\Delta_j,X_j)), \qquad (5.13)
\end{aligned}$$

where the function $J : (\mathbb{R} \times \{0,1\} \times \mathbb{R})^2 \to \mathbb{R}$ is defined by

$$J((s,r,x),(s',r',x')) = \int_0^s\int_0^{s'}A((t,x),(t',x'))dm_{s',r',x'}(t')dm_{s,r,x}(t),$$

$dm_{s,r,x}(t) = r\delta_s(t) - \mathbb{1}_{\{s \geq t\}}\mathbb{1}_{\{x \leq t\}}\tilde{\lambda}_Y(t)dt$ (notice that $dM_i(t) = dm_{T_i,\Delta_i,X_i}(t)$), and $A : (\mathbb{R} \times \mathbb{R})^2 \to \mathbb{R}$ is defined as

$$\begin{aligned}
&A((t,x),(t',x')) \\
&= \Big\langle \mathfrak{K}((x,t),\cdot)a(x,t) - \tilde{\mathbb{E}}\left(\mathfrak{K}((X_1',t),\cdot)b(X_1',x,t)\right) \\
&\qquad\qquad , \mathfrak{K}((x',t'),\cdot)a(x',t') - \tilde{\mathbb{E}}\left(\mathfrak{K}((X_2',t'),\cdot)b(X_2',x',t')\right)\Big\rangle_{\mathcal{H}} \\
&= \mathfrak{K}((x,t),(x',t'))a(x,t)a(x',t') - \tilde{\mathbb{E}}(\mathfrak{K}((X_1',t),(x',t'))b(X_1',x,t)a(x',t')) \\
&\quad - \tilde{\mathbb{E}}(\mathfrak{K}((x,t),(X_2',t'))a(x,t)b(X_2',x',t')) \\
&\quad + \tilde{\mathbb{E}}(\mathfrak{K}((X_1',t),(X_2',t'))b(X_1',x,t)b(X_2',x',t')).
\end{aligned}$$

It can be verified that the sum in Eq. (5.13) is a degenerate V-statistic. Indeed, the

degeneracy property can be verified by noticing that

$$\begin{aligned}
&\mathbb{E}(J((T_i,\Delta_i,X_i),(s',r',x')))\\
&=\mathbb{E}\left(\int_0^{T_i}\left(\int_0^{s'}A((t,X_i),(t',x'))dm_{s',r',x}(t')\right)dM_i(t)\right)\\
&=\mathbb{E}(Q(T_i)),
\end{aligned}$$

where $Q(s)=\int_0^s\left(\int_0^{s'}A((t,X_i),(t',x'))dm_{s',r',x}(t')\right)dM_i(t)$ is a zero mean $(\mathcal{F}_s)$-martingale, and thus, by the optional stopping Theorem, $\mathbb{E}(Q(T_i))=\mathbb{E}(Q(0))=0$. Then, by [Koroljuk and Borovskich, 1994, Theorem 4.3.2], we deduce

$$\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n J((T_i,\Delta_i,X_i),(T_j,\Delta_j,X_j))\xrightarrow{\mathcal{D}}\mathbb{E}(J((T_1,\Delta_1,X_1),(T_1,\Delta_1,X_1)))+\mathcal{Y},$$

where $\mathcal{Y}=\sum_{i=1}^\infty\lambda_i(\xi_i^2-1)$, $\xi_1,\xi_2,\ldots$ are independent standard normal random variables, and $\lambda_1,\lambda_2,\ldots$ are positive constants.

The previous result, together with Lemma 5.1, allow us to deduce

$$\Psi_{c,n}^2\xrightarrow{\mathcal{D}}\mu+\mathcal{Y},$$

where $\mu=\mathbb{E}(J((T_1,\Delta_1,X_1),(T_1,\Delta_1,X_1)))$. Notice that all integrability conditions are satisfied as we assume the reproducing kernel is bounded.

Proof of Lemma 5.1

In order to prove Lemma of 5.1, we require some intermediate results.

Recall that our test-statistic is computed as the supremum over $\omega\in B_1(\mathcal{H})$ of sums

$$\frac{1}{n}\sum_{i=1}^n\Delta_i\omega(X_i,T_i)\widehat{\pi}^c(X_i,T_i)-\frac{1}{n^2}\sum_{i=1}^n\sum_{k=1}^n\Delta_k\omega(X_i,T_k)\mathbb{1}_{\{X_k\leq X_i<T_k\leq T_i\}}.$$

By using the notation introduced at the beginning of Section 5.A, the previous sum

can be rewritten as

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left(\Delta_i\omega(X_i,T_i)\widehat{\pi}^c(X_i,T_i)-\frac{1}{n}\sum_{k=1}^{n}\Delta_i\omega(X_k,T_i)\mathbb{1}_{\{X_i\leq X_k<T_i\leq T_k\}}\right)\\
&=\frac{1}{n}\sum_{i=1}^{n}\int_0^{T_i}\left(\omega(X_i,y)\mathbb{1}_{\{X_i\leq y\}}\widehat{\pi}^c(X_i,y)-\frac{1}{n}\sum_{k=1}^{n}\omega(X_k,y)\mathbb{1}_{\{X_k\leq y\}}\mathbb{1}_{\{y\leq T_k\}}\mathbb{1}_{\{X_i\leq X_k\}}\right)dN_i(y)\\
&=\frac{1}{n}\sum_{i=1}^{n}\int_0^{T_i}\left(Z_i(y)\widehat{\pi}^c(X_i,y)-\frac{1}{n}\sum_{k=1}^{n}Z_k(y)Y_k(y)\mathbb{1}_{\{X_i\leq X_k\}}\right)dN_i(y)\\
&=\frac{1}{n}\sum_{i=1}^{n}\int_0^{T_i}H_i(y)dN_i(y),
\end{aligned}$$

where $H_i(y)=Z_i(y)\widehat{\pi}^c(X_i,y)-\frac{1}{n}\sum_{k=1}^{n}Z_k(y)Y_k(y)\mathbb{1}_{\{X_i\leq X_k\}}$. Thus,

$$\Psi_{n,c}=\sup_{\omega\in B_1(\mathcal{H})}\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_n}H_i(y)dN_i(y), \tag{5.14}$$

where recall that $\tau_n=\max\{T_1,\ldots,T_n\}$.

Proposition 5.6. *Assume that $\mathfrak{K}$ is bounded. Then, under the null hypothesis, the process $(W(t))_{t\geq 0}$, defined by $W(t)=\frac{1}{n}\sum_{i=1}^{n}\int_0^t H_i(y)dN_i(y)$, is an $(\mathcal{F}_t)$-martingale, and can be rewritten as*

$$W(t)=\frac{1}{n}\sum_{i=1}^{n}\int_0^t H_i(y)dM_i(y).$$

Notice that the previous Proposition, and Eq. (5.14) suggest the result of Lemma 5.1. It remains to prove that the process $H_i(y)$ may be approximated by its "population limit". We prove this result in two steps in the two lemmas below.

Lemma 5.2. *Assume that $\mathfrak{K}$ is bounded. Then, under the null hypothesis*

$$\sup_{\omega\in B_1(\mathcal{H})}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau_n}Z_i(y)\left(\widehat{\pi}^c(X_i,y)-\pi^c(X_i,y)\right)dM_i(y)=o_p(1),$$

Lemma 5.3. *Assume that $\mathfrak{K}$ is bounded. Then, under the null hypothesis*

$$\sup_{\omega\in B_1(\mathcal{H})}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau_n}\left(\frac{1}{n}\sum_{j=1}^{n}Z_j(y)Y_j(y)\mathbb{1}_{\{X_i\leq X_j\}}-\tilde{\mathbb{E}}(Z_1'(y)Y_1'(y)\mathbb{1}_{\{X_i\leq X_1'\}})\right)dM_i(y)=o_p(1),$$

Proof of Lemma 5.1: Eq. (5.14) and Lemma 5.6 yield

$$\sqrt{n}\Psi_{n,c} = \sup_{\omega\in B_1(\mathcal{H})} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^{\tau_n} H_i(y) dM_i(y),$$

where (recall) $H_i(y) = Z_i(y)\widehat{\pi}^c(X_i, y) - \frac{1}{n}\sum_{k=1}^{n} Z_k(y)Y_k(y)\mathbb{1}_{\{X_i\leq X_k\}}$. Notice that to obtain the result, we need to replace $\widehat{\pi}^c$ by its population version π^c, and, given (T_i, Δ_i, X_i), we need to replace the i.i.d. sum $\frac{1}{n}\sum_{k=1}^{n} Z_k(y)Y_k(y)\mathbb{1}_{\{X_i\leq X_k\}}$ by its limit, which is given by $\tilde{E}(Z_1'(y)Y_1'(y)\mathbb{1}_{\{X_i\leq X_1'\}})$. By the triangular inequality, this result follows from lemmas 5.2 and 5.3.

Proof of Proposition 5.6

Recall that $dM_i(y) = dN_i(y) - \mathbb{1}_{\{X_i\leq y\}}Y_i(y)\tilde{\lambda}_Y(y)dy$. A straightforward computation verifies $\frac{1}{n}\sum_{i=1}^{n}\int_0^t H_i(y)\mathbb{1}_{\{X_i\leq y\}}Y_i(y)\tilde{\lambda}_Y(y)dy = 0$ for all $t \geq 0$, and thus

$$W(t) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t H_i(y)dM_i(y).$$

Also, notice that $(H_i(t))_{t\geq 0}$ (with $\omega \in B_1(\mathcal{H})$) is bounded and $(\mathcal{F}_t)$-predictable, and that $M_i(t)$ is an $(\mathcal{F}_t)-$ martingale under the null hypothesis. Then, by standard martingale results we deduce that $(W(t))_{t\geq 0}$ is an $(\mathcal{F}_t)$-martingale.

Proof of Lemma 5.2

Observe that

$$Z_i(t)(\widehat{\pi}^c(X_i,t) - \pi^c(X_i,t)) = \langle\omega, \mathfrak{K}((X_i,t),\cdot)\rangle_{\mathcal{H}}\, \mathbb{1}_{\{X_i\leq t\}}(\widehat{\pi}^c(X_i,t) - \pi^c(X_i,t))$$

since $Z_i(t,\omega) = \omega(X_i,t)\mathbb{1}_{\{X_i\leq t\}} = \langle\omega, \mathfrak{K}((X_i,t),\cdot)\rangle_{\mathcal{H}}\mathbb{1}_{\{X_i\leq t\}}$ due to the reproducing property.

Then,

$$\begin{aligned}
&\sup_{\omega\in B_1(\mathcal{H})}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau_n} Z_i(t)\left(\widehat{\pi}^c(X_i,t) - \pi^c(X_i,t)\right)dM_i(t)\right)^2 \\
&= \sup_{\omega\in B_1(\mathcal{H})}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau_n} \langle\omega, \mathfrak{K}((X_i,t),\cdot)\rangle_{\mathcal{H}}\, \mathbb{1}_{\{X_i\leq t\}}(\widehat{\pi}^c(X_i,t) - \pi^c(X_i,t))dM_i(t)\right)^2 \\
&= \sup_{\omega\in B_1(\mathcal{H})}\left\langle\omega, \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau_n} \mathfrak{K}((X_i,t),\cdot)\mathbb{1}_{\{X_i\leq t\}}(\widehat{\pi}^c(X_i,t) - \pi^c(X_i,t))dM_i(t)\right\rangle_{\mathcal{H}}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n}\int_0^{\tau_n}\int_0^{\tau_n} J((X_i,t),(X_k,s))dM_i(t)dM_k(s),
\end{aligned}$$

where

$$J((X_i,t),(X_k,s)) \\ = \mathfrak{K}((X_i,t),(X_k,s))\mathbb{1}_{\{X_i\leq t\}}\mathbb{1}_{\{X_k\leq s\}}(\widehat{\pi}^c(X_i,t)-\pi^c(X_i,t))(\widehat{\pi}^c(X_k,s)-\pi^c(X_k,s)) \tag{5.15}$$

Define the process $(Q(y))_{y\geq 0}$ by

$$Q(y)=\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n}\int_0^y\int_0^y J((X_i,t),(X_k,s))dM_i(t)dM_k(s),$$

and notice that we wish to prove that $Q(\tau_n)=o_p(1)$. Let $\delta>0$, then, by Markov's inequality,

$$\mathbb{P}(Q(\tau_n)>\delta)\leq\frac{\mathbb{E}(Q(\tau_n))}{\delta}=\frac{\mathbb{E}(Q_D(\tau_n))}{\delta}+\frac{2\mathbb{E}(Q_{D^c}(\tau_n))}{\delta},$$

where the last equality holds since, by symmetry, $Q(y)=Q_D(y)+2Q_{D^c}(y)$, where

$$Q_D(y)=\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n}\int_0^y\int_0^y \mathbb{1}_{\{s=t\}}J((X_i,t),(X_k,s))dM_i(t)dM_k(s), \tag{5.16}$$

and

$$Q_{D^c}(y)=\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n}\int_0^y\int_{(0,s)} J((X_i,t),(X_k,s))dM_i(t)dM_k(s).$$

By [Fernández and Rivera, 2020, Theorem 6.8], $Q_{D^c}(y)$ is an $(\mathcal{F}_y)$-martingale, and, by the optional stopping theorem, $\mathbb{E}(Q_{D^c}(\tau_n))=\mathbb{E}(Q_{D^c}(0))=0$. Thus

$$\mathbb{P}(Q(\tau_n)>\delta)\leq\frac{\mathbb{E}(Q_D(\tau_n))}{\delta},$$

where

$$\begin{aligned} Q_D(\tau_n) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n}\int_0^{\tau_n} J((X_i,t),(X_k,t))d[M_i,M_k](t) \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_n} J((X_i,t),(X_i,t))d[M_i](t) \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_n} J((X_i,t),(X_i,t))N_i(t) \\ &= \frac{1}{n}\sum_{i=1}^{n}\Delta_i J((X_i,T_i),(X_i,T_i)) \end{aligned}$$

follows from considering continuous survival and censoring times.

We finish the proof by proving $\mathbb{E}(Q_D(\tau_n)) \to 0$ as n tends to infinity. Observe that

$$\begin{aligned} \mathbb{E}(Q_D(\tau_n)) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\Delta_i J((X_i,T_i),(X_i,T_i))\right) \\ &\leq c_1\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\Delta_i \mathbb{1}_{\{X_i\leq T_i\}}(\widehat{\pi}^c(X_i,T_i)-\pi^c(X_i,T_i))^2\right) \end{aligned}$$

follows from substituting the function J with the expression given in Eq. (5.15), and by assuming the reproducing kernel is bounded by some constant $c_1 > 0$. By Proposition 5.5, the sum $\frac{1}{n}\sum_{i=1}^{n}\Delta_i \mathbb{1}_{\{X_i\leq T_i\}}(\widehat{\pi}^c(X_i,T_i)-\pi^c(X_i,T_i))^2$ converges to 0 almost surely, and it is bounded by some constant $c > 0$, then the desired result follows from an application of dominated convergence.

Proof of Lemma 5.3

Notice that, by the reproducing property,

$$\begin{aligned} &\frac{1}{n}\sum_{j=1}^{n} Z_j(t)Y_j(t)\mathbb{1}_{\{X_i\leq X_j\}} - \tilde{\mathbb{E}}(Z_1'(t)Y_1'(t)\mathbb{1}_{\{X_i\leq X_1'\}}) \\ &= \frac{1}{n}\sum_{j=1}^{n}\langle\omega,\mathfrak{K}((X_j,t),\cdot)\rangle_{\mathcal{H}}Y_j(t)\mathbb{1}_{\{X_i\leq X_j\leq t\}} - \tilde{\mathbb{E}}\left(\langle\omega,\mathfrak{K}((X_1',t),\cdot)\rangle_{\mathcal{H}}Y_1'(t)\mathbb{1}_{\{X_i\leq X_1'\leq t\}}\right) \\ &= \left\langle\omega,\frac{1}{n}\sum_{j=1}^{n}\mathfrak{K}((X_j,t),\cdot)Y_j(t)\mathbb{1}_{\{X_i\leq X_j\leq t\}} - \tilde{\mathbb{E}}\left(\mathfrak{K}((X_1',t),\cdot)Y_1'(t)\mathbb{1}_{\{X_i\leq X_1'\leq t\}}\right)\right\rangle_{\mathcal{H}}. \end{aligned}$$

To ease notation, we define $a_{ij}(t) = Y_j(t)\mathbb{1}_{\{X_i\leq X_j\leq t\}}$ and $b_{i1}'(t) = Y_1'(t)\mathbb{1}_{\{X_i\leq X_1'\leq t\}}$ (similarly, we define $b_{i2}'(t) = Y_2'(t)\mathbb{1}_{\{X_i\leq X_2'\leq t\}}$, where recall that (T_1',Δ_1',X_1') and

(T_2', Δ_2', X_2') are independent copies of our data). Then, the previous term can be rewritten as

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n} Z_j(t)Y_j(t)\mathbb{1}_{\{X_i \leq X_j\}} - \tilde{\mathbb{E}}(Z_1'(t)Y_1'(t)\mathbb{1}_{\{X_i \leq X_1'\}}) \\
&= \left\langle \omega, \frac{1}{n}\sum_{j=1}^{n} \mathfrak{K}((X_j,t),\cdot)a_{ij}(t) - \tilde{\mathbb{E}}\left(\mathfrak{K}((X_1',t),\cdot)b_{i1}'(t)\right) \right\rangle_{\mathcal{H}}.
\end{aligned}$$

By using the fact we take supremum on the unit ball of an RKHS, it is not difficult to deduce,

$$\begin{aligned}
&\sup_{\omega \in B_1(\mathcal{H})} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^{\tau_n} \left(\frac{1}{n}\sum_{j=1}^{n} Z_j(t)Y_j(t)\mathbb{1}_{\{X_i \leq X_j\}} - \tilde{\mathbb{E}}(Z_1'(t)Y_1'(t)\mathbb{1}_{\{X_i \leq X_1'\}}) \right) dM_i(y) \right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n} \int_0^{\tau_n}\int_0^{\tau_n} J((X_i,t),(X_k,s))dM_i(t)dM_k(s), \qquad (5.17)
\end{aligned}$$

where

$$\begin{aligned}
&J((X_i,t),(X_k,s)) \\
=&\frac{1}{n^2}\sum_{j=1}^{n}\sum_{l=1}^{n} \mathfrak{K}\left((X_j,t),(X_l,s)\right) a_{ij}(t)a_{kl}(s) - \frac{1}{n}\sum_{l=1}^{n} \tilde{\mathbb{E}}(\mathfrak{K}((X_1',t),(X_l,s))b_{i1}'(t)a_{kl}(s) \\
&- \frac{1}{n}\sum_{j=1}^{n} \tilde{\mathbb{E}}(\mathfrak{K}((X_j,t),(X_2',s))a_{ij}(t)b_{k2}'(s) + \tilde{\mathbb{E}}(\mathfrak{K}((X_1',t),(X_2',s))b_{i1}'(t)b_{k2}'(s)), \\
&\qquad (5.18)
\end{aligned}$$

Following the same steps of the proof of Lemma 5.2, we can prove that Eq. (5.17) is $o_p(1)$ by proving that

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} J((X_i,T_i),(X_i,T_i))\right) \to 0. \qquad (5.19)$$

For this purpose, first observe that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} J((X_i,T_i),(X_i,T_i)) =& \frac{1}{n^3}\sum_{i,j,l=1}^{n} \mathfrak{K}\left((X_j,T_i),(X_l,T_i)\right) a_{ij}(T_i)a_{il}(T_i) \\ &- \frac{2}{n^2}\sum_{i,l=1}^{n} \tilde{\mathbb{E}}(\mathfrak{K}((X_1',T_i),(X_l,T_i))b_{i1}'(T_i)a_{il}(T_i) \\ &+ \frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbb{E}}(\mathfrak{K}((X_1',T_i),(X_2',T_i))b_{i1}'(T_i)b_{i2}'(T_i).\end{aligned}$$

Each sum on the right-hand side of the previous equation is a V-statistic of order 3, 2 and 1, respectively. It can easily be seen that they all converge to the same limit. Consequently, the law of large numbers for V-statistics implies that

$$\frac{1}{n}\sum_{i=1}^{n} J((X_i,T_i),(X_i,T_i)) \to 0$$

almost surely. Since the reproducing kernel is assumed to be bounded and, thus, the sum is bounded as well, we can deduce, finally, Eq. (5.19) from the dominated convergence theorem.

Proof of Theorem 5.2

The consistency proof relies on the interpretation of the test statistic $\Psi_{c,n}$ and the KQIC Ψ_c as the Hilbert space distances of embeddings of certain positive measures. These distances measure the degree of (quasi)-dependence. In this spirit, this approach is connected to the well-established Hilbert Schmidt Independence Criterion, see e.g. Chwialkowski et al. [2014]; Gretton et al. [2008]; Meynaoui et al. [2019]; Sejdinovic et al. [2013].

Now, let us become more concrete and introduce the following measures ν_0 and ν_1 on R_+^2 given by

$$\begin{aligned}\nu_0(dx,dy) &= \pi^c(x,y)\pi_1^c(dx,dy) \\ &= \pi^c(x,y)S_{C|X=x}(y)f_{XY}(x,y)dxdy,\end{aligned}$$

$$\begin{aligned}\nu_1(dx,dy) &= \pi^c(dx,y)\pi_1^c(x,dy) \\ &= \left(S_{Y|X=x}(y)S_{C|X=x}(y)f_X(x)\right)\left(\int_0^x S_{C|X=t}(y)f_{XY}(t,y)dt\right)dxdy\end{aligned}$$

as well as their empirical counterparts ν_0^n and ν_1^n defined as

$$\nu_0^n(dx,dy) = \frac{\widehat{\pi}^c(x,y)}{n}\sum_{i=1}^n \Delta_i \delta_{X_i}(x)\delta_{T_i}(y)$$
$$\nu_1^n(dx,dy) = \frac{\mathbb{1}_{\{x\le y\}}}{n^2}\left(\sum_{i=1}^n \delta_{X_i}(x)\mathbb{1}_{\{T_i\ge y\}}\right)\left(\sum_{k=1}^n \Delta_k\delta_{T_k}(y)\mathbb{1}_{\{X_k\le x\}}\right).$$

Moreover, set $\widehat{\rho}^c = \nu_0^n - \nu_1^n$, which is the empirical counterpart of the measure induced by the density ρ^c. Then the embeddings of the (empirical) measures into the underlying RKHS are given by

$$\phi_j(\cdot) = \iint_{x\le y} \mathfrak{K}((x,y),\cdot)\nu_j(dx,dy), \text{ and } \quad \phi_j^n(\cdot) = \iint_{x\le y} \mathfrak{K}((x,y),\cdot)\nu_j^n(dx,dy).$$

By straightforward calculations, we obtain

$$\Psi_{c,n}^2 = \sup_{\omega\in B_1(\mathcal{H})}\left(\iint_{x\le y}\omega(x,y)\widehat{\rho}^c(dx,dy)\right)^2 = \|\phi_0^n - \phi_1^n\|_{\mathcal{H}}^2$$

and

$$\Psi_c^2 = \sup_{\omega\in B_1(\mathcal{H})}\left(\iint_{x\le y}\omega(x,y)\rho^c(dx,dy)\right)^2 = \|\phi_0 - \phi_1\|_{\mathcal{H}}^2.$$

Consequently, the first part of Theorem 5.2 follows from convergence of the aforementioned distances:

Lemma 5.4. *We have* $\|\phi_0^n - \phi_1^n\|_{\mathcal{H}}^2 \to \|\phi_0 - \phi_1\|_{\mathcal{H}}^2$ *in probability.*

The proof of Lemma 5.4 is given below. For the second part of Theorem 5.2, recall that by assumption the chosen kernel $\mathfrak{K}$ is c_0-universal and, thus, the embedding of finite signed Borel measures is injective, see [Sriperumbudur et al., 2010] for details. In particular, $\Psi_c^2 = \|\phi_0 - \phi_1\|_{\mathcal{H}}^2$ equals zero if and only if $\nu_0 \equiv \nu_1$, or equivalently $\rho^c(x,y) = 0$ for almost all $x \le y$. Consequently, it remains to verify the following lemma, which is proven below.

