

Kernel-Based Hypothesis Tests: Large-Scale Approximations and Bayesian Perspectives



Qinyi Zhang
Wadham College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Trinity 2019

This thesis is dedicated to
Lao Tzu
for his mental support through his work “Tao Te Ching”.

Acknowledgements

I am forever grateful to my main supervisors, Dino Sejdinovic and Sarah Filippi, for their understanding, patient, support, advice and inspiration throughout my D.Phil journey. Dino has always helped me to solve problems that I have encountered and provided valuable insights that I have missed. I have always been amazed by the breadth and depth of his knowledge in the field of kernel methods. Sarah has always been there to challenge every statement made and every equation derived. She has taught me how to ask great questions and how to defend ones' own work. It has been a privilege to learn from such outstanding academics.

I would like to express my sincere gratitude to my collaborators Arthur Gretton and Seth Flaxman, for all the insightful discussions and fruitful research. I would like to thank all my colleges and friends for listening to my problems, complains, worries and for helping me through various hurdles in research and in life.

Furthermore, I would like to express my appreciation to my third supervisor, Yee Whye Teh, for providing financial supports for my travel to ICML and UAI in Sydney. I would also like to thank my funding organisation — the Engineering and Physical Sciences Research Council (EPSRC).

Last but not least, my gratitude to my family, without them, I won't be the "I", I am today.

Abstract

This thesis contributes to the field of nonparametric hypothesis testing (i.e. two-sample and independence testing) by providing a large-scale framework and developing a Bayesian perspective. We focus on nonparametric measures of homogeneity and dependence by considering the Hilbert norms between the embeddings of probability distributions in the reproducing kernel Hilbert space (RKHS). The rich representation provided by the associated kernel feature map enables the use of multivariate or non-Euclidean observations (e.g. strings and graphs) and it leads to powerful tests that are able to solve challenging problems given enough observations. However, the cost of computing the kernel matrix scales at least quadratically in the number of samples and makes it prohibitive to use in modern large-scale datasets. First, we propose three estimators of the well-known kernel dependence measure, the Hilbert Schmidt Independence Criterion (HSIC), namely the block-based estimator, the Nyström estimator and the random Fourier feature (RFF) estimator, and establish the corresponding linear time independence test for each of the estimators. Secondly, we consider a normalised version of HSIC, the NOrmalised Cross COvariance (NOCCO) statistic, and propose an RFF approximated NOCCO. This results in a distribution free test that is robust to the kernel bandwidth misspecification. Thirdly, we propose a two-step conditional independence test that extends the popular two-step approach REgression with Subsequent Independence Test (RESIT) through RKHS valued regressions. When used as a part of the classical PC algorithm for causal inference, the resulting algorithm is more robust to hidden variables that induce nonfunctional associations. Finally, we utilise the classical Bayes factor formalism for model comparison and propose a Bayesian two-sample test by modelling the witness function of the well-known kernel measure of homogeneity, the Maximum Mean Discrepancy (MMD), with a Gaussian Process.

Contents

1	Introduction	1
1.1	Thesis Outline	3
2	Background	7
2.1	Reproducing Kernel Hilbert Space	7
2.2	Kernel Mean Embeddings	11
2.3	The Maximum Mean Discrepancy	12
2.3.1	Universal and Characteristic Kernels	15
2.3.2	Relationship Between MMD and Other Discrepancy Measures	17
2.4	The Hilbert Schmidt Independence Criterion	18
3	Large-Scale Kernel Methods for Independence Testing	21
3.1	Introduction	22
3.2	Background	26
3.2.1	RKHS and Embeddings of Measures	26
3.2.2	Estimation of HSIC	29
3.2.3	Asymptotic Null Distribution of Estimators	30
3.2.4	Quadratic Time Null Distribution Estimations	32
3.2.4.1	Permutation Approach	32
3.2.4.2	Spectral Approach	33
3.3	Block-Based HSIC	33
3.3.1	The Block HSIC Statistic	34
3.3.2	Null Distribution of Block-Based HSIC	34
3.3.3	Linear Time Null Distribution Estimation	35
3.4	Approximate HSIC Through Primal Representations	36
3.4.1	Nyström HSIC	38
3.4.1.1	The Nyström HSIC Statistic	38
3.4.1.2	Null Distribution Estimations	39

3.4.2	Random Fourier Feature HSIC	40
3.4.2.1	Bochner’s Theorem	40
3.4.2.2	RFF HSIC Estimator	41
3.5	Experiments	42
3.5.1	Simple Linear Experiment	43
3.5.2	Nonlinear Experiments	44
3.5.2.1	Small-Scale Nonlinear Experiment	45
3.5.2.2	Large-Scale Nonlinear Experiment	46
3.5.2.3	Real Data Experiment	48
3.6	Discussion and Conclusion	49
3.7	Appendix: Proof of Theorem 3.2.1	50
4	Large-Scale Normalised Kernel Independence Tests	55
4.1	Background	55
4.1.1	Relationship Between Kernel Dependence Measures	57
4.1.2	Related Work	58
4.2	Empirical NOCCO Statistic	60
4.3	Random Fourier Features (RFF) Approximation	60
4.3.1	Relationship with Dauxois and Nkiet [1998]	64
4.4	Experiments	66
4.4.1	Robustness to Kernel Length-Scale	66
4.4.2	Computational Saving from Distribution Free Test	69
4.5	Conclusion	72
4.6	Appendix	73
4.6.1	Supporting Results	73
4.6.2	Simple Linear Experiment	73
5	Feature-to-Feature Regression for a Two-Step Conditional Independence Test	75
5.1	Introduction	76
5.2	Background	78
5.2.1	Weak Conditional Independence	78
5.2.2	RESIT: The Two-Step Approach	79
5.2.3	Kernel Tests for “Strong” Conditional Independence	81
5.3	Proposed Method	83
5.3.1	Related Work	87
5.4	Experiments	88

5.4.1	Simple Motivating Example Based on WHO Data	88
5.4.2	Effect of Dimensionality of Z and Sample Size	89
5.4.3	Application in Causal Discovery	91
5.4.3.1	Synthetic Data	92
5.4.3.2	Boston Housing Data	93
5.5	Conclusion	94
5.6	Appendix	94
5.6.1	Parameters Optimisation	94
5.6.2	Boston Housing Data	96
5.6.3	Ozone Data	97
6	Bayesian Two-Sample Testing	99
6.1	Introduction	99
6.2	Proposed Method	103
6.3	Alternative Model: $p(\theta M = 1, \mathcal{D})$	105
6.3.1	Covariance Matrix R_θ Computation	109
6.3.2	Covariance Matrix Σ_θ Estimation	109
6.3.3	Efficient Computation of Marginal Pseudolikelihood	112
6.3.4	Computation of the Jacobian Matrix	115
6.4	Null Model: $p(\theta M = 0, \mathcal{D})$	117
6.5	Posterior Inference	119
6.5.1	Bayes Factor with Fixed θ	119
6.5.2	Bayes Factor with Inferred θ	120
6.5.3	The Procedure for Posterior Inference	121
6.6	Experiments	121
6.6.1	Simple 1 Dimensional Gaussian Distributions	122
6.6.2	2 Dimensional Gaussian Distributions	125
6.6.3	1 Dimensional Mixture of Gaussian Distributions	126
6.6.4	2 by 2 Blobs of 2 Dimensional Gaussian Distributions	128
6.6.5	Higher Dimensional Gaussian Distributions	130
6.7	Conclusion	133
6.8	Appendix	134
7	Conclusion and Future work	138
7.1	Future Work	139
	Bibliography	143

Chapter 1

Introduction

This thesis considers the problem of nonparametric hypothesis testing. We focus on two types of hypothesis testing problems: two-sample testing and independence testing. In the two-sample testing problem, we will assume that we observe an independent and identically distributed (henceforth i.i.d.) samples $\mathbf{x} = \{x_i\}_{i=1}^{n_x}$ from random variable X and $\mathbf{y} = \{y_i\}_{i=1}^{n_y}$ from random variable Y with each $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$ respectively. $\mathcal{X}$ and $\mathcal{Y}$ are some generic non-empty domains. Assume the random variable X distributes according to probability measure P and Y distributes according to probability measure Q . We are interested in testing if we should reject the null hypothesis that the two samples come from the same distribution in favour of the general alternative hypothesis that the two samples come from different distributions, i.e.

$$\mathcal{H}_0 : P = Q \text{ v.s. } \mathcal{H}_1 : P \neq Q. \quad (1.1)$$

In the independence testing problem, on the other hand, we will assume that we are given an i.i.d. sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ with each $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ on some generic non-empty domain $\mathcal{X}$ and $\mathcal{Y}$. Assume P_X and P_Y are the marginal distributions of the random variables X and Y respectively and P_{XY} the joint distribution. We are interested in testing the following null and alternative hypotheses:

$$\mathcal{H}_0 : P_{XY} = P_X P_Y \text{ v.s. } \mathcal{H}_1 : P_{XY} \neq P_X P_Y. \quad (1.2)$$

In the literature, there exist many popular nonparametric measures of discrepancy, for example, the class of f-divergences [Csiszár, 1963, Morimoto, 1963, Ali and Silvey, 1966] and the integral probability metrics [Müller, 1997]. In terms of dependence measures, there are frameworks based on mutual information estimators [Cover and Thomas, 1991, Kraskov et al., 2003] as well as characteristic functions [Székely et al.,

2007, Székely and Rizzo, 2009] and kernel methods [Gretton et al., 2005a, 2008] which have close links [Sejdinovic et al., 2013b]. In this thesis, we focus on nonparametric measures of discrepancy and dependence based on representations of probability distributions in reproducing kernel Hilbert spaces (RKHSs). Every RKHS is uniquely associated with a positive definite kernel, i.e. an inner product between maps that could potentially be infinite dimensional. These maps are called *feature maps* in the kernel literature. Feature maps allow us to work in a nonlinearly transformed feature space (i.e. the RKHS) and hence the employment of linear machine learning methods can achieve nonlinear effects. In addition, the flexibility of kernel based methods also comes from its applicability to generic domains beyond Euclidean spaces, e.g. string, graph and images.

In Chapter 3 and 4 of this thesis, we focus on the problem of large-scale independence testing with a nonparametric measure of dependence — the Hilbert Schmidt independence criterion (HSIC) [Gretton et al., 2005a]. While many classical dependence measures are designed to capture only specific forms of dependence (e.g. linear or monotone), HSIC is able to capture more complex dependence patterns from non-standard input domains. However, this is achieved at a prohibitive computational cost that scales at least quadratically with the number of samples due to the computation of the kernel between all pairs of observations. Arguably, it is exactly in large-scale datasets that capturing any type of dependence is of interest. Through the use of large-scale kernel approximation methods, we propose large-scale kernel dependence tests and its normalised counterparts. They lead to strong improvements in performance as measured per power-per-computational unit performance and the resulting fast independence testing framework is responsive to all forms of dependence.

Following these two chapters, motivated by the application of nonparametric conditional independence tests in causal inference, we propose an improvement to the popular REgression with Subsequent Independence Test (RESIT) through RKHS-valued regressions. Noticing that RESIT can lead to an inflated number of false positives when the relationships between random variables do not follow associations that can be expressed by a single functional relationship. By regressing out the dependence on the conditioning variables, RESIT is able to utilise nonparametric independence tests to detect the dependence in the regression residuals to obtain information on conditional dependence. The large-scale framework developed in the previous two chapters readily applies here and can be used to scale up the proposed

conditional independence test.