Lemma 5.5. $\rho^c(x,y) = 0$ *for almost all* $x \le y$ *if and only if the null hypothesis of quasi independence is fulfilled.*

Proof of Lemma 5.4

First, observe that

$$\Psi_{c,n}^2 = \sup_{\omega \in B_1(\mathcal{H})} \left(\iint_{x \leq y} \omega(x,y) \widehat{\rho}^c(dx,dy) \right)^2 = \|\phi_0^n - \phi_1^n\|_{\mathcal{H}}^2 = V_{0,0} - 2V_{0,1} + V_{1,1},$$

where

$$V_{0,0} = \|\phi_0^n\|_{\mathcal{H}}^2 = \frac{1}{n^2} \sum_{j,i=1}^n \mathfrak{K}((X_i,T_i),(X_j,T_j)) \Delta_i \Delta_j \widehat{\pi}^c(X_i,T_i) \widehat{\pi}^c(X_j,T_j),$$

$$V_{0,1} = \langle \phi_0^n, \phi_1^n \rangle_{\mathcal{H}} = \frac{1}{n^3} \sum_{i,j,k=1}^n \mathfrak{K}((X_i,T_i),(X_j,T_k)) \Delta_i \Delta_k \widehat{\pi}^c(X_i,T_i) \mathbb{1}_{\{X_k \leq X_j < T_k \leq T_j\}},$$

$$V_{1,1} = \|\phi_1^n\|_{\mathcal{H}}^2 = \frac{1}{n^4} \sum_{i,j,k,\ell=1}^n \mathfrak{K}((X_i,T_\ell),(X_j,T_k)) \Delta_k \Delta_\ell \mathbb{1}_{\{X_\ell \leq X_i < T_\ell \leq T_i\}} \mathbb{1}_{\{X_k \leq X_j < T_k \leq T_j\}}.$$

By Proposition 5.5 we can replace $\widehat{\pi}_c$ by π_c for all asymptotic considerations, a detailed explanation for $V_{0,0}$ is given below. Thus, $V_{0,0}$, $V_{0,1}$ and $V_{1,1}$ are asymptotically equivalent to V-statistics of order 2, 3 and 4, respectively. For the desired statement, it remains to show that (i) $V_{0,0} \to \|\phi_0\|_{\mathcal{H}}^2$ (ii) $V_{0,1} \to \langle \phi_0, \phi_1 \rangle_{\mathcal{H}}$ (iii) $V_1 \to \|\phi_1\|_{\mathcal{H}}^2$. All three convergences follow from the strong law of large numbers for V-statistics and Proposition 5.5, as explained exemplary for (i):

Since the kernel $\mathfrak{K}$ and $\widehat{\pi}^c$ are bounded by some $c_1 > 0$ and 1, respectively, we can deduce from Proposition 5.5 and the triangular inequality that almost surely

$$\begin{aligned}
&\left| \frac{1}{n^2} \sum_{j,i=1}^n \mathfrak{K}((X_i,T_i),(X_j,T_j)) \Delta_i \Delta_j \Big(\widehat{\pi}^c(X_i,T_i) \widehat{\pi}^c(X_j,T_j) - \pi^c(X_i,T_i) \pi^c(X_j,T_j) \Big) \right| \\
&\leq c_1 \frac{1}{n^2} \sum_{j,i=1}^n \Delta_i \Delta_j \left(|\widehat{\pi}^c(X_i,T_i) - \pi^c(X_i,T_i)| + |\widehat{\pi}^c(X_j,T_j) - \pi^c(X_j,T_j)| \right) \\
&\leq \frac{2c_1}{n^2} \sum_{j,i=1}^n \Delta_i \Delta_j \left| \widehat{\pi}^c(X_i,T_i) - \pi^c(X_i,T_i) \right| \\
&\leq \frac{2c_1}{n} \sum_{i=1}^n \Delta_i \left| \widehat{\pi}^c(X_i,T_i) - \pi^c(X_i,T_i) \right| \to 0.
\end{aligned}$$

Thus, we can replace for further asymptotic investigations $\widehat{\pi}^c$ by π^c. Finally, by the

strong law of large numbers

$$V_{0,0} \to \mathbb{E}\Big(\mathfrak{K}((X_1,T_1),(X_2,T_2))\Delta_1\Delta_2\pi^c(X_1,T_1)\pi^c(X_2,T_2)\Big)$$
$$= \iint_{x_1<t_2}\iint_{x_2<t_2} \mathfrak{K}((x_1,t_1),(x_2,t_2))d\nu_0(x_1,t_1)d\nu_0(x_2,t_2).$$

Proof of Lemma 5.5

The first implication was already shown in the proof of Proposition 5.3. Now, assume that $\rho^c = 0$. Then

$$\pi^c(x,y)\frac{\partial^2\pi_1^c(x,y)}{\partial x\partial y} = \frac{\partial\pi^c(x,y)}{\partial x}\frac{\partial\pi_1^c(x,y)}{\partial y}. \tag{5.20}$$

Define $M(x,y) = \frac{\partial\pi_1^c(x,y)}{\partial y} = \int_0^x S_{C|X=x'}(y)f_{XY}(x',y)dx'$, then Eq. (5.20) can be rewritten as

$$\pi^c(x,y)\frac{\partial M(x,y)}{\partial x} = \frac{\partial\pi^c(x,y)}{\partial x}M(x,y). \tag{5.21}$$

Set $Q(x,y) = \mathbb{1}_{\{M(x,y)\neq 0\}}\pi^c(x,y)/M(x,y)$. From (5.20) we can conclude that $M(x,y) = 0$ implies $\pi^c(x,y) = 0$ or

$$0 = \frac{\partial^2\pi_1^c(x,y)}{\partial x\partial y} = -S_{C|X=x}(y)f_{XY}(x,y).$$

But, the right-hand side of the equation is positive for all observable (x,y), i.e. such that $S_{C|X=x}(y), f(x,y), f(x) > 0$. Note that only these pairs are relevant and, thus, we restrict to them subsequently. Thus, $\pi^c(x,y) = Q(x,y)M(x,y)$ and differentiation with respect to x leads to

$$\begin{aligned}\frac{\partial\pi^c(x,y)}{\partial x} &= \frac{\partial Q(x,y)}{\partial x}M(x,y) + Q(x,y)\frac{\partial M(x,y)}{\partial x}\\ &= \frac{\partial Q(x,y)}{\partial x}M(x,y) + \frac{\pi^c(x,y)}{M(x,y)}\frac{\partial M(x,y)}{\partial x}\\ &= \frac{\partial Q(x,y)}{\partial x}M(x,y) + \frac{\partial\pi^c(x,y)}{\partial x}.\end{aligned}$$

Thus, $\partial Q(x,y)/\partial x = 0$ for all (observable) $x \le y$. In particular, Q does not depend on x, and we can write $Q(y)$ instead of $Q(x,y)$. Consequently, we can deduce from the definitions of Q, M and π^c that

$$-Q(y)\int_0^x S_{C|X=x'}(y)f_{XY}(x',y)dx' = \int_0^x S_{Y|X=x'}(y)S_{C|X=x'}(y)f_X(x')dx'.$$

In particular, we can deduce that for all observable $x \leq y$

$$-Q(y)S_{C|X=x}(y)f_{XY}(x,y) = S_{Y|X=x}(y)S_{C|X=x}(y)f_X(x).$$

From this we obtain

$$\begin{aligned} f_{XY}(t,y) &= -Q(y)^{-1}S_{Y|X=t}(y)f_X(t) \\ \Leftrightarrow f_X(t)f_{Y|X=t}(y) &= -Q(y)^{-1}S_{Y|X=t}(y)f_X(t) \\ \Leftrightarrow \lambda_{Y|X=x}(y) &= -Q(y)^{-1}, \end{aligned}$$

where $\lambda_{Y|X=x}$ denotes the hazard rate function, which does not depend on x. Note that $S_{Y|X=x}(x) = 1$ and

$$S_{Y|X=x}(y) = \frac{S_{Y|X=x}(y)}{S_{Y|X=x}(x)} = \exp\Big(\int_x^y Q(s)^{-1}ds\Big).$$

Moreover, for $t < x < y$

$$\frac{S_{Y|X=x}(y)}{S_{Y|X=t}(y)} = \exp\Big(\int_x^y Q(s)^{-1}ds - \int_t^y Q(s)^{-1}ds\Big) = \frac{g(t)}{g(x)},$$

where $g(x) = \exp(\int_{l_X}^x Q(s)^{-1}ds)$ and $l_X = \inf\{s \geq 0 : f_X(s) > 0\}$ is the lower bound of the support of X (given $X \leq Y$). Differentiation with respect to y leads to

$$f_{Y|X=x}(y) = f_{Y|X=t}(y)\frac{g(t)}{g(x)}$$

and, thus,

$$f_{XY}(x,y) = \frac{f_{XY}(t,y)g(t)}{f_X(t)}\frac{f_X(x)}{g(x)}. \tag{5.22}$$

Now, let $(t_n)_{n\in\mathbb{N}}$ be a strictly decreasing sequence with $f(t_n) > 0$ and $t_n \to l_X$ as $n \to \infty$. Set $t_0 = \infty$. Then we can deduce from Eq. (5.22) that

$$f_{XY}(x,y) = \widetilde{f}_Y(y)\widetilde{f}_X(x),$$

where

$$\widetilde{f}_Y(y) = \sum_{n=1}^{\infty}\frac{f_{XY}(t_n,y)g(t_n)}{f_X(t_n)}\mathbb{1}_{\{y\in(t_n,t_{n-1})\}}, \quad \widetilde{f}_X(x) = \frac{f_X(x)}{g(x)}.$$

5.B Review of Related Quasi-independence Tests

In this section, we review the quasi-independence tests implemented in Section 5.4 of the main text.

WLR refers to the weighted log-rank test discussed in [Emura and Wang, 2010], which is defined as

$$L_W = \int_{x \le y} W(x,y) \left\{ N_{11}(dx,dy) - \frac{N_{1\bullet}(dx,y) N_{\bullet 1}(x,dy)}{R(x,y)} \right\},$$

where

$$N_{11}(dx,dy) = \sum_j \mathbb{1}(X_j = x, T_j = y, \Delta_j = 1),$$

$$N_{\bullet 1}(x,dy) = \sum_j \mathbb{1}(X_j \le x, T_j = y, \Delta_j = 1),$$

$$N_{1\bullet}(dx,y) = \sum_j \mathbb{1}(X_j = x, T_j \ge y),$$

$$R(x,y) = \sum_j \mathbb{1}(X_j \le x, T_j \ge y),$$

and $W : \mathbb{R}_+^2 \to \mathbb{R}$ is the weight function given by $W(x,y) = R(x,y)$. We note that, $R(x,y) = n\widehat{\pi}^c(x,y)$ defined in our notation. It is straightforward to see $\Psi_{c,n}^2 = \frac{1}{n^2} L_W^2$ in the case $\mathfrak{K} = 1$.

WLR_SC refers to the previous log-rank test with weight W given by $W(x,y) = \int_0^x \widehat{S}_{C_R}((y-u)-)^{-1} \widehat{\pi}^c(du,y)$, where $\widehat{S}_{C_R}$ is the Kaplan-Meier estimator based on the data $((C_i - X_i, 1-\Delta_i))_{i=1}^n$. this specific test was proposed to the general assumption $Y_i \perp C_i | X_i$.

M&B refers to the conditional Kendall's tau statistic in discussed in [Martin and Betensky, 2005]. Let

$$B_{ij} = \{\max(X_i, X_j) \le \min(T_i, T_j)\}$$
$$\cap \{(\Delta_i = \Delta_j = 1) \cup (T_j > T_i, \Delta_i = 1, \Delta_j = 0) \cup (T_i > T_j, \Delta_i = 1, \Delta_j = 0)\}.$$

The conditional Kendall's tau statistic is given by

$$\widehat{\tau}_b = \sum_{i<j} \mathbb{1}_{\{B_{ij}\}} \text{sign}((X_i - X_j)(T_i - T_j)).$$

MinP1 and **MinP2** refers to the minimal p-value selection tests which are permutation based methods proposed in [Chiou et al., 2018]. These tests are based on the

underlying principle that, under quasi-independence, the distributions of $Y|X \leq t$ and $Y|X > t$ should not differ, where t denotes some cut-point. Given a collection of possible cut-points t, the authors perform several two-sample log-rank tests for comparing $\{(T_i, \Delta_i) : X_i \leq t\}$ and $\{(T_i, \Delta_i) : X_i > t\}$ (under right-censored data), and set as their test-statistic the minimum log-rank p-value obtained. To guarantee meaningful comparisons, the authors consider cut-points that yield at least E events in each group.

The first test proposed is the following:

MinP1:

1 Set $m = 0$

2 Set $m = m + 1$ and split the data into two groups $\{i : X_i \leq X_m\}$ and $\{i : X_i > X_m\}$.

3 Check the groups are admissible by verifying $E \leq \sum_{i=1}^{n} \Delta_i \mathbb{1}_{\{X_i < X_m\}} \leq n - E$. If the latter holds, perform a two-sample log-rank test for comparing $\{(T_i, \Delta_i) : X_i \leq X_m\}$ and $\{(T_i, \Delta_i) : X_i > X_m\}$, and record the p-value obtained. If the condition is not satisfied, record a p-value equal to 1.

4 If $m < n$ return to Step 2

5 Set as test-statistic $minp_1$ the smallest p-value obtained.

Alternatively, the authors propose a second test, which splits the data according to whether or not, the entry times belong to the interval $(t - \epsilon, t + \epsilon)$, where t, again, denotes a cut-point and $\epsilon > 0$. Similarly to the previous case, we need to ensure that each group contains at least R data points, this can be done by choosing a suitable $\epsilon > 0$.

MinP2:

1 Set $m = 0$

2 Set $m = m + 1$ and split the data into two groups $\{i : X_i \in (X_m - \epsilon_m, X_m + \epsilon_m)\}$ and $\{i : X_i \notin (X_m - \epsilon_m, X_m + \epsilon_m)\}$, where ϵ_m is the smallest $\epsilon > 0$ such that there are at least E data-points in each group. Record the value ϵ_m.

3 If $m < n$ return to Step 2.

4 Set $\epsilon = \max_m \epsilon_m$ and $m = 0$

5 Set $m = m + 1$. Verify $E \leq \sum_{i=1}^{n} \Delta_i \mathbb{1}_{\{T_m - \epsilon < T_i < T_m + \epsilon\}} \leq n - E$ which checks that the partition of the data is admissible (under right-censoring). If the latter holds, perform a two-sample log-rank test for comparing each group and record the p-value. If the partition is not admissible record a p-value equal to 1.

6 If $m < n$ return to Step 5.

7 Set as test-statistic $minp_2$ the smallest p-value obtained.

The rejection regions for these tests are computed by using a permutation approach.

5.C Efficient Implementation of Wild Bootstrap

Similarly to the work of [Chwialkowski et al., 2014], we can implement our wild bootstrap efficiently by considering the identity $\text{tr}(\boldsymbol{AB}) = \sum_{ij}(\boldsymbol{A} \odot \boldsymbol{B})_{ij}$, where $\boldsymbol{A}$ and $\boldsymbol{B}$ denote $n \times n$ matrices, and $\odot$ denotes the element-wise product. By using this identity our test-statistic can be written as

$$\begin{aligned}
\Psi^2_{c,n} &= \frac{1}{n^2}\text{tr}(\boldsymbol{K}\widehat{\boldsymbol{\pi}}^c\tilde{\boldsymbol{L}}\widehat{\boldsymbol{\pi}}^c - 2\boldsymbol{K}\widehat{\boldsymbol{\pi}}^c\tilde{\boldsymbol{L}}\boldsymbol{B}^\top + \boldsymbol{K}\boldsymbol{B}\tilde{\boldsymbol{L}}\boldsymbol{B}^\top) \\
&= \sum_{ij}\left(\boldsymbol{K} \odot (\widehat{\boldsymbol{\pi}}^c\tilde{\boldsymbol{L}}\widehat{\boldsymbol{\pi}}^c - 2\widehat{\boldsymbol{\pi}}^c\tilde{\boldsymbol{L}}\boldsymbol{B}^\top + \boldsymbol{B}\tilde{\boldsymbol{L}}\boldsymbol{B}^\top)\right)_{ij} \\
&= \sum_{ij}\boldsymbol{M}_{ij},
\end{aligned}$$

where $\boldsymbol{M} = \boldsymbol{K} \odot (\widehat{\boldsymbol{\pi}}^c\tilde{\boldsymbol{L}}\widehat{\boldsymbol{\pi}}^c - 2\widehat{\boldsymbol{\pi}}^c\tilde{\boldsymbol{L}}\boldsymbol{B}^\top + \boldsymbol{B}\tilde{\boldsymbol{L}}\boldsymbol{B}^\top)$ is a V-statistic matrix. Then, the wild bootstrap version of the preceding V-statistic is $(\Psi^{\text{WB}}_{c,n})^2 = W^\top \boldsymbol{M} W$ where $W = (W_1, \ldots, W_n) \in \mathbb{R}^n$ are the wild bootstrap weights. In this way, we only need to compute $O(n^2)$ sum once, for each wild bootstrap, instead of computing several (actually 6 times) $O(n^3)$ matrix multiplications and two $O(n^2)$ matrix multiplications for K^W. In the experiments shown in this chapter, independent Rademacher variables are used for wild bootstrap weights.

5.D Additional Discussions on Empirical Results

This section provides additional information and discussions on empirical findings.

Kernel choice

In kernel-based hypothesis testing, test power (i.e., the probability of rejecting H_0 when it is false) can vary for different choices of kernel parameters, such as the

bandwidth in Gaussian kernels [Gretton et al., 2012b]. Previous works [Gretton et al., 2012b; Jitkrittum et al., 2018, 2016a, 2017; Sutherland et al., 2016] have proposed to choose the kernel parameters by maximizing a proxy for the test power. Such objective has also been discussed in Chapter 3 for kernel choice on Manifold testing. In the uncensored setting, the test power is (to a good approximation) increased by maximising the ratio of the test statistic to its standard deviation under the alternative. We conjecture that the same ratio represents a good criterion in the setting of left-truncation and right-censoring, for which we have strong empirical evidence. A formal proof remains a topic for future work.

In the censored case, the test power criterion takes the form $\frac{\Psi_c^2}{\sigma_{H_1}}$, where σ_{H_1} is the standard deviation of Ψ_c^2 under the alternative hypothesis H_1. Thus, to maximise the test power, we choose the kernel parameter θ by

$$\theta^* = \arg\max_{\theta} \frac{\Psi_c^2}{\sigma_{H_1}}.$$

In practice, we use part of the data to compute $\Psi_{c,n}^2/(\widehat{\sigma}_{H_1} + \lambda)$, where $\widehat{\sigma}_{H_1}$ is an empirical estimate of σ_{H_1} and a regularisation parameter $\lambda > 0$ is added for numerical stability. We then perform the test on the remaining data with the selected θ^*. A $20/80$ train-test split is suggested in Jitkrittum et al. [2017] for learning the parameter, which is used in the experiment[2]. We use the regulariser $\lambda = 0.01$.

We next give our empirical estimate for the variance $\widehat{\sigma}_{H_1}^2$. First, $\Psi_{c,n}^2$ can be written as $\Psi_{c,n}^2 = \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n J_n((T_i, \Delta_i, X_i), (T_j, \Delta_j, X_j))$, where J_n is defined by

$$J_n((T_i, \Delta_i, X_i), (T_j, \Delta_j, X_j)) = \Delta_i \Delta_j L(T_i, T_j) g_n(X_i, X_j),$$

where

$$\begin{aligned} g_n(X_i, X_j) =& K(X_i, X_j)\widehat{\boldsymbol{\pi}}^{\boldsymbol{c}}{}_{ii}\widehat{\boldsymbol{\pi}}^{\boldsymbol{c}}{}_{jj} - 2\sum_{l=1}^n K(X_i, X_l)\widehat{\boldsymbol{\pi}}^{\boldsymbol{c}}{}_{ii}\boldsymbol{B}_{l,j} \\ &+ \sum_{l=1}^n \sum_{k=1}^n K(X_k, X_l)\boldsymbol{B}_{k,i}\boldsymbol{B}_{l,j}, \end{aligned}$$

and $\widehat{\boldsymbol{\pi}}^{\boldsymbol{c}}{}_{ii} = \widehat{\pi}^c(X_i, T_i)$ and $B_{k,i} = \mathbb{1}_{\{X_i \le X_k < T_i \le T_k\}}/n$. This "$V$-statistic" form suggests that the variance can be estimated by

[2]We note that, no additional sample points are introduced for training the kernel parameter θ.

n	50	100	150	200	250	300	350	400	450	500
KQIC_IMQ	0.08	0.05	0.03	0.05	0.04	0.05	0.05	0.07	0.07	0.05

Table 5.4: Type-I error for IMQ kernels, with $\alpha = 0.05$, censoring level 25%, 100 trials, and increasing sample size n.

$$\widehat{\sigma}^2_{H_1} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n}\sum_{j=1}^{n} J_n(i,j)\right)^2 - \left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} J_n(i,j)\right)^2,$$

where $J_n(i,j) = J_n((T_i, \Delta_i, X_i), (T_j, \Delta_j, X_j))$.

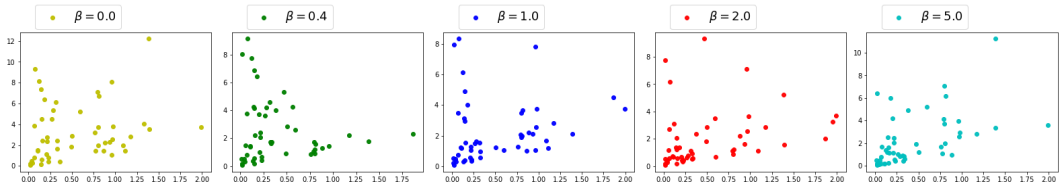
Finally, some remarks on the performance of our kernel selection heuristic in experiments. For simple cases, our kernel selection procedure makes little difference, since a broad range of kernel bandwidths yields good results, and the "median heuristic" (selection of the bandwidth as the pairwise inter-sample distance) is adequate. On the other hand, our procedure results in large power improvements for more complex cases such as periodic dependency at high frequencies, where the median distance between samples does not correspond to the length-scale at which dependence occurs. Similar phenomena have also been observed previously in [Sutherland et al., 2016].

Inverse Multi-Quadratic (IMQ) kernel We further study the performance of the IMQ kernel on our proposed test. The IMQ kernel has the form $k(x,y) = (c^2 + \|x-y\|^2)^b$, for constant $c > 0$ and $b \in (-1, 0)$. As proposed in [Gorham and Mackey, 2017], we choose $b = -\frac{1}{2}$. We select the parameter c by maximizing a heuristic proxy for test power, as discussed above. The controlled Type-I error is shown in Table 5.4, where X and Y are independent samples from $\mathrm{Exp}(1)$. Truncation and right-censoring apply with censoring time independently generated from exponential distribution. We report the test power of KQIC with IMQ kernel in later sections.

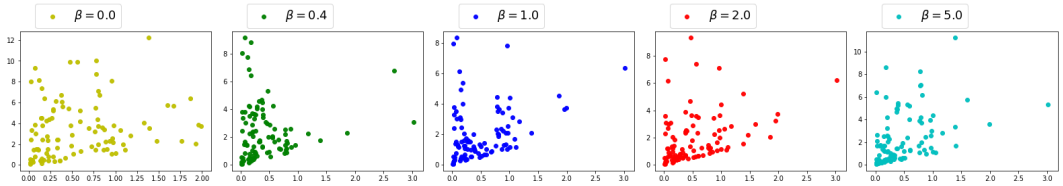
5.D.1 Periodic Dependencies

As briefly mentioned in the main text, the parameter β controls the frequency of sinusoidal dependence. At a given sample size, dependence becomes harder to detect as the frequency β increases, both for our test and for competing methods. We illustrate the datasets visually in Figure 5.7. For a fixed sample size, the test power decreases as frequency increases, which is observed in our results in Figure 5.4. For high frequency cases, larger sample size is required to correctly reject the

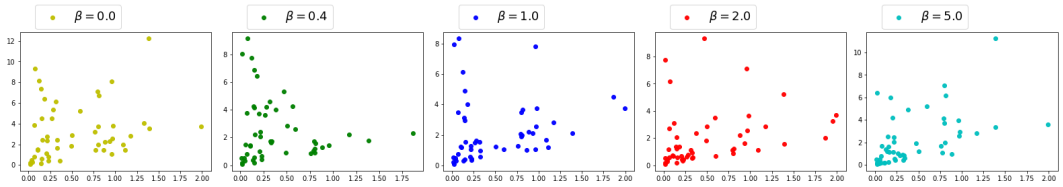
null as shown in Figure 5.5.



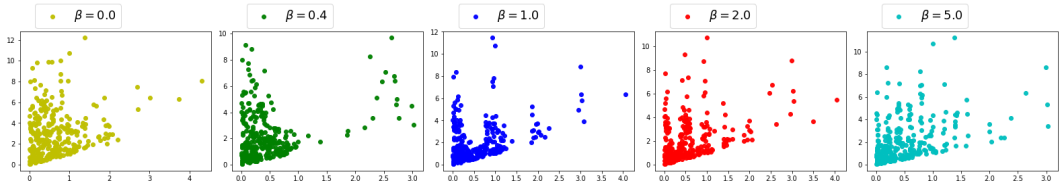
(a) Sample size: n = 50



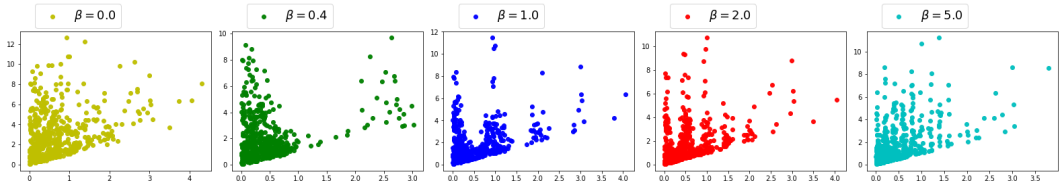
(b) Sample size: n = 100



(c) Sample size: n = 200



(d) Sample size: n = 300



(e) Sample size: n = 500

Figure 5.7: Samples from periodic dependency model w.r.t. frequency coefficient β.

Type-I error is reported in Table 5.5, and is close to the desired level (subject to finite sample effects).