While the nonparametric two-sample test and independence test have proven to be powerful in many applications, their performance depends heavily on the kernel used. When Gaussian kernel is used, this is equivalent to choosing the bandwidth for the problem at hand. Motivated by this problem, in the final chapter, we develop a framework for Bayesian two-sample testing utilising the kernel discrepancy measure — the Maximum Mean Discrepancy (MMD) [Borgwardt et al., 2006, Gretton et al., 2012a] and the Bayesian learning framework of kernel mean embeddings from Flaxman et al. [2016]. The approach provides an inference scheme for the posterior distribution of the bandwidth parameter and hence effectively selects a range of bandwidth values that are useful for testing. This work complements the existing kernel-based frequentist tests and provides a missing piece to the landscape of kernel-based hypothesis testing.

1.1 Thesis Outline

This is an integrated thesis, where Chapter 3 and 5 consist of self-contained published papers. Chapter 2 introduces the necessary concepts in kernel methods needed in the subsequent chapters. We provide definitions and properties of the reproducing kernel Hilbert space (RKHS) and the corresponding kernel. The nonparametric testing framework relies on representing probability distributions with elements of a RKHS. These representations are known as the kernel mean embeddings. We provide their formal definition and conditions on probability distributions such that the kernel mean embedding exists. We introduce the Maximum Mean Discrepancy (MMD), i.e. the kernel discrepancy measure between probability distributions in the RKHS, the important measure used in Chapter 6. We further explore the relationship of MMD with other classical measures of divergence, e.g. the class of f -divergence and integral probability metric. Related to MMD, we introduce the Hilbert Schmidt independence criterion (HSIC), the important kernel dependence measure used in Chapter 3 and 5.

In Chapter 3, we propose three large-scale estimators of HSIC: the block-based estimator, the Nyström estimator and the random Fourier feature (RFF) estimator. Suitable independence testing procedure is introduced for each of the methods to achieve an overall linear computational complexity in memory and time over the traditional HSIC-based tests. The favourable tradeoffs are verified in several challenging

high dimensional synthetic datasets and one real dataset. This is joint work with Sarah Filippi, Arthur Gretton and Dino Sejdinovic. With help from Sejdinovic, I contributed to the derivation of Nyström HSIC statistic and the RFF HSIC statistic. I implemented the three proposed large scale kernel based independence tests and conducted all the experiments. Under the guidance of Sejdinovic, I proved the asymptotic null distribution of the biased HSIC statistic. The literature review part of the paper was a collaborative effort between Sejdinovic, Gretton and myself. The other sections of the paper were written by me with revisions provided by Sejdinovic and Filippi. This work is published in:

Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic, *Large-Scale Kernel Methods for Independence Testing, Statistics and Computing*, vol. 28, no. 1, 113–130, Jan. 2018.

As an extension of Chapter 3, Chapter 4 proposes an RFF estimator of the Normalised Cross Covariance Operator (NOCCO) statistic introduced by Fukumizu et al. [2008], a normalised kernel measure of dependence that has been shown to have advantages over HSIC. We provide a simple derivation of the asymptotic null distribution of the RFF NOCCO statistic. The form of the proposed statistic and its asymptotic null distribution can be considered as a special case of the framework established in Dauxois and Nkiet [1998]. However, we also present and elaborate on the subtle differences between the two works. Through a variety of synthetic data experiments, we demonstrate the robustness of NOCCO based test on kernel bandwidth selection and verify their computational savings. This chapter is previously unpublished and is joint work with my supervisors Sarah Filippi and Dino Sejdinovic. I contributed to the implementation of the NOCCO based testing framework and its corresponding large scale version. I conducted all the experiments. The derivation of the asymptotic null distribution of the approximated NOCCO statistic was a collaborative effort between Sejdinovic and myself. I established and derived the connection of our work and that of Dauxois and Nkiet [1998]. The writing of the paper was done by me with revisions provided by Sejdinovic and Filippi. We would like to thank Arthur Gretton for providing the relevant reference of Dauxois and Nkiet [1998] and useful discussions on this work.

Chapter 5 proposes a conditional independence test that extends the popular two-step approach REgression with Subsequent Independence Test (RESIT). Through considering RKHS-valued ridge regression, we maintain the simple testing procedure

of RESIT while having a correct Type I control under challenging scenarios where the modelling assumptions required by RESIT are violated. We integrate the proposed test into the classical PC algorithm and observe that the resulting algorithm is more robust to hidden variables that induce nonfunctional associations. This is joint work with Sarah Filippi, Seth Flaxman and Dino Sejdinovic. With the idea from Sejdinovic, I derived the proposed RESIT approach and its implementation details. The connection of the proposed RESIT to conditional independence testing was established through the joint effort of Sejdinovic, Filippi and myself. I conducted all the experiments and was responsible for most of the writing with revisions by Sejdinovic, Filippi and Flaxman. This chapter is published in:

Q. Zhang, S. Filippi, S. Flaxman, and D. Sejdinovic, *Feature-to-Feature Regression for a Two-Step Conditional Independence Test, in Uncertainty in Artificial Intelligence (UAI), 2017.*

In Chapter 6, we propose a Bayesian formulation of the nonparametric two-sample testing problem utilising the Bayes factor approach. The empirical kernel mean embeddings evaluated at a set of inducing points are considered as observations, and the learning procedure of such empirical kernel mean embedding follows the framework proposed by Flaxman et al. [2016]. We derive a joint posterior inference scheme of the model label and the kernel hyperparameter, which effectively avoids the kernel bandwidth selection problem that is often crucial for the performance of nonparametric two-sample frequentist tests. Through a range of synthetic data experiments, we study the performance of the proposed Bayesian framework noting that its advantages lie in simultaneously considering multiple lengthscales. This is joint work with Sarah Filippi, Seth Flaxman and Dino Sejdinovic. I proposed and derived the framework for Bayesian two-sample testing utilising the work of Flaxman et al. [2016]. The derivation of the first covariance matrix estimation was a joint effort between Sejdinovic and myself, and I derived the second covariance matrix estimation. For the section on efficient computation of marginal pseudolikelihood, the proof was a collaborative effort between Sejdinovic, Filippi and myself. I derived the computation of the Jacobian matrix. Moreover, I presented and derived the null model of the kernel hyperparameter given the model label and observations. With guidance from Sejdinovic, I derived the procedure for the posterior inference and conducted all the experiments. The writing was done by me with revisions from Sejdinovic and Filippi.

The thesis is concluded with a summary of the main contributions and possible future research directions. In addition, we position our work against other contributions to kernel methods and related literature.

Chapter 2

Background

At the core of this thesis lies the framework of measuring distance between probability distributions using reproducing kernel Hilbert spaces. In this chapter, we briefly review key concepts of reproducing kernel Hilbert spaces and the kernel mean embeddings of probability distributions. For the rigorous treatment of the topic, we refer the readers to Berlinet and Thomas-Agnan [2004], Steinwart and Christmann [2008], Kraskov et al. [2003]. We will also review the kernel-based probability metric – the maximum mean discrepancy and the kernel-based measure of dependence – the Hilbert Schmidt independence criterion. The following chapters will assume the background material presented herein. Chapter 3 and Chapter 5 are self-contained papers which include partial overview of the background materials presented in this chapter.

2.1 Reproducing Kernel Hilbert Space

Let $\mathcal{Z}$ be any topological space on which Borel measures can be defined. By $\mathcal{M}(\mathcal{Z})$ we denote the set of all finite signed Borel measures on $\mathcal{Z}$ and by $\mathcal{M}_+^1(\mathcal{Z})$ the set of all Borel probability measures on $\mathcal{Z}$. We will now review the basic concepts of RKHS and kernel embeddings of probability measures. For further details, see Berlinet and Thomas-Agnan [2004], Steinwart and Christmann [2008], Sriperumbudur [2010].

Definition 2.1.1. *[Steinwart and Christmann, 2008, Definition 4.18] Let $\mathcal{H}$ be a Hilbert space of real-valued functions defined on $\mathcal{Z}$. A function $k : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$ is called a **reproducing kernel** of $\mathcal{H}$ if:*

1. $\forall z \in \mathcal{Z}, k(\cdot, z) \in \mathcal{H}$
2. $\forall z \in \mathcal{Z}, \forall f \in \mathcal{H}, \langle f, k(\cdot, z) \rangle_{\mathcal{H}} = f(z)$. (The Reproducing Property)

If $\mathcal{H}$ has a reproducing kernel, it is called a **Reproducing Kernel Hilbert Space (RKHS)**.

As a direct consequence of the reproducing property, for any $x, y \in \mathcal{Z}$, the reproducing kernel can be written as

$$k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}}. \quad (2.1)$$

For $x, y \in \mathbb{R}^p$, some examples of reproducing kernels are

- Linear kernel: $k(x, y) = x^T y$;
- Polynomial kernel of degree $d \in \mathbb{N}$: $k(x, y) = (x^T y + 1)^d$;
- Gaussian kernel with bandwidth $\sigma > 0$: $k(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$;
- Laplacian kernel with bandwidth $\sigma > 0$: $k(x, y) = \exp\left(-\frac{\|x-y\|}{\sigma}\right)$;
- Fractional Brownian motion covariance kernel with parameter $h \in (0, 1)$: $k(x, y) = \frac{1}{2} (\|x\|^{2h} + \|y\|^{2h} - \|x - y\|^{2h})$.

Reproducing kernels can also be defined on graphs, text, images, strings and probability distributions [Shawe-Taylor and Cristianini, 2004].

In the machine learning literature, a notion of *kernel* is understood as an inner product between feature maps.

Definition 2.1.2. [Steinwart and Christmann, 2008, Definition 4.1] *Let $\mathcal{X}$ be a non-empty set. Then a function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is called a **kernel** if there exists a Hilbert space $\mathcal{H}$ and a map $\phi : \mathcal{X} \rightarrow \mathcal{H}$ s.t. $\forall x, x' \in \mathcal{X}$,*

$$k(x, x') = \langle \phi(x), \phi(x') \rangle. \quad (2.2)$$

*We call ϕ a **feature map** and $\mathcal{H}$ a **feature space** of k .*

By (2.1), every reproducing kernel is a kernel in this sense, corresponding to a *canonical feature map* $\phi(x) : x \mapsto k(\cdot, x)$.

The most important property of kernel functions is that they are positive definite [Berlinet and Thomas-Agnan, 2004, Definition 2] [Steinwart and Christmann, 2008, Definition 4.15].

Definition 2.1.3. A function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is called **positive definite** if $\forall n \in \mathbb{N}$, $\forall \alpha_1, \dots, \alpha_n \in \mathbb{R}$ and $\forall x_1, \dots, x_n \in \mathcal{X}$, we have

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(x_j, x_i) \geq 0. \quad (2.3)$$

Further more, k is said to be **strictly positive definite**¹ if for mutually distinct $x_1, \dots, x_n \in \mathcal{X}$ equality in (2.3) only holds for $\alpha_1 = \dots = \alpha_n = 0$.

For real-valued kernel functions $k(x, x') = \langle \phi(x), \phi(x') \rangle$ defined in Definition 2.1.2,

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(x_j, x_i) = \left\langle \sum_{i=1}^n \alpha_i \phi(x_i), \sum_{j=1}^n \alpha_j \phi(x_j) \right\rangle_{\mathcal{H}} = \left\| \sum_{i=1}^n \alpha_i \phi(x_i) \right\|_{\mathcal{H}}^2 \geq 0. \quad (2.4)$$

Hence, every inner product is a positive definite function for a general Hilbert space $\mathcal{H}$. Since reproducing kernels are in fact inner products, it is immediate that they are positive definite. By convention, the $n \times n$ matrix $K := [k(x_i, x_j)]_{i,j}$ is called the **Gram matrix** for fixed $x_1, \dots, x_n \in \mathcal{X}$. Definition 2.1.3 is equivalent to saying the Gram matrix is positive definite.

As we will illustrate in Example 2.1.1, the feature map and the feature space are not uniquely defined for a given kernel.

Example 2.1.1. [Steinwart and Christmann, 2008, p.113] When $\mathcal{X} = \mathbb{R}$, consider the kernel $k(x, y) = xy$ for all $x, y \in \mathbb{R}$, then

$$k(x, y) = xy = \begin{bmatrix} \frac{x}{\sqrt{2}} & \frac{x}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{y}{\sqrt{2}} \\ \frac{y}{\sqrt{2}} \end{bmatrix}. \quad (2.5)$$

This construction implies this simple kernel can arise from the feature map $\phi(x) = x$ with the feature space $\mathcal{H} = \mathbb{R}$ but also from the feature map $\tilde{\phi}(x) = \begin{bmatrix} \frac{x}{\sqrt{2}} & \frac{x}{\sqrt{2}} \end{bmatrix}$ with the feature space $\tilde{\mathcal{H}} = \mathbb{R}^2$. Such construction can be made for arbitrary kernels and hence every kernel has many different feature spaces.

However, it can be shown that for every given reproducing kernel Hilbert space, it has a unique reproducing kernel [Steinwart and Christmann, 2008, Theorem 4.20] [Aronszajn, 1950, Section I.2 (1) & (2)] and for every symmetric positive definite function $k : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$, there exists a unique RKHS for which k is a reproducing kernel

¹Some authors call positive definite functions positive semi-definite, and strictly positive definite functions positive definite.

[Aronszajn, 1950, Moore -Aronszajn theorem, Section I.2 (4)]. In other words, there is a one-to-one correspondence between the positive definite kernel and the RKHS.

Reproducing kernel Hilbert spaces have the important property that norm convergence implies pointwise convergence. We make this precise in Proposition 2.1.1 below.

Proposition 2.1.1. *[Berlinet and Thomas-Agnan, 2004, Corollary 1] If sequence $\{f_n\}$ in $\mathcal{H}$ and f in $\mathcal{H}$ are such that $\lim_{n \rightarrow \infty} \|f_n - f\|_{\mathcal{H}} = 0$, then $\lim_{n \rightarrow \infty} f_n(z) = f(z)$, $\forall z \in \mathcal{Z}$.*

Proof. For any $z \in \mathcal{Z}$,

$$|f_n(z) - f(z)| = |\langle f_n, k(\cdot, z) \rangle_{\mathcal{H}} - \langle f, k(\cdot, z) \rangle_{\mathcal{H}}| \quad (2.6)$$

$$= |\langle f_n - f, k(\cdot, z) \rangle_{\mathcal{H}}| \quad (2.7)$$

$$\leq \|f_n - f\|_{\mathcal{H}} \|k(\cdot, z)\|_{\mathcal{H}} \quad (2.8)$$

by Cauchy-Schwarz inequality and the fact $\|k(\cdot, z)\|_{\mathcal{H}} = \sqrt{k(z, z)}$ is bounded by definition on the RKHS² [Kreyszig, 1989, p.96]. $\square$

As a consequence, RKHSs are relatively well behaved compared to other Hilbert spaces. We will present an example of a Hilbert space of continuous functions equipped with L_2 norm where the evaluation of functions is not continuous.

Example 2.1.2. *[Berlinet and Thomas-Agnan, 2004, p.2] considers the Hilbert space $\mathcal{H} = L_2([0, 1])$, and the sequence of functions $\{q_n\}_{n=1}^{\infty}$ with $q_n = x^n$. The norm:*

$$\lim_{n \rightarrow \infty} \|q_n - 0\|_{L_2([0,1])} = \lim_{n \rightarrow \infty} \left(\int_0^1 x^{2n} dx \right)^{1/2} \quad (2.9)$$

$$= \lim_{n \rightarrow \infty} \frac{1}{\sqrt{2n+1}} \quad (2.10)$$

$$= 0. \quad (2.11)$$

However, $q_n(1) = 1$ for all n . In other words, $\lim_{n \rightarrow \infty} q_n(1) \neq 0$. Hence, the evaluation of functions at point 1 is not continuous on the set $\{q_n\}_{n=1}^{\infty}$.

²An alternative definition of RKHS [Steinwart and Christmann, 2008, Definition 4.18] is through linear evaluation functionals. For a fixed point $z \in \mathcal{Z}$, the evaluation functional $\delta_z : \mathcal{H} \rightarrow \mathbb{R}$ is such that $\delta_z : f \mapsto f(z)$ for $f : \mathcal{Z} \rightarrow \mathbb{R}$ elements of the Hilbert space $\mathcal{H}$. $\mathcal{H}$ is a RKHS if δ_z is continuous at any $f \in \mathcal{H}$ or equivalently δ_z is a bounded operator on $\mathcal{H}$. $k(\cdot, z)$ is representer of evaluation at z .

2.2 Kernel Mean Embeddings

In the nonparametric testing framework, a particularly useful setup will be representing probability distributions with elements of a RKHS [Berlinet and Thomas-Agnan, 2004, Smola et al., 2007]. This is known as the kernel mean embeddings, which we define in Definition 2.2.1.

Definition 2.2.1. *Let k be a kernel on $\mathcal{Z}$, and $\nu \in \mathcal{M}_+^1(\mathcal{Z})$. The **kernel embedding** of probability measure ν into the RKHS $\mathcal{H}_k$ is $\mu_k(\nu) \in \mathcal{H}_k$ such that*

$$\int f(z) d\nu(z) = \langle f, \mu_k(\nu) \rangle_{\mathcal{H}_k}, \quad \forall f \in \mathcal{H}_k. \quad (2.12)$$

It is understood from this definition that the integral of any RKHS function f under the measure ν can be evaluated as the inner product between f and the kernel embedding $\mu_k(\nu)$ in the RKHS $\mathcal{H}_k$. As an alternative, the kernel embedding can also be defined through the use of Bochner integral $\mu_k(\nu) = \int k(\cdot, z) d\nu(z)$. Any probability measure is mapped to the corresponding expectation of the canonical feature map. A sufficient condition for the existence of an embedding of ν is that $\nu \in \mathcal{M}_+^1(\mathcal{Z}) \cap \mathcal{M}_k^{1/2}(\mathcal{Z})$ [Fukumizu et al., 2004b, p.93] [Sriperumbudur et al., 2010, Theorem 1][Gretton et al., 2012a, Lemma 3], where we adopt notation from Sejdinovic et al. [2013b]:

$$\mathcal{M}_k^\theta(\mathcal{Z}) = \left\{ \nu \in \mathcal{M}(\mathcal{Z}) : \int k^\theta(z, z) d|\nu|(z) < \infty \right\}.$$

We will make this precise in the following Lemma:

Lemma 2.2.1. *[Gretton et al., 2012a, Lemma 3] If the kernel k on $\mathcal{Z}$ is measurable and $\mathbb{E}(\sqrt{k(z, z)}) < \infty$, then $\mu_k \in \mathcal{H}_k$.*

For all $f \in \mathcal{H}$, the linear operator $T : f \mapsto \mathbb{E}(f(z)) \in \mathbb{R}$ is bounded by

$$|Tf| = |\mathbb{E}(f(z))| \leq \mathbb{E}(|f(z)|) = \mathbb{E}|\langle f, \phi(z) \rangle_{\mathcal{H}}| \leq \mathbb{E}\left(\sqrt{k(z, z)} \|f\|_{\mathcal{H}_k}\right) < \infty \quad (2.13)$$

under the assumption of Lemma 2.2.1. Consequently, T is a bounded linear operator on a normed space and hence it is continuous. Riesz representer theorem [Reed and Simon, 1980, Theorem II.4] says that all continuous linear functionals in a Hilbert space arise from an inner product. This implies that there exists $\mu_k(\nu) \in \mathcal{H}$ such that $Tf = \langle f, \mu_k(\nu) \rangle_{\mathcal{H}}$. Proposition 2 of Sriperumbudur et al. [2010] further showed that, in fact, for bounded kernel functions, i.e. $\sup_{z \in \mathcal{Z}} k(z, z) < \infty$, the mean embedding is well defined for *all* probability measures. An example of such kernel is the Gaussian

kernel.

Conveniently, this framework allows standard operations in the Hilbert spaces to be used to manipulate probability distributions. As we will see, distance between probability distributions can be easily formulated in terms of the Hilbert norm between their respective kernel mean embeddings.

2.3 The Maximum Mean Discrepancy

A well-known kernel distance between probability distributions in the RKHS is the maximum mean discrepancy (MMD) [Borgwardt et al., 2006, Gretton et al., 2012a]. In this section, we will follow the exposition presented by Gretton et al. [2012a]. For a general function space, the MMD is defined as [Gretton et al., 2012a, Definition 2]

Definition 2.3.1. *Let X and Y be random variables defined on a topological space $\mathcal{X}$ with respective Borel probability measures P and Q . Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \rightarrow \mathbb{R}$, the maximum mean discrepancy (MMD) is defined as*

$$MMD[\mathcal{F}, P, Q] := \sup_{f \in \mathcal{F}} (\mathbb{E}_X[f(X)] - \mathbb{E}_Y[f(Y)]) \quad (2.14)$$

In other words, we wish to find the function that maximises the difference in expectation between the probability distributions (i.e. mean discrepancy). The RKHS function that attains the supremum in (2.14) is called the witness function f^* . Since MMD can be written as

$$MMD[\mathcal{F}, P, Q] = \langle f, \mathbb{E}_X(k(X, \cdot)) - \mathbb{E}_Y(k(Y, \cdot)) \rangle_{\mathcal{H}_k} \quad (2.15)$$

by the linearity of expectation and inner product, the unit vector f that maximises the above inner product in a Hilbert space³ is

$$f^* = \frac{\mathbb{E}_X(k(X, \cdot)) - \mathbb{E}_Y(k(Y, \cdot))}{\|\mathbb{E}_X(k(X, \cdot)) - \mathbb{E}_Y(k(Y, \cdot))\|_{\mathcal{H}_k}} \quad (2.16)$$

$$\propto \mathbb{E}_X(k(X, \cdot)) - \mathbb{E}_Y(k(Y, \cdot)). \quad (2.17)$$

Figure.2.1 illustrates the function f^* that maximises the mean discrepancy between a standard Gaussian distribution and a t-distribution with 1 degree of freedom using a Gaussian kernel with lengthscale parameter set to 0.5. We see the function peaks above zero when the t density is smaller than the Gaussian density and drops below

³Inner product achieves supremum when the two vectors are parallel.

zero when the t density is larger than the Gaussian density. The magnitude is indicative of the degree to which the two distributions differ.