5.D.2 Dependent Censoring

In this section we show that our test achieves correct Type-I error under the null hypothesis even when considering dependent censoring times C. As stated in Assumption 5.1, we only require $Y \perp C|X$, which is a standard assumption, as also

n	100	300	500	700	900	1100	1300	1500	1700	1900
KQIC_Gauss	0.045	0.060	0.055	0.040	0.045	0.045	0.040	0.030	0.045	0.050
KQIC_IMQ	0.050	0.055	0.045	0.030	0.020	0.040	0.025	0.020	0.015	0.020
WLR	0.030	0.045	0.050	0.025	0.045	0.015	0.015	0.030	0.025	0.040
WLR_SC	0.035	0.060	0.030	0.025	0.060	0.070	0.045	0.055	0.050	0.060

Table 5.5: Type-I error with increasing sample size; $\alpha = 0.05$, censoring level 25%, 200 trials.

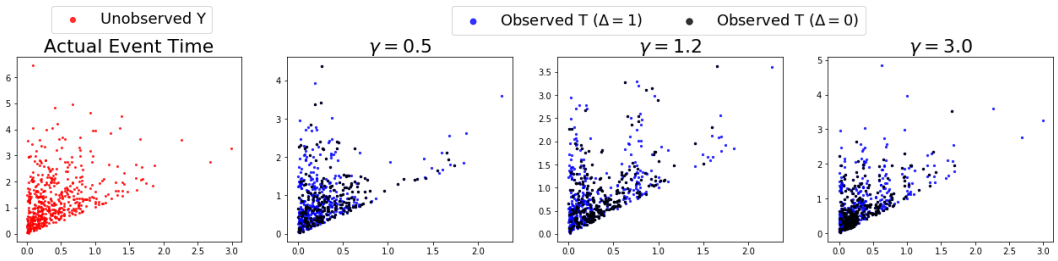


Figure 5.8: Samples generated from H_0 with periodic dependent censoring distributions.

considered in [Emura and Wang, 2010].

We generate the data as follows: Sample $X_i \sim \text{Exp}(1)$, then generate $Y_i \sim \text{Exp}(1)$ (independent of X_i) and $C_i|X_i \sim \text{Exp}(e^{\cos(2\pi\gamma X_i)})$. Generate the observed data point (T_i, Δ_i, X_i), where $T_i = \min\{Y_i, C_i\}$ and $\Delta_i = \mathbb{1}_{\{T_i=Y_i\}}$ and keep it as a valid sample only if $T_i \geq X_i$. Notice that in this case both left truncation and right-censoring are present in the data. Also, notice that the null hypothesis holds since the survival times Y_i are quasi-independent of the entry times X_i. In Figure 5.8, we show the unobserved pairs (X, Y) and the observed pairs (X, T) where the censoring variable is generated using different censoring frequencies γ. From the plot, we see that the entry times X and survival times Y look quasi-independent, but, due to the periodic dependency of the censoring distribution, the observed data (X, T) show a periodic trend, which looks similar to the observations in Figure 5.7. However, since this dependency is due to the censoring times C instead of the survival times Y, our tests are able to recover H_0 and achieve correct test level, as shown in Table 5.6. The tests proposed in [Emura and Wang, 2010] are also valid under Assumption 5.1, thus we include the results for WLR and WLR_SC as well. From Table 5.6, we observe that KQIC with both Gaussian and IMQ kernels, as well as WLR achieve the correct test level; however, WLR_SC has slightly higher type-I errors when sample size is small and achieves correct test-level when sample size becomes large (recall that WLR_SC uses a data dependent weight, thus convergence in this case might be slower).

Table 5.6: Type-I error for periodic dependent censoring distributions, with $\alpha = 0.05$ and 100 trials.

n	100	200	300	400	500	600	700	800	900	1000
KQIC_Gauss	0.07	0.06	0.03	0.03	0.06	0.05	0.04	0.04	0.03	0.07
KQIC_IMQ	0.07	0.06	0.04	0.01	0.03	0.04	0.05	0.05	0.06	0.07
WLR	0.07	0.05	0.03	0.01	0.03	0.04	0.05	0.04	0.03	0.07
WLR_SC	0.10	0.08	0.09	0.13	0.13	0.04	0.09	0.05	0.04	0.06

Table 5.7: Censoring frequency $\gamma = 0.5$. Censoring level 30%

n	100	200	300	400	500	600	700	800	900	1000
KQIC_Gauss	0.03	0.02	0.01	0.04	0.05	0.06	0.05	0.04	0.06	0.04
KQIC_IMQ	0.02	0.03	0.03	0.03	0.04	0.05	0.05	0.04	0.04	0.04
WLR	0.02	0.02	0.03	0.05	0.04	0.04	0.04	0.04	0.05	0.05
WLR_SC	0.06	0.12	0.17	0.15	0.10	0.11	0.06	0.04	0.05	0.05

Table 5.8: Censoring frequency $\gamma = 1.2$. Censoring level 35%

n	100	200	300	400	500	600	700	800	900	1000
KQIC_Gauss	0.06	0.04	0.06	0.02	0.02	0.04	0.03	0.03	0.05	0.04
KQIC_IMQ	0.05	0.04	0.05	0.01	0.02	0.04	0.04	0.03	0.03	0.04
WLR	0.04	0.02	0.04	0.02	0.02	0.04	0.04	0.03	0.05	0.03
WLR_SC	0.09	0.10	0.13	0.15	0.10	0.08	0.05	0.03	0.03	0.04

Table 5.9: Censoring frequency $\gamma = 3.0$. Censoring level 40%

5.D.3 Censoring Level

We report the Type-I error for different censoring percentages, see Table 5.10. With reasonable censoring level (e.g. $< 90\%$), the Type-I errors are well controlled. WLR_SC has higher Type-I with small sample sizes, which is similarly observed in Table 5.6. However, the Type-I error is less controlled at extremely high censoring percentages, due to the lack for useful information obtained. In practise, we may need to be careful dealing with extremely high censoring when applying the quasi-independence tests.

% censored	20	35	50	70	85	92	95
$n = 200$							
KQIC_Gauss	0.040	0.025	0.015	0.045	0.035	0.085	0.115
KQIC_IMQ	0.040	0.060	0.050	0.055	0.070	0.100	0.185
WLR	0.055	0.035	0.040	0.050	0.030	0.075	0.120
WLR_SC	0.045	0.105	0.075	0.120	0.060	0.035	0.075
$n = 300$							
KQIC_Gauss	0.055	0.040	0.055	0.045	0.060	0.090	0.065
KQIC_IMQ	0.045	0.050	0.070	0.050	0.050	0.105	0.115
WLR	0.030	0.055	0.055	0.040	0.050	0.075	0.065
WLR_SC	0.080	0.120	0.140	0.095	0.125	0.095	0.025
$n = 500$							
KQIC_Gauss	0.040	0.050	0.035	0.030	0.030	0.050	0.090
KQIC_IMQ	0.065	0.030	0.050	0.040	0.080	0.100	0.050
WLR	0.035	0.035	0.050	0.035	0.040	0.060	0.075
WLR_SC	0.060	0.035	0.055	0.075	0.065	0.035	0.015
$n = 800$							
KQIC_Gauss	0.045	0.030	0.030	0.065	0.030	0.065	0.080
KQIC_IMQ	0.065	0.050	0.050	0.060	0.060	0.090	0.140
WLR	0.015	0.010	0.025	0.055	0.065	0.085	0.100
WLR_SC	0.095	0.040	0.065	0.080	0.075	0.045	0.025

Table 5.10: Type-I error for different censoring level, with $\alpha = 0.05$ and 200 trials,

Chapter 6

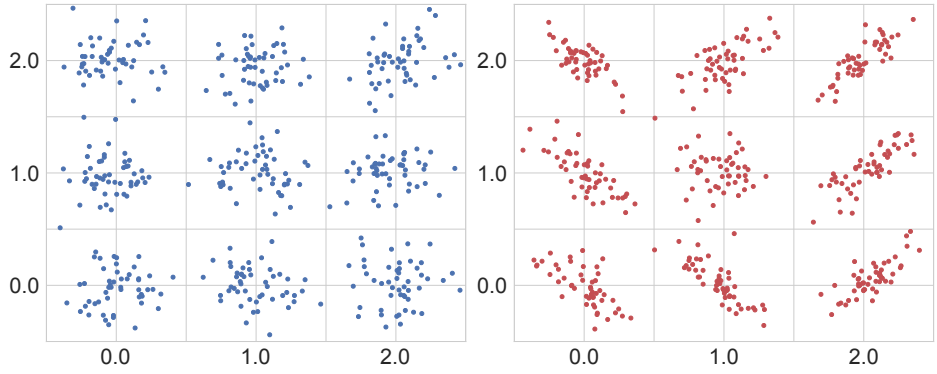
Deep Kernels for Hypothesis Testing

Summary We investigate the MMD-based two-sample testings with kernels parameterised by deep neural networks. While the deep neural networks are trained to maximise test power, these tests are able to adapt to variations in distribution smoothness as well as shape over space, and are especially suited to high dimensions and complex data. By contrast, the simpler kernels used in prior kernel-based hypothesis testings are spatially homogeneous and adaptive only in length-scale. We provide theoretical analysis for the proposed kernel learning schemes and experimentally establish the superior performance of our deep kernels in hypothesis testing on both benchmark and real-world data.

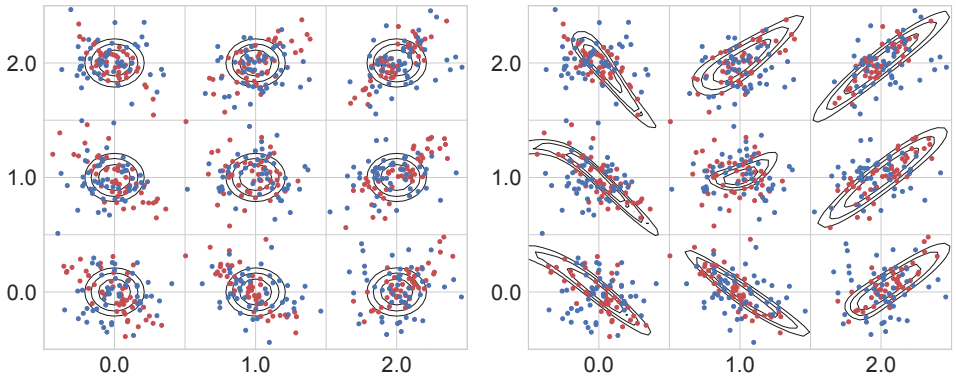
6.1 Introduction

As introduced in Chapter 2, the kernel-based non-parametric tests for two-sample problems [Gretton et al., 2012a] has been shown to have good theoretical properties as well as state-of-the-art empirical performances. Problems that we encounter in practice, however, often involve distributions with complex structure, where simple kernels will often map distinct distributions to nearby mean embeddings, making it hard to distinguish between distributions. Figure.6.1(a) shows an example of a multi-modal dataset, where the overall modes align but the sub-mode structure varies differently at each mode. A translation-invariant Gaussian kernel, Eq.(2.4), only "looks at" the data uniformly within each mode as demonstrated in Figure.6.1(b), requiring many samples to correctly distinguish the two distributions. The distributions can be distinguished more effectively if we understand the structure of each mode, as with the more complex kernel illustrated in Figure.6.1(c).

To model these complex functions, we adopt a *deep kernel* approach [Wilson et al., 2016; Sutherland et al., 2016; Jean et al., 2018; Li et al., 2017; Wenliang et al., 2018], building a kernel with a deep network. In this chapter, we use the kernel of



(a) Samples drawn from P (left) and Q (right).



(b) Contour of the Gaussian kernel (c) Deep kernel contour

Figure 6.1: Illustration of the Blob example: (a) P and Q are mixtures of nine Gaussian components with the same modes, but each component of P is an isotropic Gaussian whereas the covariance of Q differs in each component. (b) and (c) show the contours of a kernel, $k(x, \mu_i)$ for each of the nine modes μ_i; contour values are 0.7, 0.8 and 0.9. A Gaussian kernel (b) treats points isotropically throughout the space, based only on Euclidean distance $\|x - y\|$. A deep kernel (c) learned by our methods behaves differently in different parts of the space, adapting to the local structure of the data distributions and hence allowing better identification of differences between P and Q.

the form,

$$k_\omega(x, y) = [(1 - \epsilon)\kappa(\phi_\omega(x), \phi_\omega(y)) + \epsilon]q(x, y), \tag{6.1}$$

where the deep neural network ϕ_ω extracts features of samples, and κ is a simple kernel (e.g., a Gaussian) on those features, while q is a simple characteristic kernel (e.g. Gaussian) on the input space. With an appropriate choice of ϕ_ω, this allows for extremely flexible kernels which can learn complex behavior very different in different parts of space. This choice is discussed further in Section 6.4.

These complex kernels, though, cannot feasibly be specified by hand or simple

heuristics, as is typical practice in kernel methods. We select the parameters ω by maximising the ratio of the MMD to its variance, which maximises test power at large sample sizes. This procedure was proposed by Sutherland et al. [2016] as introduced in Chapter 3, but we establish for the first time that it gives consistent selection of the best kernel in the class, instead of simply choosing length-scales of a Gaussian kernel [Sutherland et al., 2016] or taking a weighted sum Gaussian kernels of different length-scales [Gretton et al., 2012b]. Previously, there were no guarantees this procedure would yield a kernel which generalised at all from the training set to a test set.

Another way to compare distributions is to train a classifier between them, and evaluate its accuracy [Lopez-Paz and Oquab, 2016]. We show, perhaps surprisingly, that our framework encompasses this approach (details in Section 6.3); but deep kernels allow for more general model classes which can use the data more efficiently.

We also train representations directly to maximise test power, rather than a cross-entropy surrogate. We test our method on several simulated and real-world datasets, including complex synthetic distributions, high-energy physics data, and challenging image problems. We find convincingly that the learned deep kernels outperform simple shallow methods, and learning by maximising test power outperforms learning through a cross-entropy surrogate loss [Lopez-Paz and Oquab, 2016].

6.2 Testing with MMD

Recall the MMD-based non-parametric tests for two-sample problem shown in Section 2.1. Using the empirical U-statistics estimator MMD_u^2 in Eq.(2.12) as the test statistics, it can be shown that under the null hypothesis H_0, $n \cdot \mathrm{MMD}_u^2$, asymptotically converges to a weighted chi-square distribution depending on distribution P and kernel k; and $\sqrt{n} \cdot \mathrm{MMD}_u^2$ is asymptotically normally distributed under the alternative hypothesis $H_1 : P \neq Q$ [Gretton et al., 2007, Theorem 8].

Proposition 6.1 (Asymptotics of MMD_u^2)**.** *Under the null hypothesis, $H_0 : P = Q$, we have if $Z_i \sim \mathcal{N}(0, 2)$,*

$$n \cdot \mathrm{MMD}_u^2 \xrightarrow{d} \sum_i w_i (Z_i^2 - 2);$$

where $\xrightarrow{d}$ denotes convergence in distribution and w_i are the eigenvalues of the P-

covariance operator of the centered kernel,

$$\tilde{k}(x,y) := \langle k(\cdot,x) - \mu_P, k(\cdot,y) - \mu_P \rangle$$
$$\int \tilde{k}(x,y)\psi_i(x)\mathrm{d}P(x) = w_i\psi_i(y).$$

Under the alternative hypothesis, $H_1 : P \neq Q$, a standard central limit theorem holds [Serfling, 2009, Section 5.5.1]:

$$\sqrt{n}\cdot(\mathrm{MMD}_u^2 - \mathrm{MMD}^2) \xrightarrow{\mathrm{d}} \mathcal{N}(0,\sigma_{H_1}^2)$$
$$\sigma_{H_1}^2 := 4\left(\mathbb{E}[\tilde{G}_{12}\tilde{G}_{13}] - \mathbb{E}[\tilde{G}_{12}]^2\right)$$

where $\tilde{G}_{ij} := k(x_i,x_j) + k(y_i,y_j) - k(x_i,y_j) - k(y_i,x_j)$[1], $x_i, x_j \sim P$, $y_i, y_j \sim Q$.

Although it is possible to construct a test based on directly estimating the null distribution [Gretton et al., 2009a], it is both simpler and, if implemented carefully, faster [Sutherland et al., 2016] to instead use a permutation test. Noting that the samples from P and Q are interchangeable under the null $H_0 : P = Q$ [Dwass, 1957; Fernández et al., 2008], we can therefore estimate the null distribution of our test statistic by repeatedly re-computing it with the samples randomly re-assigned to S_P or S_Q. With the simulated null distribution via the permutation procedure [Gretton et al., 2008], we are able to compute the p-value of our test statistics and comparing with the predefined significance level concludes the test outcome.

Test Power

The main measure of efficacy of a null hypothesis test is its *power*: the probability that, for a particular $P \neq Q$ and n, we correctly reject H_0. Proposition 6.1 implies, where Φ is the standard normal CDF, that the probability of the tests statistic exceeding the threshold r is

$$\mathbb{P}_{H_1}\left(n\cdot\mathrm{MMD}_u^2 > r\right) \xrightarrow{\mathrm{d}} \Phi\left(\frac{\sqrt{n}\,\mathrm{MMD}^2}{\sigma_{H_1}} - \frac{r}{\sqrt{n}\,\sigma_{H_1}}\right);$$

we can find the approximate test power by using the rejection threshold, found via (e.g.) permutation testing. We also know via Proposition 6.1 that this r will converge to a constant, and MMD, σ_{H_1} are also constants. For reasonably large n, the power is dominated by the first term, and the kernel yielding the most powerful

[1] With observed samples, the empirical estimate $\mathrm{MMD}_u^2(S_P, S_Q; k) = \frac{1}{n(n-1)}\sum_{i\neq j}\tilde{G}_{ij}$

test will approximately maximise [Sutherland et al., 2016]

$$J(P, Q; k) := \frac{\mathrm{MMD}^2(P, Q; k)}{\sigma_{H_1}(P, Q; k)}. \tag{6.2}$$

Selecting Kernel from Approximate Test Power

The criterion $J(P, Q; k)$ depends on the particular P and Q at hand, and thus we typically will neither be able to choose a kernel *a priori*, nor exactly evaluate J given samples. We can, however, estimate it with

$$\widehat{J}_\lambda(S_P, S_Q; k) := \frac{\mathrm{MMD}_u^2(S_P, S_Q; k)}{\hat{\sigma}_{H_1,\lambda}(S_P, S_Q; k)}, \tag{6.3}$$

where $\hat{\sigma}^2_{H_1,\lambda}$ is a regularised estimator of $\sigma^2_{H_1}$ given by[2]

$$\hat{\sigma}^2_{H_1,\lambda} = \frac{4}{n^3}\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\tilde{G}_{ij}\right)^2 - \frac{4}{n^4}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{G}_{ij}\right)^2 + \lambda. \tag{6.4}$$

The λ-regularisation [Jitkrittum et al., 2017] is to avoid the vanishing denominator. In the analysis, $\lambda = n^{-\frac{1}{3}}$ is chosen to balance the convergence rate in Theorem 6.3. It is enough to set to be a small number in practise, e.g. $\lambda = 10^{-8}$ for the presented experiments to achieve good test power.

Given S_P and S_Q, we could construct a test by choosing k to maximise $\widehat{J}_\lambda(S_P, S_Q; k)$, then using a test statistic based on $\mathrm{MMD}(S_P, S_Q; k)$. This sample re-use, however, violates the conditions of Proposition 6.1, and permutation testing would require repeatedly re-training k with permuted labels. Thus we split the data, into training set S_P^{tr}, S_Q^{tr} and testing set S_P^{te}, S_Q^{te}; and get $k^{tr} \approx \arg\max_k \widehat{J}_\lambda(S_P^{tr}, S_Q^{tr}; k)$; then we compute the test statistic and permutation threshold on S_P^{te}, S_Q^{te} using the learned k^{tr}. This procedure was proposed for MMD_u^2 by Sutherland et al. [2016], but the same technique works for a variety of tests [Gretton et al., 2012a; Jitkrittum et al., 2016a; Lopez-Paz and Oquab, 2016]. Our learning scheme adopts this framework and further studies it in Section 6.4.

Relationship to Other Approaches

One common scheme is to pick a deep network parametrised kernel k_ω based on some proxy task, such as a related classification problem [Kirchler et al., 2020;

[2]This estimator, as a V-statistic, is biased even when $\lambda = 0$, although this bias is only $O(1/N)$ (see Theorem 6.3). Although Sutherland et al. [2016]; Sutherland [2019] gave a quadratic-time estimator unbiased for $\sigma^2_{H_1}$, it is much more complicated to implement and analyse, likely has higher variance, and (being unbiased) can be negative, especially when the kernel is poor.

Lopez-Paz and Oquab, 2016] or the KID score [Bińkowski et al., 2018]. Although this approach can work quite well, it depends entirely on the features from the proxy task applying well to distinguish the differences between P and Q, which can be hard to know in general.

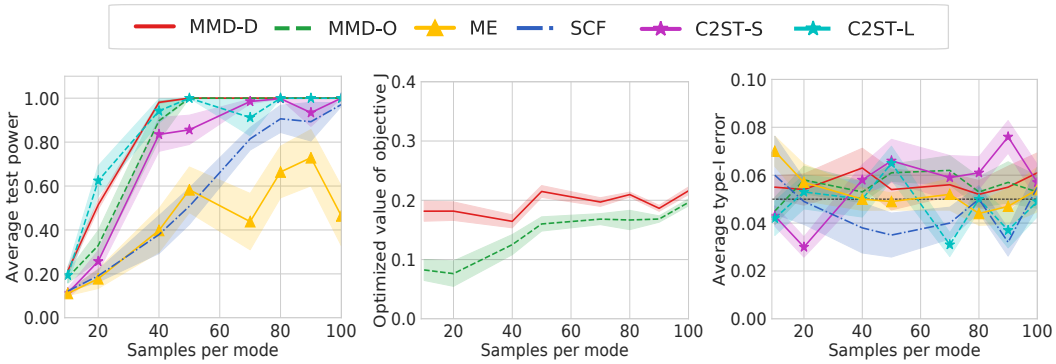
An alternative is to maximise simply MMD_u [Fukumizu et al., 2009] (proposed but not evaluated by Kirchler et al. [2020]). Ignoring σ_{H_1} means that, for instance, this approach would choose to simply scale $k \to \infty$, even though this does not change the test at all. Even when this is not possible, Sutherland et al. [2016] found this approach notably worse than maximising the objective in Eq. (6.3); we will further confirm this in our experiments.

In generative modelling tasks, MMD-GANs [Li et al., 2017; Bińkowski et al., 2018] also simply maximise MMD_u to identify the differences between their learned model Q_θ and target P. If Q_θ is quite far from P, however, an MMD-GAN requires a "weak" kernel to identify a path for improving Q_θ [Arbel et al., 2018], while the ideal kernel is one which perfectly distinguishes P and Q_θ and would likely give no signal for improvement. Our proposed algorithm, theoretical guarantees, and empirical evaluations thus all differ significantly from those for MMD-GANs.

6.2.1 Limits of Simple Kernels

We can use the criterion $\widehat{J}_\lambda$ of Eq. (6.3) even to select parameters among a simple family, such as the length-scale of a Gaussian kernel. Doing so on the *Blob* problem of Figure 6.1 illustrates the limitations of using MMD with these kernels. In Figure 6.2 (b), we show how the maximal value of $\widehat{J}$ changes as we see more samples from P and Q, for both a family of Gaussian kernels (green dashed line) and a family of deep kernels in Eq. (6.1) (red line). The optimal $\widehat{J}$ is always higher for the deep kernels; as expected, the empirical test power, shown in Figure 6.2(a), is also higher for deep kernels.

Most simple kernels used for MMD tests, whether the Gaussian we used here or Laplace, Inverse MultiQuadric (IMQ), even Automatic Relevance Determination (ARD) kernels, are all translation invariant: $k(x, y) = k(x - t, y - t)$ for any $t \in \mathbb{R}^d$. All kernels used by Sutherland et al. [2016], for instance, were also of this type. Hence the kernel behaves the same way across space, as in Figure 6.1(b). This means that for distributions whose behavior varies through space, whether because of changes in principal directions (as in Figure 6.1) where the shape becomes different, or because of some regions being much denser than others where a smaller length-scale required [e.g. Wenliang et al., 2018, Figures 1 and 2], any single global



(a) Average test power (b) $\widehat{J}_\lambda$ value in Eq.(6.3) (c) Type-I error

Figure 6.2: Results on *Blob-S* and *Blob-D* given $\alpha = 0.05$; see Section 6.6 for details. n_b is the number of samples at each mode, so $n_b = 100$ means drawing 900 samples from each of P and Q. We report, when increasing n_b, (a) average test power, (b) the value of $\widehat{J}_\lambda$, and (c) average Type-I error. (a) and (b) are on *Blob-D*, and (c) is on *Blob-S*. Shaded regions show standard errors for the mean, and the black dashed line in (c) shows α.

choice is suboptimal. Kernels which are not translation invariant, such as the deep kernels Eq. (6.1), as shown in Figure.6.1(c), are able to adapt to the different shapes necessary in different areas.

6.3 Relationship to Classifier-Based Tests

Another popular method for conducting two-sample tests is to train a classifier between two training sets S_P^{tr} and S_Q^{tr}, then assess its performance on test sets S_P^{te} and S_Q^{te}. If $P = Q$, such classification should be impossible and the classification performance will be at chance. The most common performance metric is the classification accuracy [Lopez-Paz and Oquab, 2016]. This scheme is fairly common among practitioners, and Kim et al. [2016] showed it to be optimal in rate, but suboptimal in constant, in a limited setting[3]. We will call this approach a Classifier Two-Sample Test based on Sign (C2ST-S).