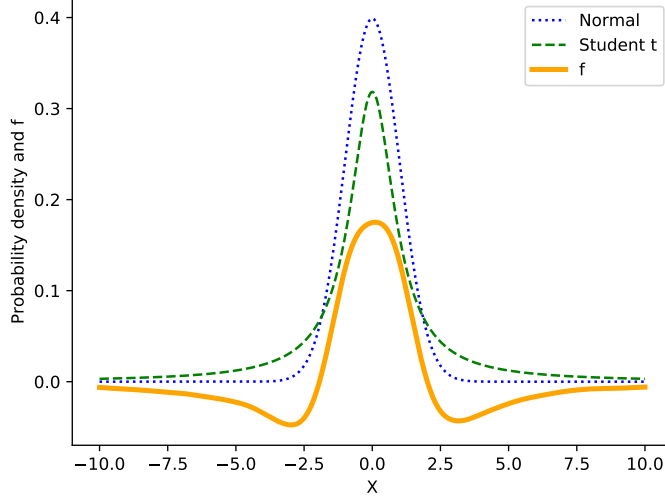


Figure 2.1: Illustration of the witness function f comparing the Gaussian distribution $\mathcal{N}(0, 1)$ and the t distribution with 1 degree of freedom. Note, f has been rescaled for ease of illustration.

When the class of function $\mathcal{F}$ is the space of bounded continuous real-valued functions (i.e. a Banach space with supremum norm [Dudley, 2002, p.158], [Gretton et al., 2012a]), the equality between $\mathbb{E}_X[f(X)]$ and $\mathbb{E}_Y[f(Y)]$ ensures the unique identification of $p = q$ [Dudley, 2002, Lemma 9.3.2]. Practically, this class of function is too rich to work with in the finite sample setting [Gretton et al., 2012a]. When restricting the class of function to the unit ball in a RKHS $\mathcal{H}_k$, Borgwardt et al. [2006] showed that the MMD can be expressed by

$$MMD^2[\mathcal{F}, P, Q] = \|\mu_P - \mu_Q\|_{\mathcal{H}_k}^2 \quad (2.18)$$

with the kernel mean embedding μ_P defined as $\mathbb{E}_X(f) = \langle f, \mu_P \rangle_{\mathcal{H}_k}$ for all $f \in \mathcal{H}_k$ and similarly for μ_Q .

The $MMD[\mathcal{F}, P, Q]$ is however not a metric on the Borel probability measures for all RKHSs $\mathcal{H}$. For example, the polynomial kernels of degree t can only distinguish between distributions that differ in moments upto and including t . Example 3 of Sriperumbudur et al. [2010] shows that a polynomial RKHS of degree two cannot

distinguish between distributions with the same mean and variance but different kurtosis. This example utilises equivalent forms of MMD, and provides helpful insights into the computation of MMD so we report it here for completeness.

Example 2.3.1. [Sriperumbudur et al., 2010, Example 3] Let $k(x, y) = (1 + x^\top y)^2$, where $x, y \in \mathbb{R}^d$. In this example, we will subscript the expectations with the relevant probability measures for clarity. Given the definition of MMD, it can be written more explicitly as [Sriperumbudur et al., 2010, p.1526]

$$\text{MMD}^2[\mathcal{F}, P, Q] := \|\mu_P - \mu_Q\|^2 \quad (2.19)$$

$$= \left\| \int_{\mathbb{R}^d} k(\cdot, x) dP(x) - \int_{\mathbb{R}^d} k(\cdot, x) dQ(x) \right\|_{\mathcal{H}_k}^2 \quad (2.20)$$

$$\begin{aligned} &= \left\langle \int_{\mathbb{R}^d} k(\cdot, x) dP(x), \int_{\mathbb{R}^d} k(\cdot, y) dP(y) \right\rangle \\ &\quad + \left\langle \int_{\mathbb{R}^d} k(\cdot, x) dQ(x), \int_{\mathbb{R}^d} k(\cdot, y) dQ(y) \right\rangle \\ &\quad - 2 \left\langle \int_{\mathbb{R}^d} k(\cdot, x) dP(x), \int_{\mathbb{R}^d} k(\cdot, y) dQ(y) \right\rangle \end{aligned} \quad (2.21)$$

$$= \int \int_{\mathbb{R}^d} k(x, y) d(P - Q)(x) d(P - Q)(y) \quad (2.22)$$

for probability measures P and Q by the reproducing property of k . Substitute the definition of the polynomial kernel of degree 2 from above, we obtain

$$\text{MMD}^2[\mathcal{F}, P, Q] := \int \int_{\mathbb{R}^d} (1 + 2x^\top y + x^\top y y^\top x) d(P - Q)(x) d(P - Q)(y) \quad (2.23)$$

$$= \int \int_{\mathbb{R}^d} 1 d(P - Q)(x) d(P - Q)(y) \quad (2.24)$$

$$+ 2 \int \int_{\mathbb{R}^d} x^\top y d(P - Q)(x) d(P - Q)(y) \quad (2.25)$$

$$+ \int \int_{\mathbb{R}^d} x^\top y y^\top x d(P - Q)(x) d(P - Q)(y). \quad (2.26)$$

(2.24) is 0 since the expectation of a constant value w.r.t. any probability measure is a constant. (2.25) is essentially considering MMD between P and Q with respect to the kernel $\tilde{k}(x, y) = x^\top y$. In this case, $\mu_P = \mathbb{E}(\tilde{k}(\cdot, X)) = \mathbb{E}(X) = \int_{\mathbb{R}^d} x dP(x)$. Similarly for y . This implies, (2.25) can be written as

$$\int \int_{\mathbb{R}^d} x^\top y d(P - Q)(x) d(P - Q)(y) = \|\mathbb{E}_P(X) - \mathbb{E}_Q(X)\|_2^2 = \|\mu_P - \mu_Q\|_2^2 \quad (2.27)$$

For (2.26), note that

$$\int \int_{\mathbb{R}^d} x^\top y y^\top x \, d(P - Q)(x) d(P - Q)(y) \quad (2.28)$$

$$= \int \int_{\mathbb{R}^d} x^\top y y^\top x \{dP(x)dP(y) + dQ(x)dQ(y) - 2dP(x)dQ(y)\} \quad (2.29)$$

$$= \text{Tr} \int \int_{\mathbb{R}^d} x x^\top y y^\top \{dP(x)dP(y) + dQ(x)dQ(y) - 2dP(x)dQ(y)\} \quad (2.30)$$

$$= \text{Tr} \{ \mathbb{E}_P(XX^\top) \mathbb{E}_P(YY^\top) + \mathbb{E}_Q(XX^\top) \mathbb{E}_Q(YY^\top) - 2\mathbb{E}_P(XX^\top) \mathbb{E}_Q(YY^\top) \} \quad (2.31)$$

$$= \text{Tr} \{ [\mathbb{E}_P(XX^\top) - \mathbb{E}_Q(YY^\top)]^\top [\mathbb{E}_P(XX^\top) - \mathbb{E}_Q(YY^\top)] \} \quad (2.32)$$

Since the matrix $\mathbb{E}_P(XX^\top) - \mathbb{E}_Q(YY^\top)$ is symmetric, by property of Frobenius norm⁴, we obtain that the above is exactly $\|\mathbb{E}_P(XX^\top) - \mathbb{E}_Q(YY^\top)\|_F^2$. If we denote the covariance matrix of P by Σ_P , then $\Sigma_P := \mathbb{E}_P(XX^\top) - \mathbb{E}_P(X)\mathbb{E}_P(X)^\top$. This implies $\mathbb{E}_P(XX^\top) = \Sigma_P + \mu_P \mu_P^\top$. The relevant quantities for Q follows similarly.

Overall, under the polynomial kernel of order 2, the MMD can be written as

$$\text{MMD}^2[\mathcal{F}, P, Q] = 2\|\mu_P - \mu_Q\|_2^2 + \|\Sigma_P - \Sigma_Q + \mu_P \mu_P^\top - \mu_Q \mu_Q^\top\|_F^2, \quad (2.33)$$

i.e. $\text{MMD}^2 = 0 \implies \mu_P = \mu_Q$ and $\Sigma_P = \Sigma_Q$. However, not necessarily $P = Q$.

2.3.1 Universal and Characteristic Kernels

The concept of universal kernels originally arises in the study of consistency of classifiers of support vector machine type [Steinwart, 2001]. They can be formally defined as

Definition 2.3.2. [Steinwart, 2001, Definition 4] A continuous kernel k on a compact metric space $(\mathcal{X}, d)$ is called **universal** if the space of all functions induced by k is dense in $C(\mathcal{X})$, i.e. for every function $f \in C(\mathcal{X})$ and every $\epsilon > 0$ there exists a function g induced by k (i.e. in $\mathcal{H}_k$) with

$$\|f - g\|_\infty \leq \epsilon. \quad (2.34)$$

When using this family of kernels, Theorem 3 of Gretton et al. [2007] and Theorem 5 of Gretton et al. [2012a] showed that $\text{MMD}[\mathcal{F}, p, q]$ is in fact a metric on Borel probability measures.

⁴ $\|A\|_F^2 = \text{trace}(AA^\top)$ for $m \times n$ real matrix A .

Theorem 2.3.1. [Gretton et al., 2007, Theorem 3] Let $\mathcal{F}$ be a unit ball in a universal RKHS $\mathcal{H}$, on compact metric space $\mathcal{X}$ with the associated universal kernel k , $MMD[\mathcal{F}, P, Q] = 0 \iff P = Q$.

Though Steinwart [2001] established that Gaussian and Laplacian kernels on compact domains are universal, the application is restrictive as many random variables are defined on non-compact domains. Fukumizu et al. [2008] introduced a related concept called the characteristic kernel. These are the kernels for which the mean embedding maps are injective.

Definition 2.3.3. The kernel k is said to be **characteristic** if the condition $\mathbb{E}_P[f(X)] = \mathbb{E}_Q[f(X)]$ for all $f \in \mathcal{H}$ implies $P = Q$.

Hence, when the mean embedding is mapped into RKHSs with characteristic kernels, $MMD[\mathcal{F}, P, Q] = 0$ becomes a metric and preserves all information of the probability distribution. We call such RKHSs associated with characteristic kernels, characteristic RKHSs. In fact, the concept of a characteristic kernel is a generalisation of a characteristic function [Fukumizu et al., 2008, Sriperumbudur et al., 2010]. Recall, the characteristic function is defined as $\mathbb{E}(e^{it^\top X})$ for $t \in \mathbb{R}$ and $i = \sqrt{-1}$. It can be viewed as the kernel mean embedding of probability distributions using the Fourier kernel $k(x, t) = e^{it^\top x}$. As the characteristic function uniquely determines a Borel probability distribution on $\mathbb{R}^D$ [Dudley, 2002, Theorem 9.5.1], i.e. $\mathbb{E}(e^{it^\top X}) = \mathbb{E}(e^{it^\top Y})$ iff $P = Q$, then $\mathbb{E}(e^{it^\top X}) = \mathbb{E}(e^{it^\top Y})$ iff $\mu_P = \mu_Q$. Such concept allows a larger family of kernels to be used for practical purposes.

For a compact metric space, the universal RKHS is dense in $L^2(P)$ for any probability P , and hence Lemma 1 of Fukumizu et al. [2008] guarantees that the RKHS is in fact characteristic (i.e. universal kernels are characteristic). Further, Fukumizu et al. [2008][Theorem 2] shows that the Gaussian and Laplacian kernels are characteristic on the entire domain of $\mathbb{R}^D$.

The conditions for characteristic RKHSs requires checking the denseness of the relevant spaces [Fukumizu et al., 2008, Gretton et al., 2007], which can be difficult to verify in practise. Sriperumbudur et al. [2010] presents a simpler characterisation of characteristic kernels when the kernel is translation invariant on $\mathbb{R}^D$. More specifically, they showed that a bounded continuous translation invariant kernel on $\mathbb{R}^D$ is characteristic if and only if the support of the Fourier transform of the kernel is the

entire domain. We refer the interested reader to the original paper for more details.

2.3.2 Relationship Between MMD and Other Discrepancy Measures

In the statistics community, two-sample testing in the Euclidean space can be achieved by the use of energy distance [Székely and Rizzo, 2004, 2005, Baringhaus and Franz, 2004] as a test statistic. The energy distance is a measure of statistical distance between two probability distributions P and Q with finite first moments. It has a compact representation in terms of certain expectations of pairwise Euclidean distances. Sejdinovic et al. [2013b] showed that the RKHS-based discrepancy measures, i.e. the maximum mean discrepancies (MMD) are the formal extensions of the energy distances i.e. that the energy distance can be viewed as MMD for a particular form of a kernel (covariance function of Brownian motion).

In the machine learning community, there are a few classes of well known discrepancy measures between two probability distributions. Kullback-Leibler divergence, Hellinger distance and total variation distance all belong to the class of f-divergences [Csiszár, 1963, Morimoto, 1963, Ali and Silvey, 1966]. For two probability distributions P and Q over a space Ω the f-divergence is defined as

$$D_f(P, Q) := \int_{\Omega} f\left(\frac{dP}{dQ}\right) dQ \quad (2.35)$$

where Ω is a measurable space, $f : [0, \infty) \rightarrow (-\infty, \infty]$ is a convex function and the probability distribution P is absolutely continuous w.r.t. Q . The well-known divergence measures can be obtained when setting f to the appropriate functions. For example, $f(t) = t \log(t)$ gives rise to the Kullback-Leibler (KL) divergence; $f(t) = \frac{1}{2}|t - 1|$ gives rise to the total variation distance; $f(t) = (\sqrt{t} - 1)^2, 2(1 - \sqrt{t})$ gives rise to the Hellinger distance [Liese and Vajda, 2006].

MMD on the other hand, falls into another popular family of distance measures on probabilities, i.e. the integral probability metric, which were already studied by Müller [1997]. This is defined as

$$\gamma_{\mathcal{F}}(P, Q) := \sup_{f \in \mathcal{F}} \left| \int_{\Omega} f dP - \int_{\Omega} f dQ \right| \quad (2.36)$$

where $\mathcal{F}$ is a class of real-valued bounded measurable functions on Ω . Different classes of functions give rise to different distance measures, which includes the well-known Wasserstein distance, Dudley metric and Fortet-Mourier metric. The RKHS distance between probability distributions, i.e. the maximum mean discrepancy (MMD) [Gretton et al., 2012a, Definition 2] as we introduced in Definition 2.3.1, is an integral probability metric $\gamma_{\mathcal{F}}$ with $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$, where $\|\cdot\|_{\mathcal{H}}$ denotes the Hilbert Schmidt norm w.r.t. the RKHS $\mathcal{H}$.

Sriperumbudur et al. [2012] showed that the total variation distance is an integral probability metric $\gamma_{\mathcal{F}}$ with $\mathcal{F} = \{f : \|f\|_{\infty} \leq 1\}$ and it is the only “non-trivial”⁵ integral probability metric that is also an f-divergence. In other words, the overlap between the class of integral probability metric and the class of f-divergence only contains the total-variation distance, hence the two classes of discrepancy measures are fundamentally different. We refer the interesting readers to Sriperumbudur et al. [2012] for the detailed study of the topic.

The empirical estimators of the Kantorovich distance, Dudley metric and the total variation distance can be obtained by solving linear programs [Sriperumbudur et al., 2012], while that of the RKHS distance can be computed in closed form. This justifies the popular and successful applications of MMD in two-sample testing and hence motivates the use in this thesis for the work on Bayesian two-sample test in Chapter 6.

2.4 The Hilbert Schmidt Independence Criterion

The notion of MMD can be used as a measure of independence when comparing the product of the marginal distributions of X and Y and the joint distribution of X and Y . Such measure is known as the Hilbert Schmidt independence criterion (HSIC) [Gretton et al., 2005a]. Note that, we need to work with the joint space $\mathcal{X} \times \mathcal{Y}$ in measuring dependencies. For any kernels $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ on the respective domains $\mathcal{X}$ and $\mathcal{Y}$, $k = k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$ given by

$$k((x, y), (x', y')) = k_{\mathcal{X}}(x, x')k_{\mathcal{Y}}(y, y') \quad (2.37)$$

⁵A trivial choice of $\mathcal{F}$ would be the set of all real-valued measurable functions on Ω and the corresponding f function is $f(t) = 0$ when $t = 1$ and $f(t) = +\infty$ when $t \neq 1$ which yields $\gamma_{\mathcal{F}}(P, Q) = D_f(P, Q) = 0$ if $P = Q$ and ∞ otherwise.

is a valid kernel on the product space $\mathcal{X} \times \mathcal{Y}$ [Steinwart and Christmann, 2008, Lemma 4.6]. Its canonical feature map $\varphi_{x,y} = k_{\mathcal{X}}(\cdot, x) \otimes k_{\mathcal{Y}}(\cdot, y)$ is understood as a function on $\mathcal{X} \times \mathcal{Y}$, i.e. $\varphi_{x,y}(x', y') = k_{\mathcal{X}}(x', x)k_{\mathcal{Y}}(y', y)$. The RKHS $\mathcal{H}_k$ of $k = k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$ is isometrically isomorphic to the tensor product space $\mathcal{H}_{k_{\mathcal{X}}} \otimes \mathcal{H}_{k_{\mathcal{Y}}}$ [Sejdinovic et al., 2013b, Section 3.3]. We define a RKHS-based measure of dependence between the random variables X and Y following Section 2.3 of Smola et al. [2007].

Definition 2.4.1. [Smola et al., 2007, Section 2.3] Let X and Y be random variables on non-empty topological spaces $\mathcal{X}$ and $\mathcal{Y}$. Let $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ be kernels separately defined on $\mathcal{X}$ and $\mathcal{Y}$ with the respective RKHS $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$. **Hilbert-Schmidt Independence Criterion (HSIC)** $\Xi_{k_{\mathcal{X}}, k_{\mathcal{Y}}}(X, Y)$ of X and Y is the MMD between the joint measure P_{XY} and the product of marginals $P_X P_Y$, computed with the product kernel $k = k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$, i.e.,

$$\Xi_{k_{\mathcal{X}}, k_{\mathcal{Y}}}(X, Y) = \|\mathbb{E}_{XY}[k_{\mathcal{X}}(\cdot, X) \otimes k_{\mathcal{Y}}(\cdot, Y)] - \mathbb{E}_X k_{\mathcal{X}}(\cdot, X) \otimes \mathbb{E}_Y k_{\mathcal{Y}}(\cdot, Y)\|_{\mathcal{H}_{k_{\mathcal{X}}} \otimes \mathcal{H}_{k_{\mathcal{Y}}}}^2. \quad (2.38)$$

Section 3.3 of Sejdinovic et al. [2013b] shows that HSIC is well defined whenever $P_X \in \mathcal{M}_{k_{\mathcal{X}}}^1(\mathcal{X})$ and $P_Y \in \mathcal{M}_{k_{\mathcal{Y}}}^1(\mathcal{Y})$. Recall from Lemma 2.2.1, that a sufficient condition for the existence of an embedding of a probability measure ν defined on $\mathcal{Z}$ is $\nu \in \mathcal{M}_k^{1/2}(\mathcal{Z})$. Consider the integral

$$\int k^{\frac{1}{2}}((x, y), (x, y)) dP_{XY}((x, y)) \quad (2.39)$$

$$= \int k_{\mathcal{X}}^{\frac{1}{2}}(x, x) k_{\mathcal{Y}}^{\frac{1}{2}}(y, y) dP_{XY}((x, y)) \quad (2.40)$$

$$\leq \left\{ \int k_{\mathcal{X}}(x, x) dP_X(x) \int k_{\mathcal{Y}}(y, y) dP_Y(y) \right\}^{\frac{1}{2}}, \quad (2.41)$$

where the last line is obtained by Cauchy Schwartz inequality. When $P_X \in \mathcal{M}_{k_{\mathcal{X}}}^1(\mathcal{X})$ and $P_Y \in \mathcal{M}_{k_{\mathcal{Y}}}^1(\mathcal{Y})$,

$$\int k_{\mathcal{X}}(x, x) dP_X(x) \int k_{\mathcal{Y}}(y, y) dP_Y(y) < \infty.$$

Hence, $\int k^{\frac{1}{2}}((x, y), (x, y)) dP_{XY}((x, y)) < \infty$ and $P_{XY} \in \mathcal{M}_{k_{\mathcal{X}} \otimes k_{\mathcal{Y}}}^{1/2}(\mathcal{X} \times \mathcal{Y})$. This implies the existence of the kernel mean embedding of the joint measure $\mu(P_{XY})$. Note that such conditions on the marginal distributions also ensure the existence of $\mu(P_X P_Y)$ (equivalently, $\mu(P_X) \otimes \mu(P_Y)$) because the existence of the marginal kernel mean embeddings $\mu(P_X)$ and $\mu(P_Y)$ are guaranteed by $P_X \in \mathcal{M}_{k_{\mathcal{X}}}^1(\mathcal{X}) \subset \mathcal{M}_{k_{\mathcal{X}}}^{\frac{1}{2}}(\mathcal{X})$

and $P_Y \in \mathcal{M}_{k_Y}^1(\mathcal{Y}) \subset \mathcal{M}_{k_Y}^{\frac{1}{2}}(\mathcal{Y})$.

The name HSIC comes from the operator view of the RKHS $\mathcal{H}_{k_X \otimes k_Y}$. More specifically, the difference between embeddings $\mathbb{E}_{XY}[k_X(\cdot, X) \otimes k_Y(\cdot, Y)] - \mathbb{E}_X k_X(\cdot, X) \otimes \mathbb{E}_Y k_Y(\cdot, Y)$ can be identified with the cross-covariance operator $\Sigma_{XY} : \mathcal{H}_{k_Y} \rightarrow \mathcal{H}_{k_X}$ for which

$$\langle f, \Sigma_{XY} g \rangle_{\mathcal{H}_{k_X}} = \text{Cov}[f(X)g(Y)],$$

$\forall f \in \mathcal{H}_{k_X}, g \in \mathcal{H}_{k_Y}$ [Gretton et al., 2005a, 2008]. The squared Hilbert-Schmidt norm of this operator $\|\Sigma_{XY}\|_{HS}^2$ is exactly the HSIC. In the sequel, we will suppress dependence on kernels k_X and k_Y in notation $\Xi_{k_X, k_Y}(X, Y)$ where there is no ambiguity.

The definition presented in (2.38) may not always be easy to work with. Repeated application of the reproducing property gives the following equivalent representation of HSIC [Smola et al., 2007, p.10]:

Proposition 2.4.1. *The HSIC of X and Y can be written as:*

$$\begin{aligned} \Xi(X, Y) &= \mathbb{E}_{XY} \mathbb{E}_{X'Y'} k_X(X, X') k_Y(Y, Y') \\ &\quad + \mathbb{E}_X \mathbb{E}_{X'} k_X(X, X') \mathbb{E}_Y \mathbb{E}_{Y'} k_Y(Y, Y') \\ &\quad - 2 \mathbb{E}_{X'Y'} [\mathbb{E}_X k_X(X, X') \mathbb{E}_Y k_Y(Y, Y')]. \end{aligned} \tag{2.42}$$

To summaries, in this chapter, we have reviewed key concepts in kernel methods that are needed in the subsequent chapters. We have provided definitions and properties of the reproducing kernel Hilbert spaces (RKHS) and the corresponding kernel. We have then reviewed the formal definition of the kernel mean embeddings, which the nonparametric testing framework relies on to represent probability distributions with elements of a RKHS and provided the conditions on probability distributions such that the kernel mean embedding exists. Following this, we have introduced the important kernel discrepancy measure between probability distributions in the RKHS, the Maximum Mean Discrepancy (MMD), which we will use in Chapter 6. We have also reviewed the concept of universal and characteristic kernels which ensures the kernel mean embedding is injective and hence each probability distribution can be uniquely represented in the RKHS with all statistical features preserved. The relationships between MMD and other classes of well-known discrepancy measures between probability distributions, the class of f-divergence and the family of integral probability metric, were also discussed. Finally, we have presented the import kernel dependency measure, the Hilbert Schmidt independence criterion (HSIC), which we will use in Chapter 3 and Chapter 5.