Let $f : \mathcal{X} \to \mathbb{R}$ output the classification scores, and the classification accuracy

$$\mathrm{acc}(P, Q; f) := \frac{1}{2}\mathbb{P}(f(X) > 0) + \frac{1}{2}\mathbb{P}(f(Y) \leq 0);$$

[3]linear discriminant analysis between high-dimensional elliptical distributions, e.g. Gaussian distributions, with identical covariances

the empirical accuracy from samples S_P and S_Q is used as the C2ST-S test statistic,

$$\widehat{\mathrm{acc}}(S_P, S_Q; f) = \frac{1}{2n} \sum_{X_i \in S_P} \mathbb{1}(f(X_i) > 0) + \frac{1}{2n} \sum_{Y_i \in S_Q} \mathbb{1}(f(Y_i) \leq 0).$$

$\widehat{\mathrm{acc}}$ is unbiased for acc and has a simple asymptotically normal distribution under the null. Although it is perhaps not immediately obvious this is the case, C2ST-S is almost a special case of the MMD. Choose kernel of the form

$$k_f^{(S)}(x, y) = \frac{1}{4}\, \mathbb{1}(f(x) > 0)\, \mathbb{1}(f(y) > 0). \tag{6.5}$$

Theorem 6.1. *A C2ST-S test with f is equivalent to an* MMD *test with $k_f^{(S)}$:*

$$\mathrm{MMD}(P, Q; k_f^{(S)}) = |\mathrm{acc}(P, Q; f) - \frac{1}{2}|$$
$$\mathrm{MMD}_b(S_P, S_Q; k_f^{(S)}) = |\widehat{\mathrm{acc}}(S_P, S_Q; f) - \frac{1}{2}|.$$

where $\mathrm{MMD}_b(S_P, S_Q; k_f^{(S)})^2 = \frac{1}{n^2} \sum_{ij} \tilde{G}_{ij}$ is the biased estimator of MMD^2.

Proof. The mean embedding μ_P under $k_f^{(S)}$ is simply

$$\frac{1}{2}\mathbb{E}\mathbb{1}(f(X) > 0) = \frac{1}{2}\mathbb{P}(f(X) > 0),$$

so the MMD is

$$\frac{1}{2}\Big|\mathbb{P}(f(X) > 0) - \mathbb{P}(f(Y) > 0)\Big| = \Big|\,\mathrm{acc}(P, Q; f) - \frac{1}{2}\Big|.$$

The empirical version follows by taking the average over samples instead of the expectation. □

In the learning phase, the C2ST-S, however, selects f to by maximising cross-entropy (approximately maximising $\widehat{\mathrm{acc}}$), while we maximise $\widehat{J}_\lambda$ in Eq. (6.3). Despite $k_f^{(S)}$ is not differentiable, maximising Eq. (6.2) would exactly maximise acc and hence maximise test power [Lopez-Paz and Oquab, 2016, Theorem 1].

Accessing f only through its sign allows for a simple null distribution, but it ignores f's measure of confidence: a highly confident output extremely far from the decision boundary is treated the same as a very uncertain one lying in an area of high overlap between P and Q, dramatically increasing the variance of the statistic.

A scheme we call C2ST-L instead tests difference in means of f on P and Q [Cheng and Cloninger, 2019]. Choose kernel of the form

$$k_f^{(L)}(x, y) = f(x)f(y). \tag{6.6}$$

Theorem 6.2. *A C2ST-L is equivalent to an* MMD *test with* $k_f^{(L)}$*:*

$$\mathrm{MMD}(P, Q; k_f^{(L)}) = |\mathbb{E}f(X) - \mathbb{E}f(Y)|$$
$$\mathrm{MMD}_b(S_P, S_Q; k_f^{(L)}) = |\frac{1}{n}\sum_{X_i \in S_P} f(X_i) - \frac{1}{n}\sum_{Y_i \in S_Q} f(Y_i)|.$$

Proof. The result follows as the kernel's feature map is $k_f^{(L)}(x, \cdot) = f(x)$. □

Now maximising accuracy (or a cross-entropy proxy) no longer directly maximises power. This kernel is differentiable, so we can directly compare the merits of maximising Eq. (6.3) to maximising cross-entropy. In Section 6.6.2 that our more direct approach maximising approximate test power is empirically superior.

Compared to using $k_f^{(L)}$, however, Section 6.6.2 shows that the learned MMD tests also obtain better performance using kernels like Eq. (6.1). This is analogous to a similar phenomenon observed in other problems by Bińkowski et al. [2018] and Wenliang et al. [2018]: C2ST methods learn a full discriminator function on the training set, and then apply only that function to the test set. Learning a deep kernel like Eq. (6.1) corresponds to learning only a powerful *representation* on the training set, and then *still learning* f itself from the test set – in a closed form that makes permutation testing simple.

6.4 Learning Deep Kernels

Choice of kernel architecture Most previous work on deep kernels has used a kernel κ directly on the output of a featurisation network ϕ_ω, $k_\omega(x, y) = \kappa(\phi_\omega(x), \phi_\omega(y))$. This is certainly also an option for us. Any such k_ω, however, is characteristic if and only if ϕ_ω is injective. If we select our kernel well, this is not really a concern.[4] Even so, it would be reassuring to know that, even if the optimisation goes awry, the resulting test will still be at least consistent. More importantly,

[4] A characteristic kernel on top of even $\phi_\omega(x) = \omega^\mathsf{T} x$ with a *random* ω will be almost surely consistent [Heller and Heller, 2016], and in general the existence of even one good ϕ_ω for a particular P, Q pair is enough that a perfect optimiser would be able to distinguish the distributions [Arbel et al., 2018, Proposition 1].

Algorithm 3 Testing with a learned deep kernel

Input: S_P, S_Q, various hyperparameters used below;
$\omega \leftarrow \omega_0$; $\lambda \leftarrow 10^{-8}$;
Split the data as $S_P = S_P^{tr} \cup S_P^{te}$ and $S_Q = S_Q^{tr} \cup S_Q^{te}$;
 # Phase 1: train the kernel parameters ω on S_P^{tr} and S_Q^{tr}
for $T = 1, 2, \ldots, T_{max}$ **do**
  $X \leftarrow$ minibatch from S_P^{tr}; $Y \leftarrow$ minibatch from S_Q^{te};
  $k_\omega \leftarrow$ kernel function with parameters ω; *# as in Eq. (6.1)*
  $M(\omega) \leftarrow \mathrm{MMD}_u^2(X, Y; k_\omega)$; *# using Eq. (2.12)*
  $V_\lambda(\omega) \leftarrow \hat{\sigma}^2_{H_1,\lambda}(X, Y; k_\omega)$; *# using Eq. (6.4)*
  $\widehat{J}_\lambda(\omega) \leftarrow M(\omega)/\sqrt{V_\lambda(\omega)}$; *# as in Eq. (6.3)*
  $\omega \leftarrow \omega + \eta \nabla_{\text{Adam}} \widehat{J}_\lambda(\omega)$; *# updates to improve $\widehat{J}_\lambda(\omega)$ with Adam optimizer*
end for
 # Phase 2: permutation test with k_ω on S_P^{te} and S_Q^{te}
$est \leftarrow \mathrm{MMD}_u^2(S_P^{te}, S_Q^{te}; k_\omega)$
for $i = 1, 2, \ldots, n_{perm}$ **do**
  Shuffle $S_P^{te} \cup S_Q^{te}$ into X and Y
  $perm_i \leftarrow \mathrm{MMD}_u^2(X, Y; k_\omega)$
end for
Output: k_ω, est, p-value $:= \frac{1}{n_{perm}} \sum_{i=1}^{n_{perm}} \mathbb{1}_{\{perm_i \geq est\}}$

it can be helpful in optimisation to add a "safeguard" preventing the learned kernel from considering extremely far-away inputs as too similar. We can achieve these goals with the form in Eq. (6.1), repeated here:

$$k_\omega(x, y) = [(1 - \epsilon)\kappa(\phi_\omega(x), \phi_\omega(y)) + \epsilon]\, q(x, y).$$

Here ϕ_ω is a deep network with parameters ω that extracts features, and κ is a base kernel on those features which we use a Gaussian kernel with length-scale σ_ϕ, i.e. $\kappa(a, b) = \exp\left(-\frac{1}{2\sigma_\phi^2}\|a - b\|^2\right)$. We choose $0 < \epsilon < 1$ and q as a Gaussian kernel on input data with length-scale σ_q.

Proposition 6.2. *Let k_ω be of the form Eq. (6.1) with $\epsilon > 0$ and q being characteristic. Then k_ω is characteristic.*

Learning procedure The kernel optimisation and testing procedure is summarised in Algorithm 3. For larger datasets, or when $n \neq m$, we use minibatches in the training procedure; for smaller datasets, we use full batches. We use the Adam optimiser [Kingma and Ba, 2014] to update all the parameters. Note that the parameters ϵ, σ_ϕ, and σ_q are included in ω.

Time complexity Let C_1 denote the cost of computing an embedding $\phi_\omega(x)$, and C_2 the cost of computing Eq. (6.1) given $\phi_\omega(x)$, $\phi_\omega(y)$. Then each iteration of training in 3 costs $\mathcal{O}\left(mC_1 + m^2C_2\right)$, where m is the minibatch size; for the moder-

ate m that fit in a GPU-sized minibatch anyway, the mC_1 term typically dominates, matching the complexity of a C2ST. Testing takes time $\mathcal{O}\left(nC_1 + n^2C_2 + n^2n_{perm}\right)$, compared to $\mathcal{O}\left(nC_1 + nn_{perm}\right)$ for permutation-based C2STs. In either case, the quadratic factors could be reduced with the block estimator approach [Zaremba et al., 2013] if necessary , at the cost of test power. In our experiments in Section 6.6, the overall runtime of our methods was scarcely different from that of C2STs.

6.5 Theoretical Analysis

We now show that optimising the regularised test power criterion based on a finite number of samples works: as n increases, our estimates converge uniformly over a ball in parameter space, and therefore if there is a unique best kernel, convergence is guaranteed. While Sutherland et al. [2016] gave no such guarantees, this result allows us to trust that, at least for reasonably large n, if our optimisation process succeeds, we will find a kernel that generalises nearly optimally rather than just overfitting to S^{tr}. We first state a generic result, then show some choices of kernels, particularly deep kernels Eq. (6.1), satisfy the conditions.

Theorem 6.3. *Let ω parametrise uniformly bounded kernel functions k_ω in a Banach space of dimension D, with $|k_\omega(x,y) - k_{\omega'}(x,y)| \leq L_k\|\omega - \omega'\|$. Let $\bar{\Omega}_s$ be a set of ω for which $\sigma^2_{H_1}(P,Q;k_\omega) \geq s^2 > 0$ and $\|\omega\| \leq R_\Omega$. Take $\lambda = n^{-1/3}$. Then, with probability at least $1-\delta$,*

$$\sup_{\omega\in\bar{\Omega}_s} |\widehat{J}_\lambda(S_P, S_Q; k_\omega) - J(P,Q;k_\omega)| \leq \left(\frac{C}{s^2 n^{1/3}}\left[\frac{1}{s} + \sqrt{D\log(R_\Omega n) + \log\frac{1}{\delta}} + L_k\right]\right),$$

where C is a constant. If there is a unique best kernel ω^*, the maximiser of $\widehat{J}_\lambda$ converges in probability to ω^* as $n \to \infty$.

A version with explicit constants and more details is given in Section 6.A (as Theorem 6.4 and Corollary 6.1); the proof is based on uniform convergence of the MMD and variance estimators using an ϵ-net argument. The following results are shown in Section 6.A.4. We first show a result on simple Gaussian bandwidth selection.

Proposition 6.3. *Suppose each $x \in \mathcal{X}$ has $\|x\| \leq R_X$, and we choose the bandwidth of a Gaussian kernel among a set whose minimum is at least $1/R_\Omega$. Then the conditions of Theorem 6.3 are met with $D = 1$ and $L_k = 2R_X/\sqrt{e}$.*

We then establish our results for fully-connected deep kernels; it also applies to convolutional networks with a slightly different R_Ω (Remark 6.2). The constants in L_k are given in Proposition 6.8.

Proposition 6.4. *Take k_ω as in 6.4, with ϕ_ω a fully-connected network with depth Λ and D total parameters, whose activations are 1-Lipschitz with $\sigma(0) = 0$ (e.g. ReLU). Suppose the operator norm of each weight matrix and L_2 norm of each bias vector is at most R_Ω, and each $x \in \mathcal{X}$ has $\|x\| \leq R_X$. Then k_ω meets the conditions of 6.3 with dimension D and $L_K = \mathcal{O}\left(\Lambda R_\Omega^{\Lambda-1} \frac{R_X+1}{\sigma_\phi}\right)$.*

The dependence on s in Theorem 6.3 is somewhat unfortunate, but the ratio structure of J means that otherwise, errors in very small variances can hurt us arbitrarily. Even so, "near-perfect" kernels (with reasonably large MMD and very small variance) will likely still be chosen as the maximiser of the regularised criterion, even if we do not estimate the (extremely large) ratio accurately. Likewise, near-constant kernels (with very small variance but still small J) will generally have their J *under*-estimated, and so are unlikely to be selected when a better kernel is available. The ϵq component in Eq. (6.1) may also help avoid extremely small variances.

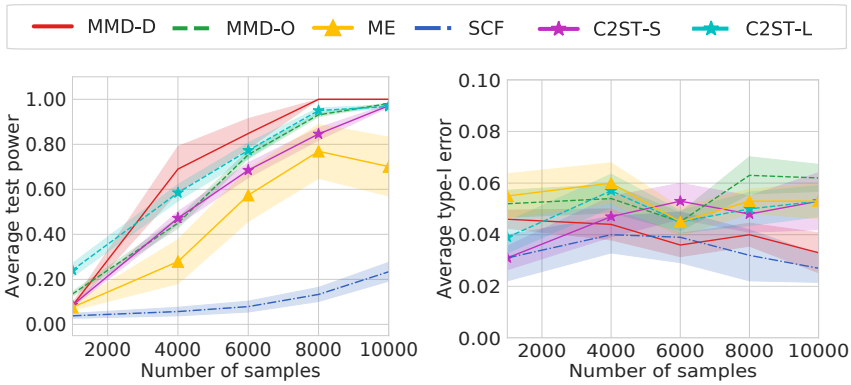
Given N data points, this result also gives insight into how many we should use to train the kernel and how many to test. With perfect optimisation, Corollary 6.3 shows a bound on the asymptotic power of the test is maximised by training on $\Theta\left(\left(N\sqrt{\log N}\right)^{\frac{3}{4}}\right)$ points, and testing on the remainder.

6.6 Experimental Results

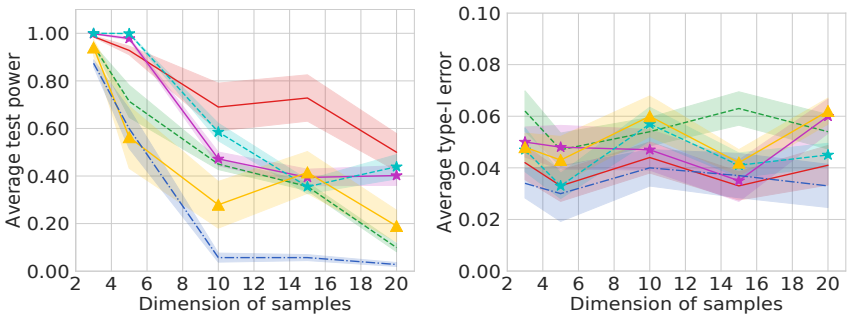
6.6.1 Comparison on Benchmark Datasets

We compare the following tests on several datasets:

- MMD-D: MMD with a deep kernel; our method described in Section 6.4.
- MMD-O: MMD with a Gaussian kernel whose length-scale is optimised as in Section 6.4. This gives better results than standard heuristics.
- Mean embedding (ME): a state-of-the-art test [Chwialkowski et al., 2015; Jitkrittum et al., 2016a] based on differences in Gaussian kernel mean embeddings at a set of optimised points.
- Smooth characteristic functions (SCF): a state-of-the-art test [Chwialkowski et al., 2015; Jitkrittum et al., 2016a] based on differences in Gaussian mean embeddings at a set of optimised frequencies.



(a) Test power v.s. n; $d = 10$ (b) Type-I error v.s. n; $d = 10$



(c) Test power v.s. d; $n = 4000$ (d) Type-I error v.s. d; $n = 4000$

Figure 6.3: Results on HDGM for $\alpha = 0.05$. Left: average test power (a) and Type-I error (b) when increasing the samples size n, keeping $d = 10$. Right: average test power (c) and Type-I error (d) when increasing the dimension d, keeping $n = 4\,000$. Shaded regions show standard errors for the mean.

- Classifier two-sample tests: tests discussed in Section 6.3 including C2STS-S [Lopez-Paz and Oquab, 2016] and C2ST-L [Cheng and Cloninger, 2019] as described in Section 6.3. We set the test thresholds via permutation for both.

For synthetic datasets, we take a single sample set for S_P^{tr} and S_Q^{tr} and learn a kernel/test locations/etc once for each method on that training set. We then evaluate its rejection rate on 100 new sample sets S_P^{te}, S_Q^{te} from the same distribution. For real datasets, we select a subset of the available data for S_P^{tr} and S_Q^{tr} and train on that; we then evaluate on 100 random subsets, disjoint from the training set, of the remaining data. We repeat this full process 10 times, and report the mean rejection rate and its standard deviation of each test. Table 6.4 shows significance tests. Further details are in Section 6.B.

***Blob* dataset** *Blob-D* is the dataset shown in Figure 6.1; *Blob-S* has Q also equal to the distribution shown in Figure 6.1(a), so that the null hypothesis holds. Details are given in Table 6.5 (Section 6.B.1). The average test power are shown in Figure 6.2(a) to demonstrate the limits of simple kernels. MMD-D and C2ST-L

Table 6.1: *Higgs* ($\alpha = 0.05$): average test power±standard error for N samples. Bold represents the highest mean per row.

n	ME	SCF	C2ST-S	C2ST-L	MMD-O	MMD-D
1 000	0.120±0.007	0.095±0.022	0.082±0.015	0.097±0.014	**0.132±0.005**	0.113±0.013
2 000	0.165±0.019	0.130±0.026	0.183±0.032	0.232±0.017	0.291±0.012	**0.304±0.035**
3 000	0.197±0.012	0.142±0.025	0.257±0.049	0.399±0.058	0.376±0.022	**0.403±0.050**
5 000	0.410±0.041	0.261±0.044	0.592±0.037	0.447±0.045	0.659±0.018	**0.699±0.047**
8 000	0.691±0.067	0.467±0.038	0.892±0.029	0.878±0.020	0.923±0.013	**0.952±0.024**
10 000	0.786±0.041	0.603±0.066	0.974±0.007	0.985±0.005	**1.000±0.000**	**1.000±0.000**
Avg.	0.395	0.283	0.497	0.506	0.564	**0.579**

are the clear winners in power, with MMD-D better in the higher-sample regime, and MMD-D is more reliable than C2STs. Figure 6.2(b) shows that $\widehat{J}$ is higher for MMD-D than MMD-O, in addition to the actual test power being better, as discussed in Section 6.2. All methods have controlled Type-I error.

High-dimensional Gaussian mixtures (*HDGM*) Here we study bi-modal Gaussian mixtures in increasing dimension. Each distribution has two Gaussian components; in *HDGM-S*, P and Q are the same, while in *HDGM-D*, P and Q differ in the covariance of a single dimension pair but are otherwise the same. Details are in Table 6.5 (Section 6.B.1). We consider both increasing n while keeping dimension $d = 10$ and increasing d while keeping $n = 4000$, with results shown in Figure 6.3. Again, MMD-D has generally the best test power across a range of problem settings, with reasonable Type-I error.

***Higgs* dataset [Baldi et al., 2014]** We compare the jet ϕ-momenta distribution ($d = 4$) of the background process, P, which lacks Higgs bosons, to the corresponding distribution Q for the process that produces Higgs bosons, following Chwialkowski et al. [2015]. As discussed in these previous works, ϕ-momenta carry very little discriminating information for recognizing whether Higgs bosons were produced. We consider a series of tests with increased number of samples n. We report average test power (comparing P to Q) in Table 6.1, and average Type-I error (comparing $P = Q$) in Table 6.6 (Section 6.B.3). As before, MMD-D generally performs the best; although the improvement over MMD-O here is not dramatic, MMD-D does notably outperform C2ST. All methods maintain reasonable Type-I errors.

***MNIST* generative model** The *MNIST* dataset contains 70 000 handwritten digit images [LeCun et al., 1998]. We compare true *MNIST* data samples P to samples Q from a pretrained deep convolutional generative adversarial network (DCGAN) [Radford et al., 2016]. Samples from both distributions are shown in Figure 6.4 (in Section 6.B.2). We consider tests for increasing numbers of samples n, and report

Table 6.2: *MNIST* ($\alpha = 0.05$): average test power±standard error for comparing N real images to N DCGAN samples.

n	ME	SCF	C2ST-S	C2ST-L	MMD-O	MMD-D
200	0.414±0.050	0.107±0.018	0.193±0.037	0.234±0.031	0.188±0.010	**0.555±0.044**
400	0.921±0.032	0.152±0.021	0.646±0.039	0.706±0.047	0.363±0.017	**0.996±0.004**
600	**1.000±0.000**	0.294±0.008	**1.000±0.000**	0.977±0.012	0.619±0.021	**1.000±0.000**
800	**1.000±0.000**	0.317±0.017	**1.000±0.000**	**1.000±0.000**	0.797±0.015	**1.000±0.000**
1 000	**1.000±0.000**	0.346±0.019	**1.000±0.000**	**1.000±0.000**	0.894±0.016	**1.000±0.000**
Avg.	0.867	0.243	0.768	0.783	0.572	**0.910**

Table 6.3: Mean test power on *Blob* ($n_b = 40$), *HDGM* ($N = 4000, d = 10$), *Higgs* ($N = 3000$) and *MNIST* ($n = 400$) for $\alpha = 0.05$. See Section 6.6.2 for the naming scheme; S+C corresponds to C2ST-S, L+C to C2ST-L, and D+J to MMD-D. L+M is the method proposed by Kirchler et al. [2020].

	S+C	L+C	G+C	D+C	L+M	G+M	D+M	L+J	G+J	D+J
Blob	0.835	0.942	0.901	0.900	0.851	0.960	0.906	0.952	0.966	**0.985**
HDGM	0.472	0.585	0.287	0.302	0.494	0.223	0.539	0.635	0.604	**0.659**
Higgs	0.257	0.399	0.353	0.384	0.321	0.254	0.379	0.295	0.364	**0.403**
MNIST	0.646	0.706	0.784	0.803	0.845	0.680	0.760	0.935	0.976	**0.996**
Avg.	0.553	0.658	0.581	0.597	0.628	0.529	0.646	0.704	0.727	**0.761**

average test power (for P to Q) in Table 6.2 and average Type-I error ($P = Q$) in Table 6.7 (in Section 6.B.3). MMD-D substantially outperforms its competitors in test power, with the desired Type-I error. ME also does well in this case: it is perhaps particularly suited to this problem, since it is capable of identifying either modes dropped by the generative model or spurious modes it inserts.

6.6.2 Ablation Study

We now study in more detail the difference between MMD-D and closely related methods. Recall from Section 6.3 that there are two main differences between MMD-D and C2STs: first, using a "full" kernel Eq. (6.1) rather than the sign-based kernel Eq. (6.5) or the intermediate linear kernel Eq. (6.6). Second, training to maximise $\widehat{J}_\lambda$ Eq. (6.3) rather than a cross-entry surrogate. MMD-D uses a full kernel Eq. (6.1) trained for test power; C2ST-S effectively uses the sign kernel Eq. (6.5) trained for cross entropy.

In this section, we consider the performance of several intermediate models empirically, demonstrating that both factors help in testing. All are based on the same feature extraction architecture ϕ_ω; some models add a classification layer with new parameters w and b,

$$f_\omega(x) = w^\mathsf{T}\phi_\omega(x) + b,$$

Table 6.4: Paired t-test results ($\alpha = 0.05$) for the results of Section 6.6.1. For *HDGM*, we fix $d = 10$ (corresponding to Figure 6.3a). ✓ indicates MMD-D achieved statistically significantly higher mean test power than the other method, $\times$ that it did not.

Dataset	ME	SCF	C2ST-S	C2ST-L	MMD-O
Blob	✓	✓	✓	×	×
HDGM	✓	✓	✓	✓	✓
Higgs	✓	✓	✓	×	×
MNIST	✓	✓	✓	✓	✓

which is treated as classification logits. The model variants we consider is

- **S** A kernel $\mathbb{1}(f_\omega(x) > 0)\mathbb{1}(f_\omega(y) > 0)$; corresponds to a test statistic of the accuracy of f (Theorem 6.1).
- **L** A kernel $f_\omega(x)f_\omega(y)$; corresponds to a test statistic comparing the mean value of f (Theorem 6.2).
- **G** A Gaussian kernel $\kappa(\phi_\omega(x), \phi_\omega(y))$.
- **D** The deep kernel Eq. (6.1) based on ϕ_ω.

We combine the model variants with a suffix describing the optimisation objective:

- **J** Choose ω, including possibly w and b, to optimise the approximate test power Eq. (6.3).
- **M** Choose ω, including possibly w and b, to maximise the value of the empirical MMD between two samples.[5]
- **C** Choose ω, including w and b, to optimise cross-entropy using the classifier that specifies the probability of x belonging to P as $1/\left(1 + \exp(-f_\omega(x))\right)$.[6]

Table 6.3 presents results for all of these methods (except for **S+J**, which is non-differentiable and hence difficult to optimise). Performance generally improves as we move from **S** to **L** to **G** to **D**, and from **C** to **J**, and from **M** to **J**.

[5]If a deep kernel is unbounded, directly maximising MMD will make optimised parameters of ϕ_ω be infinite. Thus, for L+M, we consider a normalised linear deep kernel: $\tanh(f_\omega(x)/\|S\|_{\mathrm{F}})\tanh(f_\omega(y)/\|S\|_{\mathrm{F}})$, where $S = [S_P; S_Q]$ and $\|\cdot\|_F$ is the Frobenius norm.

[6]G+C and D+C take the fixed ϕ_ω embeddings, then find the optimal length-scale/etc by optimising $\widehat{J}_\lambda$.

Architecture design of deep kernels

For *Blob*, *HDGM* and *Higgs*, ϕ_ω is a five-layer fully-connected neural network, with softplus activations. In general, we expect the fully-connected networks, to be reasonable choices for datasets where strong structural assumptions are not known. For *MNIST* dataset, ϕ_ω is a *convolutional neural network* (CNN) that contains four convolutional layers and one fully-connected layer. The structure of the CNN follows the structure of the feature extractor in the DCGAN's discriminator [Radford et al., 2016] . In general, we expect GAN discriminator architectures to work well for image datasets, as the problem is closely related.

Appendices

6.A Proofs and Derivations

Section 6.A.2 proves the main results under some assumptions about the kernel parametrisation, using intermediate results about uniform convergence of our estimators in Section 6.A.3. Section 6.A.4 then shows that these assumptions hold for different settings of kernel learning.

6.A.1 Preliminaries

Given a kernel k_ω and sample sets $\{X_i\}_{i=1}^n \sim P^n$, $\{Y_i\}_{i=1}^n \sim Q^n$, define the $n \times n$ matrix

$$\tilde{G}_{ij}^{(\omega)} = k_\omega(X_i, X_j) + k_\omega(Y_i, Y_j) - k_\omega(X_i, Y_j) - k_\omega(X_j, Y_i);$$

we will often omit ω when it is clear from context. The U-statistic estimator of the squared MMD Eq. (2.12) is

$$\hat{\eta}_\omega = \frac{1}{n(n-1)} \sum_{i \neq j} \tilde{G}_{ij}.$$

The squared MMD is $\eta_\omega = \mathbb{E}[\tilde{G}_{12}]$. The variance of $\hat{\eta}_\omega$ is given by Lemma 6.1.

Lemma 6.1. *For a fixed kernel k_ω and random sample sets $\{x_i\}_{i=1}^n$, $\{y_i\}_{i=1}^n$, we have*

$$\mathbb{V}ar[\hat{\eta}_\omega] = \frac{4(n-2)}{n(n-1)} \xi_1^{(\omega)} + \frac{2}{n(n-1)} \xi_2^{(\omega)} = \frac{4}{n} \xi_1^{(\omega)} + \frac{2\xi_2^{(\omega)} - 4\xi_1^{(\omega)}}{n(n-1)}, \tag{6.7}$$

where

$$\xi_1^{(\omega)} = \mathbb{E}\left[\tilde{G}_{12}^{(\omega)} \tilde{G}_{13}^{(\omega)}\right] - \mathbb{E}\left[\tilde{G}_{12}^{(\omega)}\right]^2, \qquad \xi_2^{(\omega)} = \mathbb{E}\left[\left(\tilde{G}_{12}^{(\omega)}\right)^2\right] - \mathbb{E}\left[\tilde{G}_{12}^{(\omega)}\right]^2.$$

Thus as $n \to \infty$,

$$n\mathbb{V}ar[\hat{\eta}_\omega] \to 4\xi_1^{(\omega)} =: \sigma_\omega^2.$$

Proof. Let U denote the pair (X, Y), and $h_\omega(U, U') = k_\omega(X, X') + k_\omega(Y, Y') - k_\omega(X, Y') - k_\omega(X', Y)$, so that $\tilde{G}_{ij}^{(\omega)} = g_\omega(U_i, U_j)$. Via Lemma A in Section 5.2.1 of Serfling [2009], we know that Eq. (6.7) holds with

$$\begin{aligned}\xi_1^{(\omega)} &= \mathbb{V}ar_U\left[\mathbb{E}_{U'}\left[g_\omega(U, U')\right]\right] \\ &= \mathbb{E}_U\left[\mathbb{E}_{U'}\left[g_\omega(U, U')\right]\mathbb{E}_{U''}\left[g_\omega(U, U'')\right]\right] - \mathbb{E}_U\left[\mathbb{E}_{U'}[g_\omega(U, U')]\right]^2 \\ &= \mathbb{E}[\tilde{G}_{12}^{(\omega)}\tilde{G}_{13}^{(\omega)}] - \mathbb{E}[\tilde{G}_{12}^{(\omega)}]^2\end{aligned}$$

and

$$\xi_2 = \mathbb{V}ar_{U,U'}\left[g_\omega(U, U')\right] = \mathbb{E}\left[\left(\tilde{G}_{12}^{(\omega)}\right)^2\right] - \mathbb{E}\left[\tilde{G}_{12}^{(\omega)}\right]^2. \qquad \square$$

We use a V-statistic estimator Eq. (6.4) for σ_ω^2:

$$\hat{\sigma}_\omega^2 = 4\left(\frac{1}{n}\sum_{i=1}^n\left(\frac{1}{n}\sum_{j=1}^n \tilde{G}_{ij}^{(\omega)}\right)^2 - \left(\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \tilde{G}_{ij}^{(\omega)}\right)^2\right).$$

As a V-statistic, $\hat{\sigma}_\omega^2$ is biased. In fact, Sutherland et al. [2016] and Sutherland [2019] provide an unbiased estimator of $\mathbb{V}ar[\hat{\eta}_\omega]$ – including the terms of order $\frac{1}{n(n-1)}$. Although this estimator takes the same quadratic time to compute as Eq. (6.4), it contains many more terms, which are cumbersome both for implementation and for analysis. Eq. (6.4) is also marginally more convenient in that it is always at least non-negative. As we show in Lemma 6.3, the amount of bias is negligible as n increases. In practice, we expect the difference to be unimportant – or the V-statistic may in fact be beneficial, since underestimating σ^2 harms the estimate of η/σ^2 more than overestimating it does.

Similarly, although we use the U-statistic estimator Eq. (2.12), it would be very similar to use the biased estimator $n^{-2}\sum_{ij}\tilde{G}_{ij}$, or the minimum variance unbiased estimator $n^{-1}(n-1)^{-1}\sum_{i\neq j}(k(X_i, X_j) + k(Y_i, Y_j)) - 2n^{-2}\sum_{ij}k(X_i, Y_J)$. Showing comparable concentration behavior to Proposition 6.5 is trivially different, and in fact it is also not difficult to show σ_ω^2 is the same for all three estimators (up to lower-order terms).

6.A.2 Main results

We will require the following assumptions. These are fairly agnostic as to the kernel form; Section 6.A.4.2 shows that these assumptions hold (and gives the constants) for the kernels Eq. (6.1) we use in the paper.

(**A**) The kernels k_ω are uniformly bounded:

$$\sup_{\omega\in\Omega}\sup_{x\in\mathcal{X}} k_\omega(x,x) \leq \nu.$$

For the kernels we use in practice, $\nu = 1$.

(**B**) The possible kernel parameters ω lie in a Banach space of dimension D. Furthermore, the set of possible kernel parameters Ω is bounded by R_ω, $\Omega \subseteq \{\omega \mid \|\omega\| \leq R_\Omega\}$.

Section 6.A.4.2 builds this space and its norm for the kernels we use in the paper.

(**C**) The kernel parametrisation is Lipschitz: for all $x, y \in \mathcal{X}$ and $\omega, \omega' \in \Omega$,

$$|k_\omega(x,y) - k_{\omega'}(x,y)| \leq L_k\|\omega - \omega'\|.$$

Proposition 6.8 in Section 6.A.4.2 gives an expression for L_k for the kernels we use in the paper.

We will first show the main results under these general assumptions, using uniform convergence results shown in Section 6.A.3, then show Assumptions (**B**) and (**C**) for particular kernels in Section 6.A.4.2.

Theorem 6.4. *Under Assumptions (**A**) to (**C**), let $\bar{\Omega}_s \subseteq \Omega$ be the set of kernel parameters for which $\sigma_\omega^2 \geq s^2$, and assume $\nu \geq 1$. Take $\lambda = n^{-1/3}$. Then, with probability at least $1-\delta$,*

$$\sup_{\omega\in\bar{\Omega}_s}\left|\frac{\hat{\eta}_\omega}{\hat{\sigma}_{\omega,\lambda}} - \frac{\eta_\omega}{\sigma_\omega}\right| \leq \frac{2\nu}{s^2 n^{1/3}}\Big(\frac{1}{s} + \frac{2304\nu^2}{\sqrt{n}} + \left[\frac{4s}{n^{1/6}} + 1024\nu\right] \cdot \left[L_k + \sqrt{2\log\frac{2}{\delta} + 2D\log\left(4R_\Omega\sqrt{n}\right)}\right]\Big),$$

and thus, treating ν as a constant,

$$\sup_{\omega\in\bar{\Omega}_s}|\frac{\hat{\eta}_\omega}{\hat{\sigma}_{\omega,\lambda}} - \frac{\eta_\omega}{\sigma_\omega}| = O_P\left(\frac{1}{s^2 n^{1/3}}\left[\frac{1}{s} + L_k + \sqrt{D}\right]\right).$$

Proof. Let $\sigma^2_{\omega,\lambda} := \sigma^2_\omega + \lambda$. Using $|\hat{\eta}_\omega| \le 4\nu$, we begin by decomposing

$$\begin{aligned}
&\sup_{\omega\in\bar{\Omega}_s} |\frac{\hat{\eta}_\omega}{\hat{\sigma}_{\omega,\lambda}} - \frac{\eta_\omega}{\sigma_\omega}| \le \sup_{\omega\in\bar{\Omega}_s} |\frac{\hat{\eta}_\omega}{\hat{\sigma}_{\omega,\lambda}} - \frac{\hat{\eta}_\omega}{\sigma_{\omega,\lambda}}| + \sup_{\omega\in\bar{\Omega}_s} |\frac{\hat{\eta}_\omega}{\sigma_{\omega,\lambda}} - \frac{\hat{\eta}_\omega}{\sigma_\omega}| + \sup_{\omega\in\bar{\Omega}_s} |\frac{\hat{\eta}_\omega}{\sigma_\omega} - \frac{\eta_\omega}{\sigma_\omega}| \\
&= \sup_{\omega\in\bar{\Omega}_s} |\hat{\eta}_\omega| \frac{1}{\hat{\sigma}_{\omega,\lambda}} \frac{1}{\sigma_{\omega,\lambda}} \frac{|\hat{\sigma}^2_{\omega,\lambda} - \sigma^2_{\omega,\lambda}|}{\hat{\sigma}_{\omega,\lambda} + \sigma_{\omega,\lambda}} + \sup_{\omega\in\bar{\Omega}_s} |\hat{\eta}_\omega| \frac{1}{\sigma_{\omega,\lambda}} \frac{1}{\sigma_\omega} \frac{|\sigma^2_{\omega,\lambda} - \sigma^2_\omega|}{\sigma_{\omega,\lambda} + \sigma_\omega} + \sup_{\omega\in\bar{\Omega}_s} \frac{1}{\sigma_\omega} |\hat{\eta}_\omega - \eta_\omega| \\
&\le \sup_{\omega\in\bar{\Omega}_s} \frac{4\nu}{\sqrt{\lambda}\, s\, (s + \sqrt{\lambda})} |\hat{\sigma}^2_\omega - \sigma^2_\omega| + \frac{4\nu\lambda}{\sqrt{s^2+\lambda}\, s\, \left(\sqrt{s^2+\lambda} + s\right)} + \sup_{\omega\in\bar{\Omega}_s} \frac{1}{s} |\hat{\eta}_\omega - \eta_\omega| \\
&\le \frac{4\nu}{s^2\sqrt{\lambda}} \sup_{\omega\in\Omega} |\hat{\sigma}^2_\omega - \sigma^2_\omega| + \frac{2\nu}{s^3}\lambda + \frac{1}{s} \sup_{\omega\in\Omega} |\hat{\eta}_\omega - \eta_\omega|.
\end{aligned}$$

Propositions 6.5 and 6.6 show uniform convergence of $\hat{\eta}_\omega$ and $\hat{\sigma}^2_\omega$, respectively. Thus, with probability at least $1 - \delta$, the error is at most

$$\begin{aligned}
&\frac{2\nu}{s^3}\lambda + \left[\frac{8\nu}{s\sqrt{n}} + \frac{1792\nu}{\sqrt{n}s^2\sqrt{\lambda}}\right] \sqrt{2\log\frac{2}{\delta} + 2D\log\left(4R_\Omega\sqrt{n}\right)} \\
&\qquad + \left[\frac{8}{s\sqrt{n}} + \frac{2048\nu^2}{\sqrt{n}s^2\sqrt{\lambda}}\right] L_k + \frac{4608\nu^3}{s^2 n\sqrt{\lambda}}.
\end{aligned}$$

Taking $\lambda = n^{-1/3}$ gives

$$\begin{aligned}
&\frac{2\nu}{s^3 n^{1/3}} + \left[\frac{8\nu}{s\sqrt{n}} + \frac{1792\nu}{s^2 n^{1/3}}\right] \sqrt{2\log\frac{2}{\delta} + 2D\log\left(4R_\Omega\sqrt{n}\right)} \\
&\qquad + \left[\frac{8}{s\sqrt{n}} + \frac{2048\nu^2}{s^2 n^{1/3}}\right] L_k + \frac{4608\nu^3}{s^2 n^{5/6}}.
\end{aligned}$$

Using $1 \le \nu$, $1792 < 2048$, we can get the slightly simpler upper bound

$$\frac{2\nu}{s^3 n^{1/3}} + \left[\frac{8\nu}{s\sqrt{n}} + \frac{2048\nu^2}{s^2 n^{1/3}}\right] \left[L_k + \sqrt{2\log\frac{2}{\delta} + 2D\log\left(4R_\Omega\sqrt{n}\right)}\right] + \frac{4608\nu^3}{s^2 n^{5/6}}. \quad \square$$

It is worth noting that, if we are particularly concerned about the s dependence, we can make some slightly different choices in the decomposition to improve the dependence on s while worsening the rate with n.

Corollary 6.1. *In the setup of Theorem 6.4, additionally assume that there is a unique population maximiser ω^* of J from Eq.* (6.2)*, i.e. for each $t > 0$ we have*

$$\sup_{\omega\in\bar{\Omega}_s : \|\omega - \omega^*\| \ge t} J(P, Q; k_\omega) < J(P, Q; k_{\omega^*}).$$

For each n, let $S_P^{(n)}$ and $S_Q^{(n)}$ be sequences of sample sets of size n, let $\widehat{J}_n(\omega)$ de-

note $J_{\lambda=n^{-1/3}}(S_P^{(n)}, S_Q^{(n)}; k_\omega)$*, and take* $\hat{\omega}_n^*$ *to be a maximiser of* $\widehat{J}_n(\omega)$.[7] *Then* $\hat{\omega}_n^*$ *converges in probability to* ω^*.

Proof. By Theorem 6.4, $\sup_{\omega\in\bar{\Omega}_s}|\widehat{J}_n(\omega) - J(\omega)| \xrightarrow{P} 0$. Then the result follows by Theorem 5.7 of Van der Vaart [2000]. □

Corollary 6.2. *In the setup of Theorem 6.4, suppose we use* n *sample points to select a kernel* $\hat{\omega}_n \in \arg\max_{\omega\in\bar{\Omega}_s} \widehat{J}_\lambda(\omega)$ *and* m *sample points to run a test of level* α*. Let* $r_{\hat{\omega}_n}^{(m)}$ *denote the rejection threshold for a test with that kernel of size* m*. Define* $J^* := \sup_{\omega\in\bar{\Omega}_s} J(\omega)$*, and constants* C*,* C'*,* C''*,* N_0 *depending on* ν*,* L_k*,* D*,* R_Ω *and* s*. For any* $n \geq N_0$*, with probability at least* $1-\delta$*, this test procedure has power*

$$\mathbb{P}\left(m\,\hat{\eta}_{\hat{\omega}_n} > r_{\hat{\omega}_n}^{(m)}\right) \geq \Phi\left(\sqrt{m}J^* - C\frac{\sqrt{m}}{n^{\frac{1}{3}}}\sqrt{\log\frac{n}{\delta}} - C'\sqrt{\log\frac{1}{\alpha}}\right) - \frac{C''}{\sqrt{m}}.$$

Proof. Let $\hat{\omega}_n \in \arg\max_{\omega\in\bar{\Omega}_s} \widehat{J}_\lambda(\omega)$. By Theorem 6.4, there are some N_0, C depending on ν, L_k, D, R_Ω, and s such that as long as $n \geq N_0$, with probability at least $1-\delta$ it holds that

$$\sup_{\omega\in\bar{\Omega}_s}|J_\lambda(\omega) - J(\omega)| \leq \tfrac{1}{2}Cn^{-\frac{1}{3}}\sqrt{\log\frac{n}{\delta}} =: \epsilon_n.$$

Assume for the remainder of this proof that this event holds. Letting $\omega^* \in \arg\max J(\omega)$, we know because $\hat{\omega}_n$ maximises $\widehat{J}_\lambda$ that $\widehat{J}_\lambda(\hat{\omega}_n) \geq \widehat{J}_\lambda(\omega^*)$. Using uniform convergence twice,

$$J(\hat{\omega}_n) \geq \widehat{J}_\lambda(\hat{\omega}_n) - \epsilon_n \geq \widehat{J}_\lambda(\omega^*) - \epsilon_n \geq (J(\omega^*) - \epsilon_n) - \epsilon_n = J^* - 2\epsilon_n.$$

Now, although Proposition 6.1 establishes that $r_\omega^{(m)} \to r_\omega$ and it is even known [Korolyuk and Borovskikh, 1988, Theorem 5] that $|r_\omega^{(m)} - r_\omega|$ is $o(1/\sqrt{m})$, the constant in that convergence will depend on the choice of ω in an unknown way. It's thus simpler to use the very loose but uniform (McDiarmid-based) bound given by Corollary 11 of Gretton et al. [2012a], which implies $r_\omega^{(m)} \leq 4\nu\sqrt{\log(\alpha^{-1})m}$ no matter the choice of ω.

We will now need a more precise characterisation of the power than that provided by the central limit theorem of Proposition 6.1. Callaert and Janssen [1978] provide such a result, a Berry-Esseen bound on U-statistic convergence: there is

[7] In fact, it suffices for the $\hat{\omega}_n^*$ to only approximately maximise $\widehat{J}_n$, as long as their suboptimality is $o_P(1)$.

some absolute constant $C'_{BS} = 2^3 4^3 C_{BS}$ such that

$$\sup_t |\mathbb{P}_{H_1}\left(\sqrt{m}\frac{\hat{\eta}_\omega - \eta_\omega}{\sigma_\omega^2} \leq t\right) - \Phi(t)| \leq \frac{C'_{BS}\mathbb{E}|\tilde{G}_{12}|^3}{(\sigma_\omega/2)^3\sqrt{m}} \leq \frac{C_{BS}\nu^3}{\sigma_\omega^3\sqrt{m}}.$$

Letting $r_\omega^{(m)}$ be the appropriate rejection threshold for k_ω with m samples, the power of a test with kernel k_ω is

$$\begin{aligned}
\mathbb{P}\left(m\hat{\eta}_\omega > r_\omega^{(m)}\right) &= \mathbb{P}\left(\sqrt{m}\frac{\hat{\eta}_\omega - \eta_\omega}{\sigma_\omega} > \frac{r_\omega^{(m)}}{\sqrt{m}\sigma_\omega} - \sqrt{m}\frac{\eta_\omega}{\sigma_\omega}\right) \\
&\geq \Phi\left(\sqrt{m}J(\omega) - \frac{r_\omega^{(m)}}{\sqrt{m}\sigma_\omega}\right) - \frac{C_{BS}\nu^3}{\sigma_\omega^3\sqrt{m}} \\
&\geq \Phi\left(\sqrt{m}J(\omega) - \frac{r_\omega^{(m)}}{s\sqrt{m}}\right) - \frac{C''}{\sqrt{m}},
\end{aligned}$$

using a new constant $C'' := C_{BS}\nu^3/s^3$. Combining the previous results on $J(\hat{\omega}_n)$ and $r_{\hat{\omega}_n}^{(m)}$ yields the claim. □

Corollary 6.3. *In the setup of Corollary 6.2, suppose we are given N data points to divide between n training points and $m = N - n$ testing points, and $\delta < 0.22$ is fixed. Ignoring the Berry-Esseen convergence term outside of Φ, the asymptotic power upper bound*

$$\Phi\left(\sqrt{m}J^* - C\frac{\sqrt{m}}{n^{\frac{1}{3}}}\sqrt{\log\frac{n}{\delta}} - C'\sqrt{\log\frac{1}{\alpha}}\right)$$

is maximised only when, as other quantities remain constant,

$$\lim_{N\to\infty} \frac{n}{\left(\frac{C}{\sqrt{3}J^*}N\sqrt{\log N}\right)^{\frac{3}{4}}} = 1.$$

Proof. Because the C' term is constant, we wish to choose

$$\arg\max_{0<n<N} \frac{J^*}{C}\sqrt{N-n} - \frac{\sqrt{N-n}}{n^{\frac{1}{3}}}\sqrt{\log\frac{n}{\delta}}.$$

Clearly neither endpoint is optimal. Relaxing n to be real-valued, the optimum must be achieved at a stationary point, where

$$\frac{-J^*}{2C\sqrt{N-n}} + \frac{\sqrt{\log\frac{n}{\delta}}}{2\sqrt{N-n}\,n^{\frac{1}{3}}} + \frac{1}{3}\sqrt{N-n}\,n^{-\frac{4}{3}}\sqrt{\log\frac{n}{\delta}} - \frac{1}{2}\sqrt{N-n}\,n^{-\frac{4}{3}}\left(\log\frac{n}{\delta}\right)^{-\frac{1}{2}} = 0.$$

Multiplying by $2\sqrt{N-n}\,n^{\frac{4}{3}}\sqrt{\log\frac{n}{\delta}}$ and rearranging, we get that a stationary point is achieved exactly when

$$\underbrace{\frac{1}{3}\left[n+2N\right]\log\frac{n}{\delta}+n}_{D}=\underbrace{\frac{J^*}{C}n^{\frac{4}{3}}\sqrt{\log\frac{n}{\delta}}+N}_{E}.$$

Now write, without loss of generality, $n=\left(A_N N\sqrt{\log N}\right)^{\frac{3}{4}}$, and so

$$D=\frac{1}{3}\left[A_N^{\frac{3}{4}}N^{\frac{3}{4}}(\log N)^{\frac{3}{8}}+2N\right]\left[\underbrace{\frac{3}{4}\log A_N+\frac{3}{4}\log N+\frac{3}{8}\log\log N}_{\log n}+\log\frac{1}{\delta}\right]$$
$$+A_N^{\frac{3}{4}}N^{\frac{3}{4}}(\log N)^{\frac{3}{8}}$$
$$E=\frac{J^*}{C}A_N N\sqrt{\log N}\sqrt{\underbrace{\frac{3}{4}\log A_N+\frac{3}{4}\log N+\frac{3}{8}\log\log N}_{\log n}+\log\frac{1}{\delta}}+N.$$

We will show that $D-E\to 0$ requires $A_N\to C/(\sqrt{3}J^*)$, implying the result.

We first suppose $A_N=\omega(1)$, further breaking into cases which result in different terms inside D and E becoming dominant:

If $A_N=\Omega(N)$, $D=\Theta\left(A_N^{\frac{3}{4}}N^{\frac{3}{4}}(\log N)^{\frac{3}{8}}\log A_N\right)$,
$E=\Theta\left(A_N N\sqrt{\log(N)\log(A_N)}\right)$.

If $A_N=\Omega\left(\frac{N^{\frac{1}{3}}}{\sqrt{\log N}}\right)$, $A_N=o(N)$, $D=\Theta\left(A_N^{\frac{3}{4}}N^{\frac{3}{4}}(\log N)^{\frac{3}{8}}\log N\right)$,
$E=\Theta\left(A_N N\log N\right)$.

If $A_N=\omega(1)$, $A_N=o\left(\frac{N^{\frac{1}{3}}}{\sqrt{\log N}}\right)$, $D=\Theta\left(N\log N\right)$,
$E=\Theta\left(A_N N\log N\right)$.

In each case, $E=\omega(D)$ and so $D-E\to-\infty$, contradicting that $D=E$. Thus a stationary point requires $A_N=\mathcal{O}(1)$ for a stationary point.