Chapter 3

Large-Scale Kernel Methods for Independence Testing

This chapter consists of a self-contained paper:

Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic, Large-Scale Kernel Methods for Independence Testing, Statistics and Computing, vol. 28, no. 1, 113–130, Jan. 2018.

Abstract

Representations of probability measures in reproducing kernel Hilbert spaces provide a flexible framework for fully nonparametric hypothesis tests of independence, which can capture any type of departure from independence, including nonlinear associations and multivariate interactions. However, these approaches come with an at least quadratic computational cost in the number of observations, which can be prohibitive in many applications. Arguably, it is exactly in such large-scale datasets that capturing *any type* of dependence is of interest, so striking a favourable tradeoff between computational efficiency and test performance for kernel independence tests would have a direct impact on their applicability in practice. In this contribution, we provide an extensive study of the use of large-scale kernel approximations in the context of independence testing, contrasting block-based, Nyström and random Fourier feature approaches. Through a variety of synthetic data experiments and one real data experiment, it is demonstrated that our large scale methods give comparable performance with existing methods whilst using significantly less computation time and memory.

3.1 Introduction

Given a paired sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ with each $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ independently and identically following the joint distribution P_{XY} on some generic domains $\mathcal{X}$ and $\mathcal{Y}$, the nonparametric independence problem consists of testing if we should reject the null hypothesis $\mathcal{H}_0 : P_{XY} = P_X P_Y$ in favour of the general alternative hypothesis $\mathcal{H}_1 : P_{XY} \neq P_X P_Y$, where P_X and P_Y are the marginal distributions for X and Y respectively. This problem is fundamental and extensively studied, with wide-ranging applications in statistical inference and modelling. Classical dependence measures, such as Pearson’s product-moment correlation coefficient, Spearman’s ρ , Kendall’s τ or methods based on contingency tables are typically designed to capture only particular forms of dependence (e.g. linear or monotone). Furthermore, they are applicable only to scalar random variables or require space partitioning limiting their use to relatively low dimensions. As availability of larger datasets also facilitates building more complex models, dependence measures are sought that capture more complex dependence patterns and those that occur between multivariate and possibly high-dimensional datasets. In this light, among the most popular dependence measures recently have been those based on characteristic functions [Székely et al., 2007, Székely and Rizzo, 2009] as well as a broader framework based on kernel methods [Gretton et al., 2005a, 2008]. A desirable property of consistency against any alternative - i.e. test power provably increasing to one with the sample size regardless of the form of dependence, is warranted for statistical tests based on such approaches. However, this is achieved at an expense of computational and memory requirements that increase at least quadratically with the sample size, which is prohibitive for many modern applications. Thus, a natural question is whether a favourable trade-off between computational efficiency and test power can be sought with appropriate large-scale approximations. As we demonstrate, several large-scale approximations are available in this context and they lead to strong improvements in power-per-computational unit performance, resulting in a fast and flexible independence testing framework responsive to all forms of dependence *and* applicable to large datasets.

The key quantity we consider is the Hilbert-Schmidt Independence Criterion (HSIC) introduced by Gretton et al. [2005a]. HSIC uses the distance between the kernel embeddings of probability measures in the Reproducing Kernel Hilbert Space (RKHS) [Gretton et al., 2008, Zhang et al., 2011, Smola et al., 2007]. By building on decades of research into kernel methods for machine learning [Schölkopf and Smola,

2002], HSIC can be applied to multivariate observations as well as to those lying in non-Euclidean and structured domains, e.g., Gretton et al. [2008] considers independence testing on text data. HSIC has also been applied to clustering and learning taxonomies [Song et al., 2007, Blaschko and Gretton, 2009], feature selection [Song et al., 2012], causal inference [Peters et al., 2014, Flaxman et al., 2015, Zaremba and Aste, 2014] and computational linguistics [Nguyen and Eisenstein, 2017]. A closely related dependence coefficient that measures all types of dependence between two random vectors of arbitrary dimensions is the distance covariance (dCov) of Székely et al. [2007], Székely and Rizzo [2009], which measures distances between empirical characteristic functions or equivalently measures covariances with respect to a stochastic process [Székely and Rizzo, 2009], and its normalised counterpart, distance correlation (dCor). RKHS-based dependence measures like HSIC are in fact extensions of dCov – Sejdinovic et al. [2013b] showed that dCov can be understood as a form of HSIC with a particular choice of kernel. Moreover, dCor can be viewed as an instance of kernel matrix alignment of Cortes et al. [2012]. As we will see, statistical tests based on estimation of HSIC and dCov are computationally expensive and require at least $\mathcal{O}(m^2)$ time and storage complexity, where m is the number of observations, just to compute an HSIC estimator which serves as a test statistic. In addition, the complicated form of the asymptotic null distribution of the test statistics necessitates either permutation testing [Arcones and Gine, 1992] (further increasing the computational cost) or even more costly direct sampling from the null distribution, requiring eigen-decompositions of kernel matrices using the spectral test of Gretton et al. [2009], with a cost of $\mathcal{O}(m^3)$ ¹. These memory and time requirements often make the HSIC-based tests infeasible for practitioners.

In this paper, we consider several ways to speed up the computation in HSIC-based tests. More specifically, we introduce three fast estimators of HSIC: the block-based estimator, the Nyström estimator and the random Fourier feature (RFF) estimator and study the resulting independence tests. In the block-based setting, we obtain a simpler asymptotic null distribution as a consequence of the Central Limit Theorem in which only asymptotic variance needs to be estimated - we discuss possible approaches for this. RFF and Nyström estimators correspond to the primal finite-dimensional approximations of the kernel functions and as such also warrant estimation of the

¹An alternative approach applicable to scalar variables uses kernel-based methods combined with a copula transformation [Póczos et al., 2012] can be used to tabulate the null distribution, but this approach is not straightforward to generalise to modelling dependence between random vectors.

null distribution in linear time – we introduce a novel spectral tests based on eigendecompositions of primal covariance matrices, which avoid permutation approach and significantly reduce the computational expense for the direct sampling from the null distribution.

Related Work

Some of the approximation methods considered in this paper were inspired by their use in a related context of two-sample testing. In particular, the block-based approach for two-sample testing was studied in Gretton et al. [2012a,b], Zaremba et al. [2013] under the name of linear-time MMD (Maximum Mean Discrepancy), i.e. the distance between the mean embeddings of the probability distributions in the RKHS. The approach estimates MMD on a small block of data and then averages the estimates over blocks to obtain the final test statistic. Our block-based estimator of HSIC follows exactly the same strategy. On the other hand, the Nyström method [Williams and Seeger, 2001, Snelson and Ghahramani, 2006] is a classical low-rank kernel approximation technique, where data is projected into lower-dimensional subspaces of RKHS (spanned by so called *inducing variables*). Such an idea is popular in fitting sparse approximations to Gaussian process (GP) regression models, allowing reduction of the computational cost from $\mathcal{O}(m^3)$ to $\mathcal{O}(n^2m)$ where $n \ll m$ is the number of inducing variables. To the best of our knowledge, Nyström approximation was not studied in the context of hypothesis testing. Random Fourier feature (RFF) approximations [Rahimi and Recht, 2007], however, due to their relationship with evaluations of empirical characteristic functions, do have a rich history in the context of statistical testing – as discussed in Chwialkowski et al. [2015], which also proposes an approach to scale up kernel-based two sample tests by additional smoothing of characteristic functions, thereby improving the test power and its theoretical properties. Moreover, the approximation strategy of MMD and two-sample testing through primal representation using RFF have also been studied in Zhao and Meng [2015], Sutherland and Schneider [2015], Lopez-Paz [2016]. In addition, Lopez-Paz et al. [2013] first proposed the idea of applying RFF in order to construct an approximation to a kernel-based dependence measure. More specifically, they develop Randomised Canonical Correlation Analysis (RCCA) (see also Lopez-Paz et al. [2014], Lopez-Paz [2016]) approximating the non-linear kernel-based generalisation of the Canonical Correlation Analysis [Lai and Fyfe, 2000, Bach and Jordan, 2002] and using a further copula transformation, construct a test statistic termed RDC (randomised dependence coefficient) requiring $\mathcal{O}(m \log m)$ time to compute. Under suitable assumptions, Bartlett’s approximation

[Mardia et al., 1979] provides a closed form expression for the asymptotic null distribution of this statistic which further results in a distribution free test, leading to an attractive option for large-scale independence testing. We extend these ideas based on RFF to construct approximations of HSIC and dCov/dCor, which are conceptually distinct kernel-based dependence measures from that of kernel CCA, i.e., they measure different types of norms of RKHS operators (operator norm vs Hilbert-Schmidt norm).

In fact, the Nyström and RFF approximations can also be viewed through the lens of nonlinear canonical analysis framework introduced by Dauxois and Nkiet [1998]. This is the earliest example we know where nonlinear dependence measures based on spectra of appropriate Hilbert space operators are studied. In particular, the cross-correlation operator with respect to a dictionary of basis functions in L_2 (e.g. B-splines) is considered in Dauxois and Nkiet [1998]. Huang et al. [2009] links this framework to the RKHS perspective. The functions of the spectra that were considered in Dauxois and Nkiet [1998] are very general, but the simplest one (sum of the squared singular values) can be recast as the NOrmalised Cross-Covariance Operator (NOCCO) of Fukumizu et al. [2008], which considers the Hilbert-Schmidt norm of the cross-correlation operator on RKHSs and as such extends kernel CCA to consider the entire spectrum. While in this work we focus on HSIC (Hilbert-Schmidt norm of the *cross-covariance operator*), which is arguably the most popular kernel dependence measure in the literature, a similar Nyström or RFF approximation can be applied to NOCCO as well - we leave this as a topic for future work.

The paper is structured as follows: in Section 3.2, we first provide some necessary definitions from the RKHS theory and review the aforementioned Hilbert-Schmidt Independence Criterion (HSIC) and discuss its biased and unbiased quadratic time estimators. Then, Section 3.2.3 gives the asymptotic null distributions of estimators (proofs provided in Section 3.7). In Section 3.3, we develop a block-based HSIC estimator and derive its asymptotic null distribution. Following this, a linear time asymptotic variance estimation approach is proposed. In Section 3.4.1 and 3.4.2, we propose Nyström HSIC and RFF HSIC estimator respectively, both with the corresponding linear time null distribution estimation approaches. Finally, in Section 3.5, we explore the performance of the three testing approaches on a variety of challenging synthetic data and one real data experiment.

3.2 Background

This section starts with a brief overview of the key concepts and notations required to understand the RKHS theory and kernel embeddings of probability distributions into the RKHS. It then provides the definition of HSIC which will serve as a basis for later independence tests. We review the quadratic time biased and unbiased estimators of HSIC as well as their respective asymptotic null distributions. As the final part of this section, we outline the construction of independence tests in quadratic time.