We now do the same for $A_N=o(1)$. First, clearly $n\geq 1$; suppose that in fact $n=\Theta(1)$, i.e. $A_N=\Theta\left(1/(N\sqrt{\log N})\right)$. In this case, we would have $D=\frac{2}{3}N\log\frac{n}{\delta}+\Theta(1)$ and $E=N+\Theta(1)$, so that $D=E$ requires $\frac{2}{3}\log\frac{n}{\delta}\to 1$, i.e. $n\to\delta\exp\frac{3}{2}\approx 4.5\,\delta$. For $\delta<0.22$, this contradicts $n\geq 1$. So we know that

$\log n = \omega(1)$. Now, the remaining options for A_N all yield $D - E \to \infty$:

$$\text{If } A_N = o(1),\ A_N = \Omega\left(\frac{1}{\log N}\right), D = \Theta\left(N \log n\right), E = \Theta\left(A_N N \log n\right).$$
$$\text{If } A_N = o\left(\frac{1}{\log N}\right),\ A_N = \omega\left(\frac{1}{N\sqrt{\log N}}\right), D = \Theta\left(N \log n\right), E = \Theta\left(N\right).$$

Thus we have established that $A_N = \Theta(1)$. Thus, we obtain that

$$D = \frac{1}{2} N \log N + \mathcal{O}\left(N\right) \qquad E = \frac{\sqrt{3}J^*}{2C} A_N N \log N + \mathcal{O}\left(N\sqrt{\log N}\right).$$

Asymptotic equality hence requires $A_N \to C/(\sqrt{3}J^*)$. □

6.A.3 Uniform convergence results

These results, on the uniform convergence of $\hat{\eta}$ and $\hat{\sigma}^2$, were used in the proof of Theorem 6.4.

Proposition 6.5. *Under Assumptions (**A**) to (**C**), we have that with probability at least* $1 - \delta$*,*

$$\sup_{\omega} |\hat{\eta}_\omega - \eta_\omega| \leq \frac{8}{\sqrt{n}} \left[\nu\sqrt{2\log\frac{2}{\delta} + 2D\log\left(4R_\Omega\sqrt{n}\right)} + L_k\right].$$

Proof. Theorem 7 of Sriperumbudur et al. [2009] gives a similar bound in terms of Rademacher chaos complexity, but for ease of combination with our bound on convergence of the variance estimator, we use a simple ϵ-net argument instead. We study the random error function

$$\Delta(\omega) := \hat{\eta}_\omega - \eta_\omega.$$

First, we place T points $\{\omega_i\}_{i=1}^T$ such that for any point $\omega \in \Omega$, $\min_i \|\omega - \omega_i\| \leq q$; Assumption (**B**) ensures this is possible with at most $T = (4R_\Omega/q)^D$ points [Cucker and Smale, 2001, Proposition 5].

Now, $\mathbb{E}\Delta = 0$, because $\hat{\eta}$ is unbiased. Recall that $\hat{\eta} = \frac{1}{n(n-1)}\sum_{i\neq j}\tilde{G}_{ij}$, and via Assumption (**A**) we know $|\tilde{G}_{ij}| \leq 4\nu$. This $\hat{\eta}$, and hence Δ, satisfies bounded differences: if we replace (X_1, Y_1) with (X_1', Y_1'), obtaining $\hat{\eta}' = \frac{1}{n(n-1)}\sum_{i\neq j}\tilde{F}_{ij}$

where $\tilde{F}$ agrees with $\tilde{G}$ except when i or j is 1, then

$$|\hat{\eta} - \hat{\eta}'| \leq \frac{1}{n(n-1)} \sum_{i \neq j} |\tilde{G}_{ij} - \tilde{F}_{ij}| = \frac{1}{n(n-1)} \sum_{i>1} |\tilde{G}_{i1} - \tilde{F}_{i1}| + \frac{1}{n(n-1)} \sum_{j>1} |\tilde{G}_{1j} - \tilde{F}_{1j}|$$
$$\leq \frac{2}{n(n-1)} \sum_{i>1} 8\nu = \frac{16\nu}{n}.$$

Using McDiarmid's inequality for each $\Delta(\omega_i)$ and a union bound, we then obtain that with probability at least $1 - \delta$,

$$\max_{i \in \{1,\dots,T\}} |\Delta(\omega)| \leq \frac{16\nu}{\sqrt{2n}} \sqrt{\log \frac{2T}{\delta}} \leq \frac{8\nu}{\sqrt{n}} \sqrt{2 \log \frac{2}{\delta} + 2D \log \frac{4R_\Omega}{q}}.$$

We also have via Assumption (**C**), for any two $\omega, \omega' \in \Omega$,

$$|\hat{\eta}_\omega - \hat{\eta}_{\omega'}| \leq \frac{1}{n(n-1)} \sum_{i \neq j} |\tilde{G}_{ij}^{(\omega)} - \tilde{G}_{ij}^{(\omega')}|$$
$$\leq \frac{1}{n(n-1)} \sum_{i \neq j} 4L_k \|\omega - \omega'\| = 4L_k \|\omega - \omega'\|$$
$$|\eta_\omega - \eta_{\omega'}| = |\mathbb{E}\left[\tilde{G}_{12}^{(\omega)}\right] - \mathbb{E}\left[\tilde{G}_{12}^{(\omega')}\right]| \leq \mathbb{E}|\tilde{G}_{12}^{(\omega)} - \tilde{G}_{12}^{(\omega')}| \leq 4L_k \|\omega - \omega'\|$$

so that $\|\Delta\|_L \leq 8L_k$. Combining these two results, we know that with probability at least $1 - \delta$

$$\sup_\omega |\Delta(\omega)| \leq \max_{i \in \{1,\dots,T\}} |\Delta(\omega_i)| + 8L_k q \leq \frac{8\nu}{\sqrt{n}} \sqrt{2 \log \frac{2}{\delta} + 2D \log \frac{4R_\Omega}{q}} + 8L_k q;$$

setting $q = 1/\sqrt{n}$ yields the desired result. □

Proposition 6.6. *Under Assumptions (**A**) to (**C**), with probability at least* $1 - \delta$,

$$\sup_{\omega \in \Omega} \left|\hat{\sigma}_\omega^2 - \sigma_\omega^2\right| \leq \frac{64}{\sqrt{n}} \left[7\sqrt{2 \log \frac{2}{\delta} + 2D \log\left(4R_\Omega \sqrt{n}\right)} + \frac{18\nu^2}{\sqrt{n}} + 8L_k\nu\right].$$

Proof. We again use an ϵ-net argument on the (random) error function

$$\Delta(\omega) := \hat{\sigma}_{k_\omega}^2 - \sigma_{k_\omega}^2.$$

First, choose T points $\{\omega_i\}_{i=1}^T$ such that for any point $\omega \in \Omega$, $\min_i \|\omega - \omega_i\| \leq q$; again, via Assumption (**B**) and Proposition 5 of Cucker and Smale [2001] we have $T \leq (4R_\Omega/q)^D$. By Lemmas 6.2 and 6.3 and a union bound, with probability at

least $1-\delta$,

$$\max_{i\in\{1,\dots,T\}}|\Delta(\omega)| \le 448\sqrt{\frac{2}{n}\log\frac{2T}{\delta}} + \frac{1152\nu^2}{n}$$
$$\le 448\sqrt{\frac{2}{n}\log\frac{2}{\delta} + \frac{2}{n}D\log\frac{4R_\Omega}{q}} + \frac{1152\nu^2}{n}.$$

Lemma 6.4 shows that $\|\Delta\|_L \le 512L_k\nu$, which means that with probability at least $1-\delta$,

$$\sup_{\omega\in\Omega}|\Delta(\omega)| \le 448\sqrt{\frac{2}{n}\log\frac{2}{\delta} + \frac{2}{n}D\log\frac{4R_\Omega}{q}} + \frac{1152\nu^2}{n} + 512L_k\nu q. \tag{6.8}$$

Taking $q = 1/\sqrt{n}$ gives the desired result. □

Lemma 6.2. *For any kernel k bounded by ν (Assumption (**A**)), with probability at least $1-\delta$,*

$$|\hat\sigma_k^2 - \mathbb{E}\hat\sigma_k^2| \le 448\sqrt{\frac{2}{n}\log\frac{2}{\delta}}.$$

Proof. We simply apply McDiarmid's inequality to $\hat\sigma_k^2$. Suppose we change (X_1, Y_1) to (X_1', Y_1'), giving a new $\tilde G$ matrix $\tilde F$ which agrees with $\tilde G$ on all but the first row and column. Note that $|\tilde G_{ij}| \le 4\nu$, and recall

$$\hat\sigma_k^2 = 4\left(\frac{1}{n^3}\sum_i\left(\sum_j \tilde G_{ij}\right)^2 - \left(\frac{1}{n^2}\sum_{ij}\tilde G_{ij}\right)^2\right).$$

The first term in the parentheses of $\hat\sigma_k^2$ changes by

$$\left|\frac{1}{n^3}\sum_i\left(\sum_j \tilde G_{ij}\right)^2 - \frac{1}{n^3}\sum_i\left(\sum_j \tilde F_{ij}\right)^2\right| \le \frac{1}{n^3}\sum_{ij\ell}|\tilde G_{ij}\tilde G_{i\ell} - \tilde F_{ij}\tilde F_{i\ell}|.$$

In this sum, if none of i, j, or ℓ are one, the term is zero. The n^2 terms for which $i=1$ are each upper-bounded by $32\nu^2$, simply bounding each $\tilde G$ or $\tilde F$ by 4ν. Of the remainder, there are $(n-1)$ terms where $j=\ell=1$, each $|\tilde G_{i1}^2 - \tilde F_{i1}^2| \le 16\nu^2$. We are left with $2(n-1)^2$ terms which have exactly one of j or ℓ equal to 1; the $j=1$ terms are $|\tilde G_{i1}\tilde G_{i\ell} - \tilde F_{i1}\tilde G_{i\ell}| \le |\tilde G_{i1} - \tilde F_{i1}||\tilde G_{i\ell}| \le (8\nu)(4\nu)$, so each of these terms is at most $32\nu^2$. The total sum is thus at most

$$\frac{1}{n^3}\left(n^2 32\nu^2 + (n-1)16\nu^2 + 2(n-1)^2 32\nu^2\right) = \left(\frac{6}{n} - \frac{7}{n^2} + \frac{3}{n^3}\right)16\nu^2.$$

The remainder of the change in $\hat{\sigma}_k^2$ can be determined by bounding

$$\begin{aligned}|\sum_{ij}\tilde{G}_{ij}-\sum_{ij}\tilde{F}_{ij}| &\le \sum_{ij}|\tilde{G}_{ij}-\tilde{F}_{ij}| = \sum_{j}|\tilde{G}_{1j}-\tilde{F}_{1j}|+\sum_{i>1}|\tilde{G}_{i1}-\tilde{F}_{i1}| \\ &\le n(8\nu)+(n-1)(8\nu) = (8\nu)(2n-1),\end{aligned}$$

which then gives us

$$\begin{aligned}&\left|\left(\frac{1}{n^2}\sum_{ij}\tilde{G}_{ij}\right)^2-\left(\frac{1}{n^2}\sum_{ij}\tilde{F}_{ij}\right)^2\right| \\ &= \left|\frac{1}{n^2}\sum_{ij}\tilde{G}_{ij}+\frac{1}{n^2}\sum_{ij}\tilde{F}_{ij}\right|\cdot\left|\frac{1}{n^2}\sum_{ij}\tilde{G}_{ij}-\frac{1}{n^2}\sum_{ij}\tilde{F}_{ij}\right| \\ &\le (2\cdot 4\nu)\,\frac{2n-1}{n^2}\,(8\nu) = 64\nu^2\left(\frac{2}{n}-\frac{1}{n^2}\right).\end{aligned}$$

Thus

$$\begin{aligned}|\hat{\sigma}_k^2-(\hat{\sigma}_k')^2| &\le 4\left[\left(\frac{6}{n}-\frac{7}{n^2}+\frac{3}{n^3}\right)16\nu^2+\left(\frac{2}{n}-\frac{1}{n^2}\right)64\nu^2\right] \\ &= \frac{64\nu^2}{n^3}\left[14n^2-11n+3\right] \le \frac{896\nu^2}{n}.\end{aligned}$$

Because the same holds for changing any of the (X_i, Y_i) pairs, the result follows by McDiarmid's inequality. □

Lemma 6.3. *For any kernel k bounded by ν (Assumption (**A**)), the estimator $\hat{\sigma}_k^2$ satisfies*

$$|\mathbb{E}\hat{\sigma}_k^2-\sigma_k^2| \le \frac{1152\nu^2}{n}.$$

Proof. We have that $\mathbb{E}\hat{\sigma}_k^2 = 4\left(\frac{1}{n^3}\sum_{ij\ell}\mathbb{E}\left[\tilde{G}_{i\ell}\tilde{G}_{j\ell}\right]-\frac{1}{n^4}\sum_{ijab}\mathbb{E}\left[\tilde{G}_{ij}\tilde{G}_{ab}\right]\right)$. Most terms in these sums have their indices distinct; these are the ones that we care about. (We could evaluate the expectations of the other terms exactly, but it would be tedious.) We can thus break down the first term as

$$\begin{aligned}\frac{1}{n^3}\sum_{ij\ell}\mathbb{E}[\tilde{G}_{i\ell}\tilde{G}_{j\ell}] &= \frac{1}{n^3}\sum_{ij\ell:|\{i,j,\ell\}|=3}\mathbb{E}[\tilde{G}_{i\ell}\tilde{G}_{j\ell}]+\frac{1}{n^3}\sum_{ij\ell:|\{i,j,\ell\}|<3}\mathbb{E}[\tilde{G}_{i\ell}\tilde{G}_{j\ell}] \\ &= \frac{n(n-1)(n-2)}{n^3}\mathbb{E}[\tilde{G}_{12}\tilde{G}_{13}]+\left(1-\frac{n(n-1)(n-2)}{n^3}\right)q,\end{aligned}$$

where q is the appropriately-weighted mean of the various $\mathbb{E}[\tilde{G}_{i\ell}\tilde{G}_{j\ell}]$ terms for which i, j, ℓ are not mutually distinct. Since $|\tilde{G}_{ij}| \le 4\nu$, $\mathbb{E}[\tilde{G}_{i\ell}\tilde{G}_{j\ell}] < 16\nu^2$ and

so $|q| \leq 16\nu^2$ as well. Noting that

$$\frac{n(n-1)(n-2)}{n^3} = 1 - \frac{3}{n} + \frac{2}{n^2}$$

we obtain

$$|\frac{1}{n^3}\sum_{ij\ell}\mathbb{E}[\tilde{G}_{i\ell}\tilde{G}_{j\ell}] - \mathbb{E}[\tilde{G}_{12}\tilde{G}_{13}]| = \left(\frac{3}{n} - \frac{2}{n^2}\right)|-\mathbb{E}[\tilde{G}_{12}\tilde{G}_{13}]+q| \leq \left(\frac{3}{n} - \frac{2}{n^2}\right)32\nu^2. \tag{6.9}$$

The second term can be handled similarly:

$$\begin{aligned}\frac{1}{n^4}\sum_{ijab}\mathbb{E}[\tilde{G}_{ij}\tilde{G}_{ab}] &= \frac{1}{n^4}\sum_{ijab:|\{i,j,a,b\}|=4}\mathbb{E}[\tilde{G}_{ij}\tilde{G}_{ab}] + \frac{1}{n^4}\sum_{ijab:|\{i,j,a,b\}|<4}\mathbb{E}[\tilde{G}_{ij}\tilde{G}_{ab}] \\ &= \frac{n(n-1)(n-2)(n-3)}{n^4}\mathbb{E}[\tilde{G}_{ij}\tilde{G}_{ab}] \\ &\quad + \left(1 - \frac{n(n-1)(n-2)(n-3)}{n^4}\right)r,\end{aligned}$$

where r is the appropriately-weighted mean of the non-distinct terms, $|r| \leq 16\nu^2$. For i, j, a, b all distinct, $\mathbb{E}[\tilde{G}_{ij}\tilde{G}_{ab}] = \mathbb{E}[\tilde{G}_{12}]^2$. Here

$$\frac{n(n-1)(n-2)(n-3)}{n^4} = \frac{(n-1)(n^2-5n+6)}{n^3} = 1 - \frac{6}{n} + \frac{11}{n} - \frac{6}{n^3}$$

and so

$$|\frac{1}{n^4}\sum_{ijab}\mathbb{E}[\tilde{G}_{ij}\tilde{G}_{ab}] - \mathbb{E}[\tilde{G}_{12}]^2| \leq \left(\frac{6}{n} - \frac{11}{n^2} + \frac{6}{n^3}\right)32\nu^2. \tag{6.10}$$

Recalling $\sigma_k^2 = 4(\mathbb{E}[\tilde{G}_{12}\tilde{G}_{13}] - \mathbb{E}[\tilde{G}_{12}]^2)$,

$$|\mathbb{E}\hat{\sigma}_k^2 - \sigma_k^2| \leq 128\nu^2\left(\frac{9}{n} - \frac{13}{n^2} + \frac{6}{n^3}\right),$$

and since $n \geq 1$, we have $13/n^2 > 6/n^3$, yielding the result. □

Lemma 6.4. *Under Assumptions (**A**) and (**C**), we have*

$$\sup_{\omega,\omega'\in\Omega}\frac{|\hat{\sigma}_\omega^2 - \hat{\sigma}_{\omega'}^2|}{\|\omega-\omega'\|} \leq 256L_k\nu \qquad \textit{and} \qquad \sup_{\omega,\omega'\in\Omega}\frac{|\sigma_\omega^2 - \sigma_{\omega'}^2|}{\|\omega-\omega'\|} \leq 256L_k\nu.$$

Proof. We first handle the change in $\hat{\sigma}_k$:

$$\begin{aligned}|\hat{\sigma}^2_{k_\omega} - \hat{\sigma}^2_{k_{\omega'}}| &= 4\Big|\frac{1}{n^3}\sum_{ij\ell}\tilde{G}^{(\omega)}_{i\ell}\tilde{G}^{(\omega)}_{j\ell} - \frac{1}{n^3}\sum_{ij\ell}\tilde{G}^{(\omega')}_{i\ell}\tilde{G}^{(\omega')}_{j\ell} \\ &\quad - \frac{1}{n^4}\sum_{ijab}\tilde{G}^{(\omega)}_{ij}\tilde{G}^{(\omega)}_{ab} + \frac{1}{n^4}\sum_{ijab}\tilde{G}^{(\omega')}_{ij}\tilde{G}^{(\omega')}_{ab}\Big| \\ &\leq \frac{4}{n^3}\sum_{ij\ell}|\tilde{G}^{(\omega)}_{i\ell}\tilde{G}^{(\omega)}_{j\ell} - \tilde{G}^{(\omega')}_{i\ell}\tilde{G}^{(\omega')}_{j\ell}| + \frac{4}{n^4}\sum_{ijab}|\tilde{G}^{(\omega)}_{ij}\tilde{G}^{(\omega)}_{ab} - \tilde{G}^{(\omega')}_{ij}\tilde{G}^{(\omega')}_{ab}|.\end{aligned}$$

We can handle both terms by bounding

$$\begin{aligned}|\tilde{G}^{(\omega)}_{ij}\tilde{G}^{(\omega)}_{ab} - \tilde{G}^{(\omega')}_{ij}\tilde{G}^{(\omega')}_{ab}| &\leq |\tilde{G}^{(\omega)}_{ij}\tilde{G}^{(\omega)}_{ab} - \tilde{G}^{(\omega)}_{ij}\tilde{G}^{(\omega')}_{ab}| + |\tilde{G}^{(\omega)}_{ij}\tilde{G}^{(\omega')}_{ab} - \tilde{G}^{(\omega')}_{ij}\tilde{G}^{(\omega')}_{ab}| \\ &= |\tilde{G}^{(\omega)}_{ij}||\tilde{G}^{(\omega)}_{ab} - \tilde{G}^{(\omega')}_{ab}| + |\tilde{G}^{(\omega)}_{ij} - \tilde{G}^{(\omega')}_{ij}||\tilde{G}^{(\omega')}_{ab}| \\ &\leq 4\nu\left(|\tilde{G}^{(\omega)}_{ab} - \tilde{G}^{(\omega')}_{ab}| + |\tilde{G}^{(\omega)}_{ij} - \tilde{G}^{(\omega')}_{ij}|\right).\end{aligned}$$

Using Assumption (**C**) and the definition of H,

$$|\tilde{G}^{(\omega)}_{ij} - \tilde{G}^{(\omega')}_{ij}| \leq 4L_k\|\omega - \omega'\|$$

so

$$|\tilde{G}^{(\omega)}_{ij}\tilde{G}^{(\omega)}_{ab} - \tilde{G}^{(\omega')}_{ij}\tilde{G}^{(\omega')}_{ab}| \leq 32\nu L_k\|\omega - \omega'\| \tag{6.11}$$

and hence

$$|\hat{\sigma}^2_{\omega} - \hat{\sigma}^2_{\omega'}| \leq 256\nu L_k\|\omega - \omega'\|.$$

Again using Eq. (6.11), we also have

$$\begin{aligned}|\sigma^2_{\omega} - \sigma^2_{\omega'}| &\leq 4|\mathbb{E}\left[\tilde{G}^{(\omega)}_{12}\tilde{G}^{(\omega)}_{13}\right] - \mathbb{E}\left[\tilde{G}^{(\omega')}_{12}\tilde{G}^{(\omega')}_{13}\right]| + 4|\mathbb{E}\left[\tilde{G}^{(\omega)}_{12}\right]^2 - \mathbb{E}\left[\tilde{G}^{(\omega')}_{12}\right]^2| \\ &\leq 4\mathbb{E}|\tilde{G}^{(\omega)}_{12}\tilde{G}^{(\omega)}_{13} - \tilde{G}^{(\omega')}_{12}\tilde{G}^{(\omega')}_{13}| + 4\mathbb{E}|\tilde{G}^{(\omega)}_{12}\tilde{G}^{(\omega)}_{34} - \tilde{G}^{(\omega')}_{12}\tilde{G}^{(\omega')}_{34}| \\ &\leq 256\nu L_k\|\omega - \omega'\|.\end{aligned}$$

□

6.A.4 Constructing appropriate kernels

We now show Propositions 6.3 and 6.4, which each state that Assumption (**C**) is satisfied by various choices of kernel. The following assumption will be useful for different kernel schemes.

(**I**) The domain $\mathcal{X}$ is Euclidean and bounded, $\mathcal{X} \subseteq \left\{x \in \mathbb{R}^d : \|x\| \leq R_X\right\}$ for some constant $R_X < \infty$.

We begin by recalling a well-known property of the Gaussian kernel, useful for both Gaussian bandwidth selection and deep kernels. A proof is in Section 6.A.5.

Lemma 6.5. *The Gaussian kernel* $\kappa(a,b) = \exp\left(-\frac{\|a-b\|^2}{2\sigma^2}\right)$ *satisfies*

$$|\kappa(a,b) - \kappa(a',b')| \leq \frac{1}{\sigma\sqrt{e}}\left(\|a-b\| + \|a'-b'\|\right) \leq \frac{1}{\sigma\sqrt{e}}\left(\|a-a'\| + \|b-b'\|\right).$$

6.A.4.1 Gaussian bandwidth selection (Proposition 6.3)

Lemma 6.5 immediately gives us Assumption (**C**) when we use Gaussian kernels:

Proposition 6.7. *Define a one-dimensional Banach space for inverse length-scales of Gaussian kernels* $\gamma > 0$*, so that* $k_\gamma(x,y) = \kappa_{1/\gamma}(x,y)$*, with standard addition and multiplication and norms defined by the absolute value, and* k_0 *taken to be the constant* 1 *function. Let* Ω *be any subset of this space. Under Assumption (**I**), Assumption (**C**) holds: for any* $x, y \in \mathcal{X}$ *and* $\gamma, \gamma' \in \Gamma$*,*

$$|k_\gamma(x,y) - k_{\gamma'}(x,y)| \leq \frac{2R_X}{\sqrt{e}}|\gamma - \gamma'|.$$

Proof.

$$\begin{aligned}|k_\gamma(x,y) - k_{\gamma'}(x,y)| &= |\kappa_1(\gamma x, \gamma y) - \kappa_1(\gamma' x, \gamma' y)| \\ &\leq \frac{1}{\sqrt{e}}|\gamma\|x-y\| - \gamma'\|x-y\|| = \frac{\|x-y\|}{\sqrt{e}}|\gamma - \gamma'|.\end{aligned}$$

□

6.A.4.2 Deep kernels (Proposition 6.4)

To handle the deep kernel case, we will need some more assumptions on the form of the kernel.

(**II**) $\phi_\omega(x) = \phi_\omega^{(\Lambda)}$ is a feedforward neural network with Λ layers given by

$$\phi_\omega^{(0)}(x) = x \qquad \phi_\omega^{(\ell)}(x) = \sigma^{(\ell)}\left(W_\omega^{(\ell)}\phi_\omega^{(\ell-1)}(x) + b_\omega^{(\ell)}\right),$$

where the network parameter ω consists of all the weight matrices $W_\omega^{(\ell)}$ and biases $b_\omega^{(\ell)}$, and the activation functions $\sigma^{(\ell)}$ are each 1-Lipschitz, $\|\sigma^{(\ell)}(x) - \sigma^{(\ell)}(y)\| \leq \|x-y\|$, with $\sigma^{(\ell)}(0) = 0$ so that $\|\sigma^{(\ell)}(x)\| \leq \|x\|$. Define a Banach space on ω, with addition and scalar multiplication component-wise, and

$$\|\omega\| = \max_{\ell \in \{1,\ldots,\Lambda\}} \max\left(\|W_\omega^{(\ell)}\|, \|b_\omega^{(\ell)}\|\right),$$

where the matrix norm denotes operator norm $\|W\| = \sup_x \|Wx\|/\|x\|$. (For convolutional networks, see Remark 6.2.)