3.2.1 RKHS and Embeddings of Measures

Let $\mathcal{Z}$ be any topological space on which Borel measures can be defined. By $\mathcal{M}(\mathcal{Z})$ we denote the set of all finite signed Borel measures on $\mathcal{Z}$ and by $\mathcal{M}_+^1(\mathcal{Z})$ the set of all Borel probability measures on $\mathcal{Z}$. We will now review the basic concepts of RKHS and kernel embeddings of probability measures. For further details, see Berlinet and Thomas-Agnan [2004], Steinwart and Christmann [2008], Sriperumbudur [2010].

Definition 3.2.1. *Let $\mathcal{H}$ be a Hilbert space of real-valued function defined on $\mathcal{Z}$. A function $k : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$ is called a **reproducing kernel** of $\mathcal{H}$ if:*

1. $\forall z \in \mathcal{Z}, k(\cdot, z) \in \mathcal{H}$
2. $\forall z \in \mathcal{Z}, \forall f \in \mathcal{H}, \langle f, k(\cdot, z) \rangle_{\mathcal{H}} = f(z)$.

*If $\mathcal{H}$ has a reproducing kernel, it is called a **Reproducing Kernel Hilbert Space (RKHS)**.*

As a direct consequence, for any $x, y \in \mathcal{Z}$,

$$k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}}. \quad (3.1)$$

In machine learning literature, a notion of *kernel* is understood as an inner product between feature maps [Steinwart and Christmann, 2008]. By (3.1), every reproducing kernel is a kernel in this sense, corresponding to a *canonical feature map* $x \mapsto k(\cdot, x)$.

For $x, y \in \mathbb{R}^p$, some examples of reproducing kernels are

- Linear kernel: $k(x, y) = x^T y$;
- Polynomial kernel of degree $d \in \mathbb{N}$: $k(x, y) = (x^T y + 1)^d$;

- Gaussian kernel with bandwidth $\sigma > 0$: $k(x, y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2})$;
- Fractional Brownian motion covariance kernel with parameter $h \in (0, 1)$:
 $k(x, y) = \frac{1}{2} (\|x\|^{2h} + \|y\|^{2h} - \|x - y\|^{2h})$.

Checking whether a given function k is a valid reproducing kernel can be onerous. Fortunately, the Moore-Aronszajn theorem [Aronszajn, 1950] gives a simple characterisation: for any symmetric, positive definite function $k : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$, there exists a unique Hilbert space of functions $\mathcal{H}$ defined on $\mathcal{Z}$ such that k is reproducing kernel of $\mathcal{H}$ [Berlinet and Thomas-Agnan, 2004]. RKHS are precisely the space of functions where norm convergence implies pointwise convergence and are as a consequence relatively well behaved comparing to other Hilbert spaces. In nonparametric testing, as we consider here, a particularly useful setup will be representing probability distributions and, more broadly, finite signed Borel measures $\nu \in \mathcal{M}(\mathcal{Z})$ with elements of an RKHS [Smola et al., 2007].

Definition 3.2.2. *Let k be a kernel on $\mathcal{Z}$, and $\nu \in \mathcal{M}(\mathcal{Z})$. The **kernel embedding** of measure ν into the RKHS $\mathcal{H}_k$ is $\mu_k(\nu) \in \mathcal{H}_k$ such that*

$$\int f(z) d\nu(z) = \langle f, \mu_k(\nu) \rangle_{\mathcal{H}_k}, \quad \forall f \in \mathcal{H}_k. \quad (3.2)$$

It is understood from this definition that the integral of any RKHS function f under the measure ν can be evaluated as the inner product between f and the kernel embedding $\mu_k(\nu)$ in the RKHS $\mathcal{H}_k$. As an alternative, the kernel embedding can be defined through the use of Bochner integral $\mu_k(\nu) = \int k(\cdot, z) d\nu(z)$. Any probability measure is mapped to the corresponding expectation of the canonical feature map. By Cauchy-Schwarz inequality and the Riesz representation theorem, a sufficient condition for the existence of an embedding of ν is that $\nu \in \mathcal{M}_k^{1/2}(\mathcal{Z})$, where we adopt notation from Sejdinovic et al. [2013b]: $\mathcal{M}_k^\theta(\mathcal{Z}) = \{\nu \in \mathcal{M}(\mathcal{Z}) : \int k^\theta(z, z) d|\nu|(z) < \infty\}$, which is, e.g. satisfied for all finite measures if k is a bounded function (such as Gaussian kernel).

Embeddings allow measuring distances between probability measures, giving rise to the notion of Maximum Mean Discrepancy (MMD) [Borgwardt et al., 2006, Gretton et al., 2012a].

Definition 3.2.3. Let k be a kernel on $\mathcal{Z}$. The squared distance between the kernel embeddings of two probability measures P and Q in the RKHS,

$$\text{MMD}_k(P, Q) = \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k}^2 \quad (3.3)$$

is called **Maximum Mean Discrepancy (MMD)** between P and Q with respect to k .

When the corresponding kernels are *characteristic* [Sriperumbudur, 2010], embedding is injective and MMD is a metric on probability measures. The estimators of MMD are useful statistics in nonparametric two-sample testing [Gretton et al., 2012a], i.e. testing if two given samples are drawn from the same probability distribution. For any kernels $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ on the respective domains $\mathcal{X}$ and $\mathcal{Y}$, it is easy to check that $k = k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$ given by

$$k((x, y), (x', y')) = k_{\mathcal{X}}(x, x')k_{\mathcal{Y}}(y, y') \quad (3.4)$$

is a valid kernel on the product domain $\mathcal{X} \times \mathcal{Y}$. Its canonical feature map is $(x, y) \mapsto k_{\mathcal{X}}(\cdot, x) \otimes k_{\mathcal{Y}}(\cdot, y)$ where $\varphi_{x,y} = k_{\mathcal{X}}(\cdot, x) \otimes k_{\mathcal{Y}}(\cdot, y)$ is understood as a function on $\mathcal{X} \times \mathcal{Y}$, i.e. $\varphi_{x,y}(x', y') = k_{\mathcal{X}}(x', x)k_{\mathcal{Y}}(y', y)$. The RKHS of $k = k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$ is in fact isometric to $\mathcal{H}_{k_{\mathcal{X}}} \otimes \mathcal{H}_{k_{\mathcal{Y}}}$, which can be viewed as the space of Hilbert-Schmidt operators between $\mathcal{H}_{k_{\mathcal{Y}}}$ and $\mathcal{H}_{k_{\mathcal{X}}}$ [Steinwart and Christmann, 2008, Lemma 4.6]. We are now ready to define a RKHS-based measure of dependence between random variables X and Y .

Definition 3.2.4. Let X and Y be random variables on domains $\mathcal{X}$ and $\mathcal{Y}$ (non-empty topological spaces). Let $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ be kernels on $\mathcal{X}$ and $\mathcal{Y}$ respectively. **Hilbert-Schmidt Independence Criterion (HSIC)** $\Xi_{k_{\mathcal{X}}, k_{\mathcal{Y}}}(X, Y)$ of X and Y is MMD between the joint measure P_{XY} and the product of marginals $P_X P_Y$, computed with the product kernel $k = k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$, i.e.,

$$\Xi_{k_{\mathcal{X}}, k_{\mathcal{Y}}}(X, Y) = \|\mathbb{E}_{XY}[k_{\mathcal{X}}(\cdot, X) \otimes k_{\mathcal{Y}}(\cdot, Y)] - \mathbb{E}_X k_{\mathcal{X}}(\cdot, X) \otimes \mathbb{E}_Y k_{\mathcal{Y}}(\cdot, Y)\|_{\mathcal{H}_{k_{\mathcal{X}} \otimes k_{\mathcal{Y}}}}^2. \quad (3.5)$$

HSIC is well defined whenever $P_X \in \mathcal{M}_{k_{\mathcal{X}}}^1(\mathcal{X})$ and $P_Y \in \mathcal{M}_{k_{\mathcal{Y}}}^1(\mathcal{Y})$ as this implies $P_{XY} \in \mathcal{M}_{k_{\mathcal{X}} \otimes k_{\mathcal{Y}}}^{1/2}(\mathcal{X} \times \mathcal{Y})$ [Sejdinovic et al., 2013b]. The name of HSIC comes from the operator view of the RKHS $\mathcal{H}_{k_{\mathcal{X}} \otimes k_{\mathcal{Y}}}$. Namely, the difference between embeddings $\mathbb{E}_{XY}[k_{\mathcal{X}}(\cdot, X) \otimes k_{\mathcal{Y}}(\cdot, Y)] - \mathbb{E}_X k_{\mathcal{X}}(\cdot, X) \otimes \mathbb{E}_Y k_{\mathcal{Y}}(\cdot, Y)$ can be identified with the cross-covariance operator $C_{XY} : \mathcal{H}_{k_{\mathcal{Y}}} \rightarrow \mathcal{H}_{k_{\mathcal{X}}}$ for which $\langle f, C_{XY}g \rangle_{\mathcal{H}_{k_{\mathcal{X}}}} = \text{Cov}[f(X)g(Y)]$, $\forall f \in \mathcal{H}_{k_{\mathcal{X}}}, g \in \mathcal{H}_{k_{\mathcal{Y}}}$ [Gretton et al., 2005a, 2008]. HSIC is then simply the squared

Hilbert-Schmidt norm $\|C_{XY}\|_{HS}^2$ of this operator, while distance correlation (dCor) of Székely et al. [2007], Székely and Rizzo [2009] can be cast as $\|C_{XY}\|_{HS}^2 / \|C_{XX}\|_{HS} \|C_{YY}\|_{HS}$ [Sejdinovic et al., 2013b, Appendix A]. In the sequel, we will suppress dependence on kernels k_X and k_Y in notation $\Xi_{k_X, k_Y}(X, Y)$ where there is no ambiguity.

Repeated application of the reproducing property gives the following equivalent representation of HSIC [Smola et al., 2007]:

Proposition 3.2.1. *The HSIC of X and Y can be written as:*

$$\begin{aligned} \Xi(X, Y) &= \mathbb{E}_{XY} \mathbb{E}_{X'Y'} k_X(X, X') k_Y(Y, Y') \\ &\quad + \mathbb{E}_X \mathbb{E}_{X'} k_X(X, X') \mathbb{E}_Y \mathbb{E}_{Y'} k_Y(Y, Y') \\ &\quad - 2 \mathbb{E}_{X'Y'} [\mathbb{E}_X k_X(X, X') \mathbb{E}_Y k_Y(Y, Y')]. \end{aligned} \quad (3.6)$$

3.2.2 Estimation of HSIC

Using the form of HSIC in (3.6), given an iid sample of $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ from the joint distribution P_{XY} , an unbiased estimator of HSIC can be obtained as a sum of three U -statistics [Gretton et al., 2008]:

$$\begin{aligned} \Xi_u(\mathbf{z}) &= \frac{(m-2)!}{m!} \sum_{(i,j) \in \mathbf{i}_2^m} (K_x)_{ij} (K_y)_{ij} \\ &\quad + \frac{(m-4)!}{m!} \sum_{(i,j,q,r) \in \mathbf{i}_4^m} (K_x)_{ij} (K_y)_{qr} \\ &\quad - 2 \frac{(m-3)!}{m!} \sum_{(i,j,q) \in \mathbf{i}_3^m} (K_x)_{ij} (K_y)_{iq}, \end{aligned} \quad (3.7)$$

where the index set $\mathbf{i}_r^m$ denotes the set of all r -tuples drawn without replacement from $\{1, \dots, m\}$, $(K_x)_{ij} := k_X(x_i, x_j)$ and $(K_y)_{ij} := k_Y(y_i, y_j)$. Naïve computation of (3.7) would require $\mathcal{O}(m^4)$ operations. However, an equivalent form which needs $\mathcal{O}(m^2)$ operations is given in Song et al. [2012] as

$$\Xi_u(\mathbf{z}) = \frac{1}{m(m-3)} \left[\text{tr}(\tilde{K}_x \tilde{K}_y) + \frac{\mathbf{1}^T \tilde{K}_x \mathbf{1} \mathbf{1}^T \tilde{K}_y \mathbf{1}}{(m-1)(m-2)} - \frac{2}{m-2} \mathbf{1}^T \tilde{K}_x \tilde{K}_y \mathbf{1} \right] \quad (3.8)$$

where $\tilde{K}_x = K_x - \text{diag}(K_x)$ (i.e. the kernel matrix with diagonal elements set to zero) and similarly for $\tilde{K}_y$. $\mathbf{1}$ is a vector of 1s of relevant dimension.