(III) k_ω is a kernel of the form Eq. (6.1),

$$k_\omega(x, y) = [(1 - \epsilon)\kappa(\phi_\omega(x), \phi_\omega(y)) + \epsilon]\, q(x, y),$$

with $0 \leq \epsilon \leq 1$, κ a kernel function, and $q(x, y)$ a kernel with $\sup_x q(x, x) \leq Q$.

Note that this includes kernels of the form $k_\omega(x, y) = \kappa(\phi_\omega(x), \phi_\omega(y))$: take $\epsilon = 0$ and $q(x, y) = 1$.

(IV) κ in Assumption (**III**) is a kernel function satisfying

$$|\kappa(a, b) - \kappa(a', b')| \leq L_\kappa \left(\|a - a'\| + \|b - b'\|\right).$$

This holds for a Gaussian κ via Lemma 6.5.

We now turn to proving Assumption (**C**) for deep kernels. First, we will need some smoothness properties of the network ϕ.

Lemma 6.6. *Under Assumption (**II**), suppose ω, ω' have $\|\omega\| \leq R$, $\|\omega'\| \leq R$, with $R \neq 1$. Then, for any x,*

$$\|\phi_\omega(x)\| \leq R^\Lambda \|x\| + \frac{R}{R-1}(R^\Lambda - 1) \tag{6.12}$$

$$\|\phi_\omega(x) - \phi_{\omega'}(x)\| \leq \left(\Lambda R^{\Lambda-1}\left(\|x\| + \frac{R}{R-1}\right) - \frac{R^\Lambda - 1}{(R-1)^2}\right)\|\omega - \omega'\|. \tag{6.13}$$

If $R \geq 2$, we furthermore have

$$\|\phi_\omega(x)\| \leq R^\Lambda(\|x\| + 2) \tag{6.14}$$

$$\|\phi_\omega(x) - \phi_{\omega'}(x)\| \leq \Lambda R^{\Lambda-1}\left(\|x\| + 2\right)\|\omega - \omega'\|. \tag{6.15}$$

The proof, by recursion, is given in Section 6.A.5. We are now ready to prove Assumption (**C**) for deep kernels.

Proposition 6.8. *Make Assumptions (**I**) to (**IV**) and Assumption (**B**), with $R_\Omega \geq 2$.*[8]

[8]Of course, if we know a bound of $R_\Omega < 2$, the result will still hold using $R_\Omega = 2$. It is also possible to show a tighter result, via Eq. (6.12) and Eq. (6.13) or their analogue for $R = 1$; the expression is simply less compact.

*Then Assumption (**C**) holds: for any $x, y \in \mathcal{X}$ and $\omega, \omega' \in \Omega$,*

$$|k_\omega(x,y) - k_{\omega'}(x,y)| \leq 2Q(1-\epsilon)L_\kappa \Lambda R_\Omega^{\Lambda-1}(R_X+2)\|\omega - \omega'\|.$$

Proof.

$$\begin{aligned}
|k_\omega(x,y) - k_{\omega'}(x,y)| &= (1-\epsilon)|\kappa(\phi_\omega(x),\phi_\omega(y)) - \kappa(\phi_{\omega'}(x),\phi_{\omega'}(y))|q(x,y) \\
&\leq Q(1-\epsilon)L_\kappa\left(|\phi_\omega(x) - \phi_{\omega'}(x)| + |\phi_\omega(y) - \phi_{\omega'}(y)|\right) \\
&\leq Q(1-\epsilon)L_\kappa \Lambda R_\Omega^{\Lambda-1}(\|x\| + \|y\| + 4)\|\omega - \omega'\| \\
&\leq Q(1-\epsilon)L_\kappa \Lambda R_\Omega^{\Lambda-1}(2R_X + 4)\|\omega - \omega'\|. \qquad \square
\end{aligned}$$

Remark 6.1. *For the deep kernels we use in the paper (Assumptions (**II**) to (**IV**)) on bounded domains (Assumption (**I**)), we know L_k via Proposition 6.8; Theorem 6.3 combines Theorem 6.4, Corollary 6.1, and Proposition 6.8. If we further use a Gaussian kernel q of bandwidth σ_ϕ, the last bracketed term in the error bound of Theorem 6.4 becomes*

$$\frac{2(1-\epsilon)}{\sigma_\phi\sqrt{e}}\Lambda R_\Omega^{\Lambda-1}(R_X+2) + \sqrt{2\log\frac{2}{\delta} + 2D\log\left(4R_\Omega\sqrt{n}\right)}.$$

The component $R_\Omega^{\Lambda-1}(R_X+2)$, from Eq. (6.14)*, is approximately the largest that ϕ_ω could make its outputs' norms; σ_ϕ will generally be on a comparable scale to the norm of the actual outputs of the network, so their ratio is something like the "unused capacity" of the network to blow up its inputs. This term is weighted about equally in the convergence bound with the square root of the total number of parameters in the network.*

Remark 6.2. *We can handle convolutional networks as follows. We define Ω in essentially the same way, letting $W_\omega^{(\ell)}$ denote the convolutional kernel (the set of parameters being optimised), but define $\|\omega\|$ in terms of the operator norm of the linear transform corresponding to the convolution operator. This is given in terms of the operator norm of various discrete Fourier transforms of the kernel matrix by Lemma 2 of Bibi et al. [2019]; see also Theorem 6 of Sedghi et al. [2019]. The number of parameters D is then the actual number of parameters optimised in gradient descent, but the radius R_Ω is computed differently.*

6.A.5 Miscellaneous Proofs

The following lemma was used for Propositions 6.7 and 6.8.

Lemma 6.5. *The Gaussian kernel* $\kappa(a,b) = \exp\left(-\frac{\|a-b\|^2}{2\sigma^2}\right)$ *satisfies*

$$|\kappa(a,b) - \kappa(a',b')| \leq \frac{1}{\sigma\sqrt{e}}\left(\|a-b\| + \|a'-b'\|\right) \leq \frac{1}{\sigma\sqrt{e}}\left(\|a-a'\| + \|b-b'\|\right).$$

Proof. We have that

$$\begin{aligned}|\kappa(a,b) - \kappa(a',b')| &= \left|\exp\left(-\frac{\|a-b\|^2}{2\sigma^2}\right) - \exp\left(-\frac{\|a'-b'\|^2}{2\sigma^2}\right)\right| \\ &\leq \|x \mapsto \exp\left(-\frac{x^2}{2\sigma^2}\right)\|_L \cdot \left|\|a-b\| - \|a'-b'\|\right|.\end{aligned}$$

We can bound the Lipschitz constant as its maximal derivative norm,

$$\sup_x \frac{|x|}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right).$$

Noting that

$$\frac{\mathrm{d}}{\mathrm{d}x}\log\left(\frac{|x|}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right)\right) = \frac{1}{x} - \frac{x}{\sigma^2}$$

vanishes only at $x = \pm\sigma$, the supremum is achieved by using that value, giving

$$\|x \mapsto \exp\left(-\frac{x^2}{2\sigma^2}\right)\|_L = \frac{1}{\sigma\sqrt{e}}.$$

The result follows from

$$\left|\|a-b\| - \|a'-b'\|\right| \leq \|a-b-a'+b'\| \leq \|a-a'\| + \|b-b'\|. \qquad \square$$

This next lemma was used in Proposition 6.8.

Lemma 6.6. *Under Assumption (**II**), suppose* ω, ω' *have* $\|\omega\| \leq R$, $\|\omega'\| \leq R$, *with* $R \neq 1$. *Then, for any x,*

$$\|\phi_\omega(x)\| \leq R^\Lambda\|x\| + \frac{R}{R-1}(R^\Lambda - 1) \tag{6.12}$$

$$\|\phi_\omega(x) - \phi_{\omega'}(x)\| \leq \left(\Lambda R^{\Lambda-1}\left(\|x\| + \frac{R}{R-1}\right) - \frac{R^\Lambda - 1}{(R-1)^2}\right)\|\omega - \omega'\|. \tag{6.13}$$

If $R \geq 2$, *we furthermore have*

$$\|\phi_\omega(x)\| \leq R^\Lambda(\|x\| + 2) \tag{6.14}$$

$$\|\phi_\omega(x) - \phi_{\omega'}(x)\| \leq \Lambda R^{\Lambda-1}\left(\|x\| + 2\right)\|\omega - \omega'\|. \tag{6.15}$$

Proof. First, $\|\phi_\omega^{(0)}(x)\| = \|x\|$, showing Eq. (6.12) when $\Lambda = 0$. In general,

$$\begin{aligned}\|\phi_\omega^{(\ell)}(x)\| &= \|\sigma^{(\ell)}\left(W_\omega^{(\ell)}\phi_\omega^{(\ell-1)}(x) + b_\omega^{(\ell)}\right)\| \\ &\leq \|W_\omega^{(\ell)}\phi_\omega^{(\ell-1)}(x) + b_\omega^{(\ell)}\| \\ &\leq \|W_\omega^{(\ell)}\|\|\phi_\omega^{(\ell-1)}(x)\| + \|b_\omega^{(\ell)}\| \\ &\leq R\|\phi_\omega^{(\ell-1)}(x)\| + R,\end{aligned}$$

and expanding this recursion gives

$$\|\phi_\omega^{(\ell)}(x)\| \leq R^\ell\|x\| + \sum_{m=1}^{\ell} R^m = R^\ell\|x\| + \frac{R}{R-1}(R^\ell - 1).$$

Now, we have Eq. (6.13) for $\Lambda = 0$ because $\phi_\omega^{(0)}(x) - \phi_{\omega'}^{(0)}(x) = 0$. For $\ell \geq 1$,

$$\begin{aligned}&\|\phi_\omega^{(\ell)}(x) - \phi_{\omega'}^{(\ell)}(x)\| = \|\sigma^{(\ell)}\left(W_\omega^{(\ell)}\phi_\omega^{(\ell-1)}(x) + b_\omega^{(\ell)}\right) - \sigma^{(\ell)}\left(W_{\omega'}^{(\ell)}\phi_{\omega'}^{(\ell-1)}(x) - b_{\omega'}^{(\ell)}\right)\| \\ &\leq \|W_\omega^{(\ell)}\phi_\omega^{(\ell-1)}(x) - W_{\omega'}^{(\ell)}\phi_\omega^{(\ell-1)}(x)\| + \|W_{\omega'}^{(\ell)}\phi_\omega^{(\ell-1)}(x) - W_{\omega'}^{(\ell)}\phi_{\omega'}^{(\ell-1)}(x)\| + \|b_\omega^{(\ell)} - b_{\omega'}^{(\ell)}\| \\ &\leq \|W_\omega^{(\ell)} - W_{\omega'}^{(\ell)}\|\|\phi_\omega^{(\ell-1)}(x)\| + \|W_{\omega'}^{(\ell)}\|\|\phi_\omega^{(\ell-1)}(x) - \phi_{\omega'}^{(\ell-1)}(x)\| + \|\omega - \omega'\| \\ &\leq \|\omega - \omega'\|\left(R^{\ell-1}\|x\| + \frac{R}{R-1}(R^{\ell-1} - 1) + 1\right) + R\|\phi_\omega^{(\ell-1)}(x) - \phi_{\omega'}^{(\ell-1)}(x)\|.\end{aligned}$$

Expanding the recursion yields

$$\begin{aligned}\|\phi_\omega^{(\ell)}(x) - \phi_{\omega'}^{(\ell)}(x)\| &\leq \sum_{m=0}^{\ell-1} R^m\left(R^{\ell-1-m}\|x\| + \frac{R}{R-1}(R^{\ell-m-1} - 1) + 1\right)\|\omega - \omega'\| \\ &= \sum_{m=0}^{\ell-1}\left(R^{\ell-1}\|x\| + \frac{R^\ell}{R-1} - \frac{R^{m+1}}{R-1} + R^m\right)\|\omega - \omega'\| \\ &= \left(\ell R^{\ell-1}\|x\| + \frac{\ell R^\ell}{R-1} - \left(\frac{R}{R-1} - 1\right)\sum_{m=0}^{\ell-1} R^m\right)\|\omega - \omega'\| \\ &= \left(\ell R^{\ell-1}\left(\|x\| + \frac{R}{R-1}\right) - \frac{1}{R-1}\frac{R^\ell - 1}{R-1}\right)\|\omega - \omega'\|.\end{aligned}$$

When $R \geq 2$, we have that $R/(R-1) \leq 2$ and $R^\ell > 1$, giving Eq. (6.14) and Eq. (6.15). □

6.B Experimental Details

6.B.1 Details of Synthetic Datasets

Table 6.5 shows details of four synthetic datasets. *Blob* datasets are often used to validate two-sample test methods [Gretton et al., 2012a; Jitkrittum et al., 2016a; Sutherland et al., 2016], although we rotate each blob to show the benefits of non-homogeneous kernels. *HDGM* datasets can be regarded as *high-dimension Blob* which contains two modes with the same variance and different covariance.

Datasets	P	Q
Blob-S	$\sum_{i=1}^{9} \frac{1}{9}\mathcal{N}(\mu_i^b, 0.03 \times I_2)$	$\sum_{i=1}^{9} \frac{1}{9}\mathcal{N}(\mu_i^b, 0.03 \times I_2)$
Blob-D	$\sum_{i=1}^{9} \frac{1}{9}\mathcal{N}(\mu_i^b, 0.03 \times I_2)$	$\sum_{i=1}^{9} \frac{1}{9}\mathcal{N}\left(\mu_i^b, \begin{bmatrix} 0.03 & \Delta_i^b \\ \Delta_i^b & 0.03 \end{bmatrix}\right)$
HDGM-S	$\sum_{i=1}^{2} \frac{1}{2}\mathcal{N}(\mu_i^h, I_d)$	$\sum_{i=1}^{2} \frac{1}{2}\mathcal{N}(\mu_i^h, I_d)$
HDGM-D	$\sum_{i=1}^{2} \frac{1}{2}\mathcal{N}(\mu_i^h, I_d)$	$\sum_{i=1}^{2} \frac{1}{2}\mathcal{N}\left(\mu_i^h, \begin{bmatrix} 1 & \Delta_i^h & \mathbf{0}_{d-2} \\ \Delta_i^h & 1 & \mathbf{0}_{d-2} \\ \mathbf{0}_{d-2}^T & \mathbf{0}_{d-2}^T & I_{d-2} \end{bmatrix}\right)$

Table 6.5: Specifications of P and Q of synthetic datasets. $\mu_1^b = [0, 0], \mu_2^b = [0, 1], \mu_3^b = [0, 2], \ldots, \mu_8^b = [2, 1], \mu_9^b = [2, 2]$, the same with Figure 6.1(a). $\mu_1^h = \mathbf{0}_d$, $\mu_2^h = 0.5 \times \mathbf{1}_d$, I_d is an identity matrix with size d. $\Delta_i^b = -0.02 - 0.002 \times (i-1)$ if $i < 5$ and $\Delta_i^b = 0.02 + 0.002 \times (i-6)$ if $i > 5$. if $i = 5$, $\Delta_i^b = 0$ (same with Figure 6.1a). Δ_1^h and Δ_2^h are set to 0.5 and -0.5, respectively.

6.B.2 Real Datasets and Visualizations

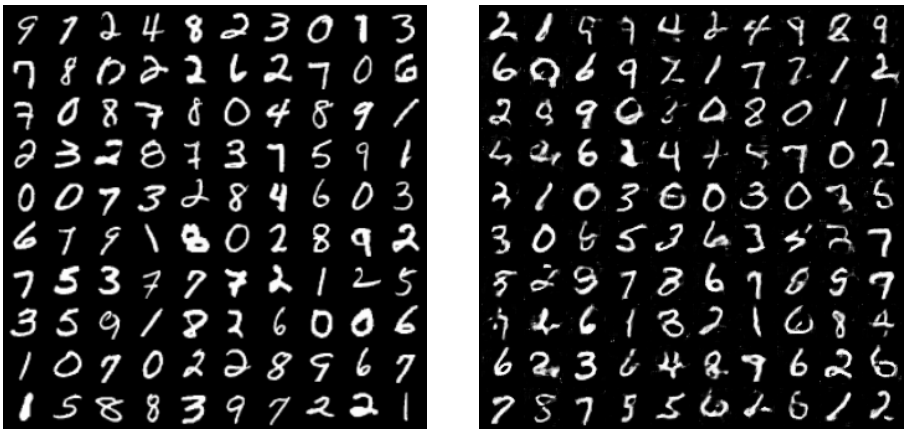
Higgs dataset can be downloaded from UCI Machine Learning Repository [9]. *MNIST* dataset can be downloaded via Pytorch [10]. Figure 6.4 shows images from two sets of *MNIST* digits: the Real-*MNIST* and "Fake"-*MNIST*.

6.B.3 Type-I errors on *Higgs* and *MNIST*

Table 6.6 shows average Type-I error on *Higgs* dataset when increasing number of samples (n). Table 6.7 shows average Type-I error on Real-*MNIST* vs. Real-*MNIST* when increasing number of samples (n).

[9] https://archive.ics.uci.edu/ml/datasets/HIGGS

[10] https://github.com/eriklindernoren/PyTorch-GAN/blob/master/implementations/dcgan/dcgan.py



(a) Real-*MNIST* (b) "Fake"-*MNIST*

Figure 6.4: Images from Real-*MNIST* [LeCun et al., 1998] and "Fake"-*MNIST* generated from DCGAN [Radford et al., 2016].

n	ME	SCF	C2ST-S	C2ST-L	MMD-O	MMD-D
1000	0.048	0.040	0.043	0.048	0.059	0.037
2000	0.043	0.032	0.060	0.056	0.055	0.053
3000	0.049	0.043	0.046	0.053	0.051	0.069
5000	0.056	0.035	0.052	0.065	0.049	0.062
8000	0.050	0.034	0.065	0.067	0.056	0.037
10000	0.059	0.032	0.057	0.058	0.045	0.048
Avg.	0.051	0.036	0.054	0.058	0.050	0.051

Table 6.6: Results on *Higgs* ($\alpha = 0.05$). We report average Type-I error on *Higgs* dataset when increasing number of samples (N). Note that, in *Higgs*, we have two types of Type-I errors: 1) Type-I error when two samples drawn from P (no Higgs bosons) and 2) Type-I error when two samples drawn from Q (having Higgs bosons). Type-I reported here is the average value of 1) and 2). Since Type-I error reported here is the average value of two average Type-I errors, we do not report standard errors of the average Type-I error in this table.

n	ME	SCF	C2ST-S	C2ST-L	MMD-O	MMD-D
200	$0.076_{\pm 0.011}$	$0.075_{\pm 0.010}$	$0.035_{\pm 0.006}$	$0.045_{\pm 0.005}$	$0.068_{\pm 0.004}$	$0.056_{\pm 0.003}$
400	$0.062_{\pm 0.010}$	$0.056_{\pm 0.007}$	$0.044_{\pm 0.006}$	$0.040_{\pm 0.004}$	$0.053_{\pm 0.005}$	$0.056_{\pm 0.005}$
600	$0.051_{\pm 0.003}$	$0.049_{\pm 0.009}$	$0.039_{\pm 0.005}$	$0.054_{\pm 0.007}$	$0.066_{\pm 0.008}$	$0.056_{\pm 0.008}$
800	$0.054_{\pm 0.006}$	$0.046_{\pm 0.006}$	$0.043_{\pm 0.005}$	$0.042_{\pm 0.007}$	$0.051_{\pm 0.005}$	$0.054_{\pm 0.007}$
1000	$0.047_{\pm 0.006}$	$0.045_{\pm 0.010}$	$0.038_{\pm 0.006}$	$0.046_{\pm 0.005}$	$0.041_{\pm 0.007}$	$0.062_{\pm 0.006}$
Avg.	0.058	0.054	0.040	0.045	0.056	0.057

Table 6.7: Results on *MNIST* given $\alpha = 0.05$. We report average Type-I error$\pm$standard errors on Real-*MNIST* vs. Real-*MNIST* when increasing number of samples (N).

Chapter 7

Conclusions and Future Directions

In this thesis, we investigate the non-parametric hypothesis testing problems including goodness-of-fit tests, two-sample tests and quasi-independence tests based on kernel methods. In the setting of testing goodness-of-fit, the test statistics are developed from relevant kernelised Stein discrepancies for practical data scenarios such as Riemannian manifold data (Chapter 3) and data with censoring (Chapter 4). We analyse and compare the advantages and disadvantages of the effect of different Stein operators in performing the KSD-based tests for goodness-of-fit. In the non-Euclidean data scenario, we compare the Stein discrepancies from different differential orders with respect to the test functions; and in the censored data scenario, we compare the Stein discrepancies based on important functions in survival analysis. For non-Euclidean data, we perform model criticism based on the optimised test locations adapted from Finite Set Stein Discrepancy (FSSD), to extract the interpretable information when the proposed model fails to fit the observed data.

In the setting of testing quasi-independence, where the observed data are left-truncated and right censored, our developed test is equivalent to *simultaneously* take infinitely manly weighted log-rank tests by taking supremum over a rich-enough class of unit ball RKHS function. Such a scheme alleviates the sensitivity of the test performances from choosing appropriate weight functions and the proposed test can achieve state-of-the-art performance with high test power and well-controlled Type-I error at high censoring rate and with complicated censoring structures.

In the setting of the two-sample problem, the proposed *translation non-invariant* deep kernel for MMD-based test adaptively learns the distribution features to compare the samples more efficiently with smaller sample size, achieving better test performances compared to existing state-of-the-art kernel-based tests. With the deep kernel architectures that is capable of extracting useful features, complicated data such as images of *MNIST* digits can be tested more efficiently.

Future Work

Generalised Kernel Stein Discrepancy

The KSD-based goodness-of-fit tests developed in Chapter 3 and Chapter 4 rely on identifying appropriate Stein operators for the specific data scenario. However, there is not yet a systematic way to develop new Stein operators for practical scenarios that did not appear in the literature before. As such, a unifying framework can be useful to develop Stein operators and corresponding KSD-based tests or learning scheme. Eq. (2.22) provides a Langevin-type diffusion Stein operator. For fixed function g, consider the construction of Stein operator $\mathcal{A}_{q,g}f(x) = \mathcal{T}_q(f(x)g(x)) = g(x)(f'(x) + f(x)\log q(x)') + g'(x)f(x)$. The technique to formulate the Stein operator is referred to as standardisation [Anastasiou et al., 2021; Mijoule et al., 2018], which was developed to analyse Gaussian approximation. The corresponding KSD is defined as $\mathrm{KSD}_g(p\|q;\mathcal{H}) = \sup_{f\in B_1(\mathcal{H})} |\mathbb{E}_p[\mathcal{T}_q(f(x)g(x))]|$. Different choice of g may be appropriate in dealing with different scenarios. The Diffusion KSD (DKSD) [Barp et al., 2019] had shed a light on using different diffusion function for construction of Stein discrepancy in the context of density estimation. In addition, as the Stein operators are usually not unique, investigating whether there is a better discrepancy beyond $g \equiv 1$, which is the KSD in Eq. (2.22) can be an interesting direction to improve the proposed testing procedures.

Deep Kernel with Stein Discrepancy

With the success of using *deep kernel* architecture for MMD-based two-sample tests where the non-translation invariant kernels are able to extract distribution features to perform the test more efficiently compared to using the simple kernels, it is an interesting direction to consider such deep kernel architecture on KSD-based test for goodness-of-fit. When the observed data is of small sample size, the goodness-of-fit test using KSD with simple (translation invariant) kernels may see very low power, and much larger sample sizes are required to see the asymptotic trend. Moreover, KSD is known to be less robust when the underlying distribution exhibits multimodal structures. Learning deep kernel architectures aim to extract more effective features to compare distributions and improve the test efficiency at the scheme of lower sample sizes. In addition, as a discrepancy measure between distributions, another interesting direction to investigate is using deep kernel KSD for training generative models.

Censoring in Higher Dimensions

In Chapter 4 and 5, the hypothesis testing methods have been developed for univariate data, where the censoring is naturally defined in $\mathbb{R}^+$. Despite such univariate case of time-to-event data sees a wide range of applications such as clinical data, e-commerce data, insurance data, etc., the method developed is only useful for one dimensional setting. It will be very interesting to explore model validation or test of independence with censoring in higher dimensional space with the presence of censoring. As the natural ordering is not straight-forward to define, the development and analysis need to be treated carefully w.r.t. to the corresponding definition of "censoring" notion. The generalisation of quantile notion to spatial quantile [Chakraborty et al., 2014] can be potentially a useful tool to incorporate censoring notion in higher dimensions.

Bibliography

Aalen, O., Borgan, O., and Gjessing, H. (2008). *Survival and event history analysis: a process point of view*. Springer Science & Business Media.

Akritas, M. G. (1988). Pearson-type goodness-of-fit tests: the univariate case. *Journal of the American Statistical Association*, 83(401):222–230.

Allen, A. M., Therneau, T. M., Larson, J. J., Coward, A., Somers, V. K., and Kamath, P. S. (2018). Nonalcoholic fatty liver disease incidence and impact on metabolic burden and death: a 20 year-community study. *Hepatology*, 67(5):1726–1736.

Anastasiou, A., Barp, A., Briol, F.-X., Ebner, B., Gaunt, R. E., Ghaderinezhad, F., Gorham, J., Gretton, A., Ley, C., Liu, Q., et al. (2021). Stein's method meets statistics: A review of some recent developments. *arXiv preprint arXiv:2105.03481*.

Andersen, P., Geskus, R., de Witte, T., and Putter, H. (2012). Competing risks in epidemiology: possibilities and pitfalls. *International Journal of Epidemiology*, 41(3):861–870.

Arbel, M., Sutherland, D. J., Bińkowski, M., and Gretton, A. (2018). On gradient regularizers for mmd gans. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 6701–6711.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404.

Bahadur, R. R. et al. (1960). Stochastic comparison of tests. *Annals of Mathematical Statistics*, 31(2):276–295.

Baldi, P., Sadowski, P., and Whiteson, D. (2014). Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5:4308.

Banerjee, A., Dhillon, I. S., Ghosh, J., and Sra, S. (2005). Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382.

Barbour, A. D. and Chen, L. H. Y. (2005). *An introduction to Stein's method*, volume 4. World Scientific.

Barbour, A. D., Holst, L., and Janson, S. (1992). *Poisson approximation*, volume 2. The Clarendon Press Oxford University Press.

Barp, A., Briol, F.-X., Duncan, A., Girolami, M., and Mackey, L. (2019). Minimum stein discrepancy estimators. In *Advances in Neural Information Processing Systems*, pages 12964–12976.

Barp, A., Oates, C., Porcu, E., and Girolami, M. (2018). A riemannian-stein kernel method. *arXiv preprint arXiv:1810.04946*.

Berlinet, A. and Thomas, C. (2004). *Reproducing kernel Hilbert spaces in Probability and Statistics*. Kluwer Academic Publishers.

Bibi, A., Ghanem, B., Koltun, V., and Ranftl, R. (2019). Deep layers as stochastic solvers. In *ICLR*.

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. (2018). Demystifying mmd gans. In *International Conference on Learning Representations*.

Biswas, A., Datta, S., Fine, J., and Segal, M. (2007). *Statistical advances in the biomedical science*. Wiley Online Library.

Boente, G., Rodriguez, D., and Manteiga, W. G. (2014). Goodness-of-fit test for directional data. *Scandinavian Journal of Statistics*, 41(1):259–275.

Bradburn, M. J., Clark, T. G., Love, S. B., and Altman, D. G. (2003). Survival analysis part ii: multivariate data analysis–an introduction to concepts and methods. *British journal of cancer*, 89(3):431–436.

Brendel, M., Janssen, A., Mayer, C.-D., and Pauly, M. (2014). Weighted logrank permutation tests for randomly right censored life science data. *Scandinavian Journal of Statistics*, 41(3):742–761.

Byrne, S. and Girolami, M. (2013). Geodesic monte carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40(4):825–845.

Callaert, H. and Janssen, P. (1978). The Berry-Esseen theorem for u-statistics. *The Annals of Statistics*, 6(2):417–421.

Carmeli, C., De Vito, E., Toigo, A., and Umanitá, V. (2010). Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61.

Chaieb, L., Rivest, L.-P., and Abdous, B. (2006). Estimating survival under a dependent truncation. *Biometrika*, 93(3):655–669.

Chakraborty, A., Chaudhuri, P., et al. (2014). The spatial distribution in infinite dimensional spaces and related quantiles and depths. *The Annals of Statistics*, 42(3):1203–1231.

Chansky, K., Subotic, D., Foster, N. R., and Blum, T. (2016). Survival analyses in lung cancer. *Journal of thoracic disease*, 8(11):3457.

Chen, L. H. Y., Goldstein, L., and Shao, Q. M. (2010). *Normal approximation by Stein's method*. Springer.

Chen, W. Y., Mackey, L., Gorham, J., Briol, F.-X., and Oates, C. J. (2018). Stein points. *arXiv preprint arXiv:1803.10161*.

Cheng, X. and Cloninger, A. (2019). Classification logit two-sample testing by neural networks. *arXiv preprint arXiv:1909.11298*.

Chikuse, Y. (2003). Concentrated matrix langevin distributions. *Journal of Multivariate Analysis*, 2(85):375–394.

Chikuse, Y. (2012). *Statistics on special manifolds*, volume 174. Springer Science & Business Media.

Chikuse, Y. and Jupp, P. E. (2004). A test of uniformity on shape spaces. *Journal of multivariate analysis*, 88(1):163–176.

Chiou, S., Qian, J., Mormino, E., Betensky, R., Australian, I. B. L. F. S., Harvard, A. B. S., and Initiative, A. D. N. (2018). Permutation tests for general dependent truncation. *Computational Statistics & Data Analysis*, 128:308.

Chung, C.-F., Schmidt, P., and Witte, A. D. (1991). Survival analysis: A survey. *Journal of Quantitative Criminology*, 7(1):59–98.

Chwialkowski, K. and Gretton, A. (2014). A kernel independence test for random processes. In *International Conference on Machine Learning*.

Chwialkowski, K., Sejdinovic, D., and Gretton, A. (2014). A wild bootstrap for degenerate kernel tests. In *Advances in Neural Information Processing Systems*, pages 3608–3616.

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit. In *International Conference on Machine Learning*, pages 2606–2615.

Chwialkowski, K. P., Ramdas, A., Sejdinovic, D., and Gretton, A. (2015). Fast two-sample testing with analytic representations of probability measures. *Advances in Neural Information Processing Systems*, 28:1981–1989.

Collett, D. (2015). *Modelling survival data in medical research*. Chapman and Hall/CRC.

Crowley, J. and Hu, M. (1977). Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association*, 72(357):27–36.

Cucker, F. and Smale, S. (2001). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49.

Dehling, H. and Mikosch, T. (1994). Random quadratic forms and the bootstrap for U-statistics. *Journal of Multivariate Analysis*, 51(2):392–413.

Downs, T. D. (1972). Orientation statistics. *Biometrika*, 59(3):665–676.

Dryden, I. L. and Mardia, K. V. (2016). *Statistical shape analysis: with applications in R*, volume 995. John Wiley & Sons.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, 28(1):181–187.

Edmonson, J. H., Fleming, T. R., Decker, D., Malkasian, G., Jorgensen, E., Jefferies, J., Webb, M., and Kvols, L. (1979). Different chemotherapeutic sensitivities and host factors affecting prognosis in advanced ovarian carcinoma versus minimal residual disease. *Cancer treatment reports*, 63(2):241–247.

Emura, T. and Wang, W. (2010). Testing quasi-independence for truncation data. *Journal of Multivariate Analysis*, 101:223–239.

Eric, M., Bach, F., and Harchaoui, Z. (2007). Testing for homogeneity with kernel fisher discriminant analysis. *Advances in Neural Information Processing Systems*, 20:609–616.

Fallaize, C. J. and Kypraios, T. (2016). Exact bayesian inference for the bingham distribution. *Statistics and Computing*, 26(1-2):349–360.

Fernandez, T. and Gretton, A. (2019). A maximum-mean-discrepancy goodness-of-fit test for censored data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2966–2975.

Fernandez, T., Gretton, A., Rindt, D., and Sejdinovic, D. (2019). A kernel log-rank test of independence for right-censored data. *arXiv preprint arXiv:1912.03784*.

Fernandez, T. and Rivera, N. (2019). A reproducing kernel hilbert space log-rank test for the two-sample problem. *arXiv preprint arXiv:1904.05187*.

Fernández, T. and Rivera, N. (2020). Kaplan-Meier V- and U-statistics. *Electronic Journal of Statistics*, 14(1):1872–1916.

Fernandez, T., Rivera, N., Xu, W., and Gretton, A. (2020). Kernelized stein discrepancy tests of goodness-of-fit for time-to-event data. In *International Conference on Machine Learning*, pages 3112–3122. PMLR.

Fernández, V. A., Gamero, M. J., and Garcia, J. M. (2008). A test for the two-sample problem based on empirical characteristic functions. *Computational statistics & data analysis*, 52(7):3730–3748.

Figueiredo, A. (2007). Comparison of tests of uniformity defined on the hypersphere. *Statistics & probability letters*, 77(3):329–334.

Figueiredo, A. M. S. (2012). Goodness-of-fit for a concentrated von mises-fisher distribution. *Computational Statistics*, 27(1):69–82.

Flanders, H. (1963). *Differential Forms with Applications to the Physical Sciences*. Dover.

Fleming, T. R. and Harrington, D. P. (2011). *Counting processes and survival analysis*, volume 169. John Wiley & Sons.

Fukumizu, K., Bach, F. R., and Gretton, A. (2007a). Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8(2).

Fukumizu, K., Gretton, A., Lanckriet, G., Schölkopf, B., and Sriperumbudur, B. K. (2009). Kernel choice and classifiability for rkhs embeddings of probability distributions. *Advances in Neural Information Processing Systems*, 22:1750–1758.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2007b). Kernel measures of conditional dependence. In *NIPS*, volume 20, pages 489–496.

García-Portugués, E. and Verdebout, T. (2018). An overview of uniformity tests on the hypersphere. *arXiv preprint arXiv:1804.00286*.

Garreau, D., Jitkrittum, W., and Kanagawa, M. (2017). Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*.

Genton, M. and Hering, A. (2007). Blowing in the wind. *Significance*, 4(1):11–14.

Gill, R. (1983). Large sample behaviour of the product-limit estimator on the whole line. *Ann. Statist.*, 11(1):49–58.

Gill, R. D. (1980). *Censoring and stochastic integrals*, volume 124 of *Mathematical Centre Tracts*. Mathematisch Centrum, Amsterdam.

Giné, E. (1975). Invariant tests for uniformity on compact riemannian manifolds based on sobolev norms. *The Annals of statistics*, pages 1243–1266.

Gleser, L. J. (1966). The comparison of multivariate tests of hypothesis by means of bahadur efficiency. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 157–174.

Gneiting, T. et al. (2013). Strictly and non-strictly positive definite functions on spheres. *Bernoulli*, 19(4):1327–1349.

Gorham, J. and Mackey, L. (2015). Measuring sample quality with stein's method. In *Advances in Neural Information Processing Systems*, pages 226–234.

Gorham, J. and Mackey, L. (2017). Measuring sample quality with kernels. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1292–1301. JMLR. org.

Gretton, A. (2015). A simpler condition for consistency of a kernel independence test. *arXiv preprint arXiv:1501.06103*.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. J. (2007). A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520.

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

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005a). Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer.

Gretton, A., Bousquet, O., Smola, A. J., and Schölkopf, B. (2005b). Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT*, pages 63–78.

Gretton, A., Fukumizu, K., Harchaoui, Z., and Sriperumbudur, B. (2009a). A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems*, pages 673–681.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. (2008). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, pages 585–592.

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., and Sriperumbudur, B. K. (2012b). Optimal kernel choice for large-scale two-sample tests. In *Advances in neural information processing systems*, pages 1205–1213.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. (2009b). Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5.

Gretton, A., Smola, A. J., Bousquet, O., Herbrich, R., Belitski, A., Augath, M., Murayama, Y., Pauls, J., Schölkopf, B., and Logothetis, N. K. (2005c). Kernel constrained covariance for dependence measurement. In *AISTATS*, volume 10, pages 112–119. Citeseer.

Gu, S. S., Lheureux, S., Sayad, A., Cybulska, P., Hogen, L. B.-D., Vyarvelska, I., Tu, D., Parulekar, W., Levine, D. A., Bernardini, M. Q., et al. (2019). Computational modeling of ovarian cancer: Implications for therapy and screening. *medRxiv*, page 19009712.

Gutmann, M. U. and Hyvärinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13:307–361.

Hamelryck, T., Kent, J. T., and Krogh, A. (2006). Sampling realistic protein conformations using local structural bias. *PLoS Comput. Biol.*, 2:e131.

Heller, R. and Heller, Y. (2016). Multivariate tests of association based on univariate tests. In *Advances in Neural Information Processing Systems*, pages 208–216.

Hering, A. S. and Genton, M. G. (2010). Powering up with space-time wind forecasting. *Journal of the American Statistical Association*, 105(489):92–104.

Hoff, P. (2019). Package 'rstiefel'.

Hoff, P. D. (2009). Simulation of the matrix bingham–von mises–fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2):438–456.

Hollander, M. and Proschan, F. (1979). Testing to determine the underlying distribution using randomly censored data. *Biometrics*, pages 393–401.

Huggins, J. H. and Mackey, L. (2018). Random feature stein discrepancies. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1903–1913.

Hyde, J. (1977). Testing survival under right censoring and left truncation. *Biometrika*, 64(2):225–230.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709.

Jakob, W. (2012). Numerically stable sampling of the von mises-fisher distribution on s^2 (and other tricks). *Interactive Geometry Lab, ETH Zürich, Tech. Rep*.

Jean, N., Xie, S. M., and Ermon, S. (2018). Semi-supervised deep kernel learning: Regression with unlabeled data by minimizing predictive variance. In *Advances in Neural Information Processing Systems*, pages 5322–5333.

Jitkrittum, W., Kanagawa, H., Sangkloy, P., Hays, J., Schölkopf, B., and Gretton, A. (2018). Informative features for model comparison. In *Advances in Neural Information Processing Systems*, pages 808–819.

Jitkrittum, W., Kanagawa, H., and Schölkopf, B. (2020). Testing goodness of fit of conditional density models with kernels. *arXiv preprint arXiv:2002.10271*.

Jitkrittum, W., Szabó, Z., Chwialkowski, K. P., and Gretton, A. (2016a). Interpretable distribution features with maximum testing power. In *Advances in Neural Information Processing Systems*, pages 181–189.

Jitkrittum, W., Szabó, Z., and Gretton, A. (2016b). An adaptive test of independence with analytic kernel embeddings. *arXiv preprint arXiv:1610.04782*.

Jitkrittum, W., Xu, W., Szabó, Z., Fukumizu, K., and Gretton, A. (2017). A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems*, pages 262–271.

Jupp, P. et al. (2005). Sobolev tests of goodness of fit of distributions on compact riemannian manifolds. *The Annals of Statistics*, 33(6):2957–2966.

Jupp, P. et al. (2008). Data-driven sobolev tests of uniformity on compact riemannian manifolds. *The Annals of Statistics*, 36(3):1246–1260.

Jupp, P. and Kume, A. (2018). Measures of goodness of fit obtained by canonical transformations on riemannian manifolds. *arXiv preprint arXiv:1811.04866*.

Jupp, P., Mardia, K., et al. (1979). Maximum likelihood estimators for the matrix von mises-fisher and bingham distributions. *The Annals of Statistics*, 7(3):599–606.

Kanagawa, H., Jitkrittum, W., Mackey, L., Fukumizu, K., and Gretton, A. (2019). A kernel stein test for comparing latent variable models. *arXiv preprint arXiv:1907.00586*.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Kent, J., Mardia, K., and Rao, J. (1979). A characterization of the uniform distribution on the circle. *The Annals of Statistics*, pages 882–889.

Kent, J. T. (1982). The fisher–bingham distribution on the sphere. *J. Royal. Stat. Soc. B*, 44:71–80.

Kent, J. T., Ganeiber, A. M., and Mardia, K. V. (2013). A new method to simulate the bingham and related distributions in directional data analysis with applications. *arXiv preprint arXiv:1310.8110*.

Kim, I., Ramdas, A., Singh, A., and Wasserman, L. (2016). Classification accuracy as a proxy for two sample testing. *arXiv preprint arXiv:1602.02210*.

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

Kirchler, M., Khorasani, S., Kloft, M., and Lippert, C. (2020). Two-sample testing using deep learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1387–1398. PMLR.

Klein, J. and Moeschberger, M. (2006). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Kobayashi, S. and Nomizu, K. (1963). *Foundations of differential geometry*, volume 1. New York, London.

Koroljuk, V. S. and Borovskich, Y. V. (1994). *Theory of U-statistics*, volume 273 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht.

Korolyuk, V. S. and Borovskikh, Y. V. (1988). Asymptotic theory of U-statistics. *Ukrainian Mathematical Journal*, 40(2):142–154.

Lagakos, S., Barraj, L., and De Gruttola, V. (1988). Nonparametric analysis of truncated survival data, with application to AIDS. *Biometrika*, 75:515–523.

Le, H., Lewis, A., Bharath, K., and Fallaize, C. (2020). A diffusion approach to stein's method on riemannian manifolds. *arXiv preprint arXiv:2003.11497*.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lee, A. J. (1990). *U-Statistics: Theory and Practice*. CRC Press.

Lee, J. M. (2018). *Introduction to Riemannian manifolds*. Springer.

Leucht, A. and Neumann, M. H. (2013). Dependent wild bootstrap for degenerate u-and v-statistics. *Journal of Multivariate Analysis*, 117:257–280.

Ley, C., Reinert, G., Swan, Y., et al. (2017). Stein's method for comparison of univariate distributions. *Probability Surveys*, 14:1–52.

Ley, C. and Verdebout, T. (2017). *Modern directional statistics*. Chapman and Hall/CRC.

Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Póczos, B. (2017). Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pages 2203–2213.

Li, Y. and Turner, R. E. (2017). Gradient estimators for implicit models. *arXiv preprint arXiv:1705.07107*.

Lilliefors, H. W. (1967). On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association*, 62(318):399–402.

Liu, C. and Zhu, J. (2018). Riemannian stein variational gradient descent for bayesian inference. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Liu, Q., Lee, J., and Jordan, M. (2016). A kernelized stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, pages 276–284.

Liu, Q. and Wang, D. (2016). Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances In Neural Information Processing Systems*, pages 2378–2386.

Liu, S. and Kanamori, T. (2019). Estimating density models with complex truncation boundaries. *arXiv preprint arXiv:1910.03834*.

Lopez-Paz, D. and Oquab, M. (2016). Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*.

Loprinzi, C. L., Laurie, J. A., Wieand, H. S., Krook, J. E., Novotny, P. J., Kugler, J. W., Bartel, J., Law, M., Bateman, M., and Klatt, N. E. (1994). Prospective evaluation of prognostic variables from patient-completed questionnaires. north central cancer treatment group. *Journal of Clinical Oncology*, 12(3):601–607.

Ma, Y.-A., Chen, T., and Fox, E. (2015). A complete recipe for stochastic gradient mcmc. In *Advances in Neural Information Processing Systems*, pages 2917–2925.

Mardia, K. V. (2018). In *A new estimation methodology for standard directional distributions*.

Mardia, K. V., Holmes, D., and Kent, J. (1984). A goodness-of-fit test for the von mises-fisher distribution. *J. Royal. Stat. Soc. B*, 46:72–78.

Mardia, K. V. and Jupp, P. E. (1999). *Directional Statistics*. Wiley, New York, NY.

Mardia, K. V., Kent, J., and Laha, A. (2016). Score matching estimators for directional distributions. *arXiv:1604.08470*.

Martin, E. and Betensky, R. (2005). Testing quasi-independence of failure and truncation via conditional Kendall's tau. *Journal of the American Statistical Association*, 100:484–492.

Meister, R. and Schaefer, C. (2008). Statistical methods for estimating the probability of spontaneous abortion in observational studies—analyzing pregnancies exposed to coumarin derivatives. *Reproductive Toxicology*, 26(1):31–35.

Meynaoui, A., Albert, M., Laurent, B., and Marrel, A. (2019). Adaptive test of independence based on HSIC measures. *arXiv preprint arXiv:1902.06441*.

Mijoule, G., Reinert, G., and Swan, Y. (2018). Stein operators, kernels and discrepancies for multivariate continuous distributions. *arXiv preprint arXiv:1806.03478*.

Miller Jr, R. G. (2011). *Survival analysis*, volume 66. John Wiley & Sons.

Mirabello, L., Troisi, R., and Savage, S. (2009). Osteosarcoma incidence and survival rates from 1973 to 2004: data from the surveillance, epidemiology, and end results program. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 115(7):1531–1543.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966.

Oates, C. J., Cockayne, J., Briol, F.-X., and Girolami, M. (2019). Convergence rates for a class of estimators based on stein's method. *Bernoulli*, 25(2):1141–1159.

Pawlowsky-Glahn, V. and Buccianti, A. (2011). *Compositional data analysis: Theory and applications*. John Wiley & Sons.

Pourzanjani, A. A., Jiang, R. M., Mitchell, B., Atzberger, P. J., and Petzold, L. R. (2017). General bayesian inference over the stiefel manifold via the givens representation. *arXiv preprint arXiv:1710.09443*.

Radford, A., Metz, L., and Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*.

Ramdas, A., Trillos, N. G., and Cuturi, M. (2017). On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47.

Rindt, D., Sejdinovic, D., and Steinsaltz, D. (2019). Nonparametric independence testing for right-censored data using optimal transport. *arXiv preprint arXiv:1906.03866*.

Sedghi, H. and Anandkumar, A. (2014). Provable methods for training neural networks with sparse connectivity. *arXiv preprint arXiv:1412.2693*.

Sedghi, H., Gupta, V., and Long, P. M. (2019). The singular values of convolutional layers. In *ICLR*.

Sei, T., Shibata, H., Takemura, A., Ohara, K., and Takayama, N. (2013). Properties and applications of fisher distribution on the rotation group. *Journal of Multivariate Analysis*, 116:440–455.

Sejdinovic, D., Gretton, A., and Bergsma, W. (2013). A kernel test for three-variable interactions. In *Advances in Neural Information Processing Systems*, pages 1124–1132.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons.

Singh, H., Hnizdo, V., and Demchuk, E. (2002). Probabilistic model for two dependent circular variables. *Biometrika*, 89(3):719–723.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, 19(2):279–281.

Song, L., Huang, J., Smola, A., and Fukumizu, K. (2009). Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968.

Spivak, M. (2018). *Calculus on manifolds: a modern approach to classical theorems of advanced calculus*. CRC press.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Lanckriet, G. R., and Schölkopf, B. (2009). Kernel choice and classifiability for rkhs embeddings of probability distributions. In *23rd Annual Conference on Neural Information Processing Systems, NIPS 2009*, pages 1750–1758.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. (2011). Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(7).

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010). Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561.

Stein, C. et al. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California.

Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.

Student (1908). The probable error of a mean. *Biometrika*, pages 1–25.

Sutherland, D., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A., and Gretton, A. (2016). Generative models and model criticism via optimized maximum mean discrepancy. *arXiv:1611.04488*.

Sutherland, D. J. (2019). Unbiased estimators for the variance of mmd estimators. *arXiv preprint arXiv:1906.02104*.

Szabó, Z., Sriperumbudur, B. K., Póczos, B., and Gretton, A. (2016). Learning theory for distribution regression. *The Journal of Machine Learning Research*, 17(1):5272–5311.

Tolstikhin, I., Sriperumbudur, B. K., and Muandet, K. (2017). Minimax estimation of kernel mean embeddings. *The Journal of Machine Learning Research*, 18(1):3002–3048.

Tsai, W.-Y. (1988). Estimation of the survival function with increasing failure rate based on left truncated and right censored data. *Biometrika*, 75(2):319–324.

Tsai, W.-Y. (1990). Testing the assumption of independence of truncation time and failure time. *Biometrika*, 77:169–177.

Tsai, W.-Y., Jewell, N., and Wang, M. (1987). A note on the product-limit estimator under right censoring and left truncation. *Biometrika*, 74:883–886.

Turnbull, B. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 38:290–295.

Uehara, M., Matsuda, T., and Kim, J. K. (2020). Imputation estimators for unnormalized models with missing data. In *International Conference on Artificial Intelligence and Statistics*, pages 831–841. PMLR.

Van der Vaart, A. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wang, F., Xiang, X., Cheng, J., and Yuille, A. L. (2017). Normface: l_2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049. ACM.

Wang, M. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*, 86:130–143.

Wenliang, L., Sutherland, D., Strathmann, H., and Gretton, A. (2018). Learning deep kernels for exponential family densities. *arXiv preprint arXiv:1811.08357*.

Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. (2016). Deep kernel learning. In *Artificial intelligence and statistics*, pages 370–378.

Wood, A. T. A. (1994). Simulation of the von mises fisher distribution. *PLoS Comput. Biol.*, 23:157–164.

Woodroofe, M. (1985). Estimating a distribution function with truncated data. *The Annals of Statistics*, 13:163–177.

Xu, W. and Matsuda, T. (2020). A stein goodness-of-fit test for directional distributions. In *International Conference on Artificial Intelligence and Statistics*, pages 831–841. PMLR.

Yang, J., Liu, Q., Rao, V., and Neville, J. (2018). Goodness-of-fit testing for discrete distributions via stein discrepancy. In *International Conference on Machine Learning*, pages 5557–5566.

Yang, J., Rao, V., and Neville, J. (2019). A stein–papangelou goodness-of-fit test for point processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 226–235.

Yang, S. (1994). A central limit theorem for functionals of the Kaplan-Meier estimator. *Statist. Probab. Lett.*, 21(5):337–345.

Yuan, T. (2019). The 8-parameter fisher-bingham distribution on the sphere. *arXiv preprint arXiv:1906.08247*.

Zaremba, W., Gretton, A., and Blaschko, M. (2013). B-tests: Low variance kernel two-sample tests. In *NeurIPS*.

Zhang, Q., Filippi, S., Gretton, A., and Sejdinovic, D. (2018). Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130.