We will refer to the above as the quadratic time estimator. Gretton et al. [2008] note that the V -statistic estimator (or the quadratic time biased estimator) of HSIC can be an easier-to-use alternative for the purposes of independence testing, since the bias is accounted for in the asymptotic null distribution. The V -statistic is given by

$$\Xi_b(\mathbf{z}) = \frac{1}{m^2} \sum_{i,j}^m (K_x)_{ij} (K_y)_{ij} + \frac{1}{m^4} \sum_{i,j,q,r}^m (K_x)_{ij} (K_y)_{qr} - 2 \frac{1}{m^3} \sum_{i,j,q}^m (K_x)_{ij} (K_y)_{iq},$$

where the summation indices are now drawn with replacement. Further, it can be simplified as follows to reduce the computation:

$$\Xi_b(\mathbf{z}) = \frac{1}{m^2} \text{Trace}(K_x H K_y H) = \frac{1}{m^2} \langle H K_x H, H K_y H \rangle \quad (3.9)$$

where $H = I_m - \frac{1}{m} \mathbf{1}\mathbf{1}^T$ is an $m \times m$ centering matrix. (3.9) gives an intuitive understanding of the HSIC statistic: it measures average similarity between the centered kernel matrices, which are in turn similarity patterns within the samples.²

3.2.3 Asymptotic Null Distribution of Estimators

The asymptotic null distribution of the biased HSIC statistic defined in (3.9) computed using a given data set converges in distribution in Theorem 3.2.1 below. This asymptotic distribution builds the theoretical foundation for the spectral testing approach (described in Section 3.2.4.2) that we will use throughout the paper.

Theorem 3.2.1. (Asymptotic Null Distribution of the Biased HSIC) *Under the null hypothesis, let the dataset $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \stackrel{i.i.d.}{\sim} P_{XY} = P_X P_Y$, with $P_X \in \mathcal{M}_{k_X}^2(\mathcal{X})$ and $P_Y \in \mathcal{M}_{k_Y}^2(\mathcal{Y})$, then*

$$m \Xi_{b,k_X,k_Y}(\mathbf{Z}) \xrightarrow{D} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \lambda_i \eta_j N_{i,j}^2 \quad (3.10)$$

where $N_{i,j} \stackrel{i.i.d.}{\sim} \mathcal{N}(0,1)$, $\forall i, j \in \mathbb{N}$ and $\{\lambda_i\}_{i=1}^{\infty}$, $\{\eta_j\}_{j=1}^{\infty}$ are the eigenvalues of the integral kernel operators $S_{\tilde{k}_{P_X}}$ and $S_{\tilde{k}_{P_Y}}$, where the integral kernel operator $S_{\tilde{k}_P} : L_P^2(\mathcal{Z}) \rightarrow L_P^2(\mathcal{Z})$ is given by

$$S_{\tilde{k}_P} g(z) = \int_{\mathcal{Z}} \tilde{k}_P(z, w) g(w) dP(w). \quad (3.11)$$

²A straightforward estimator of dCor [Székely et al., 2007, Székely and Rizzo, 2009] is then given by normalising $\Xi_b(\mathbf{z})$ by the Frobenius norms of $H K_x H$ and $H K_y H$, i.e., $\widehat{dCor}(\mathbf{z}) = \frac{\langle H K_x H, H K_y H \rangle}{\|H K_x H\|_F \|H K_y H\|_F}$

where $\tilde{k}_P(z, z')$ is the kernel centred at probability measure P :

$$\begin{aligned}\tilde{k}_P(z, z') &:= \langle k(z, \cdot) - \mathbb{E}_W k(W, \cdot), k(z', \cdot) - \mathbb{E}_W k(W, \cdot) \rangle \\ &= k(z, z') + \mathbb{E}_{WW'} k(W, W') - \mathbb{E}_W k(z, W) - \mathbb{E}_W k(z', W),\end{aligned}\quad (3.12)$$

with $W, W' \stackrel{i.i.d.}{\sim} P$.

For completeness, the proof of this theorem, which is a consequence of [Lyons, 2013, Theorem 2.7] and the equivalence between distance-based and RKHS-based independence statistics [Sejdinovic et al., 2013b] is given in the Appendix 3.7. As remarked by Sejdinovic et al. [2013b], the finite marginal moment conditions imply that the integral operators $S_{\tilde{k}_X}$ and $S_{\tilde{k}_Y}$ are trace class and hence Hilbert-Schmidt [Reed and Simon, 1980]. Anderson et al. [1994] noted that the form of the asymptotic distribution of the V -statistic requires the integral operators being trace class but that of the U -statistic only requires them being Hilbert-Schmidt [Anderson et al., 1994, Sejdinovic et al., 2013b]. Using the same notation as in the case of the V -statistic, the asymptotic distribution of the U -statistic in (3.7) can be written as:

$$m\Xi_{u, k_X, k_Y}(\mathbf{Z}) \xrightarrow{D} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \lambda_i \eta_j (N_{i,j}^2 - 1) \quad (3.13)$$

under the null hypothesis.

We note that [Chwialkowski and Gretton, 2014, Lemma 2, Theorem 1] proves a more general result, applicable to dependent observations under certain mixing conditions where the i.i.d. setting is a special case. Moreover, [Rubenstein et al., 2016, Theorem 5, 6] provides another elegant proof in the context of three-variable interaction testing from Sejdinovic et al. [2013a]. However, both Chwialkowski and Gretton [2014] and Rubenstein et al. [2016] assume boundedness of k_X and k_Y , while our proof in the Appendix 3.7 assumes a stronger condition of finite second moments for both k_X and k_Y , thus making the result applicable to unbounded kernels such as the Brownian motion covariance kernel.

Under the alternative hypothesis that $P_X P_Y \neq P_{XY}$, Gretton et al. [2008] remarked that $m\Xi_{b, k_X, k_Y}(\mathbf{Z})$ converges to $HSIC$ with the corresponding appropriately centred and scaled Gaussian distribution as $m \rightarrow \infty$:

$$\sqrt{m}(\Xi_{b, k_X, k_Y}(\mathbf{Z}) - HSIC) \xrightarrow{D} \mathcal{N}(0, \sigma_u^2) \quad (3.14)$$

where the variance $\sigma_u^2 = 16(\mathbb{E}_{z_i}(\mathbb{E}_{z_j, z_q, z_r}(h_{ijqr}))^2 - HSIC)$ and h_{ijqr} is defined as

$$h_{ijqr} = \frac{1}{4!} \sum_{(t,u,v,w)}^{(i,j,q,r)} (K_x)_{tu}(K_y)_{tu} + (K_x)_{tu}(K_y)_{vw} + (K_x)_{tu}(K_y)_{tv} \quad (3.15)$$

with all ordered quadruples (t, u, v, w) drawn without replacement from (i, j, q, r) and assuming $\mathbb{E}(h^2) < \infty$. In fact, under the alternative hypothesis, the difference between $m\Xi_b(\mathbf{Z})$ (i.e. the V -statistic) and the U -statistic drops as $1/m$ and hence asymptotically the two statistics converges to the same distribution [Gretton et al., 2008].

3.2.4 Quadratic Time Null Distribution Estimations

We would like to design independence tests with an asymptotic Type I error of α and hence we need an estimate of the $(1 - \alpha)$ quantile of the null distribution. Here, we consider two frequently used approaches, namely the permutation approach and the spectral approach, that require at least quadratic time both in terms of memory and computation time. The biased V -statistic will be used because of its neat and compact formulation.

3.2.4.1 Permutation Approach

Consider an iid sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ with chosen kernels $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ respectively, the permutation/bootstrap approach [Arcones and Gine, 1992] proceed in the following manner. Suppose the total number of shuffles is fixed at N_p , we first compute $\Xi_{k_{\mathcal{X}}, k_{\mathcal{Y}}}(\mathbf{z})$ using $\mathbf{z}$, $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$. Then, for each shuffle, we fix the $\{x_i\}_{i=1}^m$ and randomly permute the $\{y_i\}_{i=1}^m$ to obtain $\mathbf{z}^* = \{(x_i, y_i^*)\}_{i=1}^m$ and subsequently compute $\Xi_{k_{\mathcal{X}}, k_{\mathcal{Y}}}^*(\mathbf{z}^*)$. The one-sided p-value in this instance is the proportion of HSIC values computed on the permuted data that is greater than or equal to $\Xi_{k_{\mathcal{X}}, k_{\mathcal{Y}}}(\mathbf{z})$.

The computational time is $\mathcal{O}(\text{number of shuffles} \times m^2)$ for this approach, where the number of shuffles determines the extend to which we have explored the sampling distribution. In other words, a small number of shuffles means that we may only obtained realisations from the mode of the distribution and hence the tail structure is not adequately captured. Although a larger number of shuffles ensures the proper exploration of the sampling distribution, the computation cost can be high.

3.2.4.2 Spectral Approach

Theorem 1 of Gretton et al. [2009] have shown that the empirical finite sample estimate of the null distribution converges in distribution to its population counterpart provided that the eigenspectrums $\{\gamma_r\}_{r=1}^{\infty}$ of the integral operator $S_k: L^2_{\theta}(\mathcal{X} \times \mathcal{Y}) \rightarrow L^2_{\theta}(\mathcal{X} \times \mathcal{Y})$ is square root summable, i.e.

$$\sum_{r=1}^{\infty} \sqrt{\gamma_r} = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \sqrt{\lambda_i \eta_j} < \infty.$$

Note, the integral operator S_k is the tensor product of the operators $S_{\tilde{k}_{\mathcal{X}}}$ and $S_{\tilde{k}_{\mathcal{Y}}}$:

$$S_k g(x, y) = \int_{\mathcal{X} \times \mathcal{Y}} \tilde{k}_{\mu}(x, x') \tilde{k}_{\nu}(y, y') g(x', y') d\theta(x', y')$$

and the eigenvalues of this operator is hence the product of the eigenvalues of these two operators.

The spectral approach [Gretton et al., 2009, Zhang et al., 2011] requires that we first calculate the centred Gram matrices $\tilde{K}_X = HK_X H$ and $\tilde{K}_Y = HK_Y H$ for the chosen kernel $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$. Then, we compute the $m\Xi_{b, k_{\mathcal{X}}, k_{\mathcal{Y}}}(\mathbf{z})$ statistic according to (3.9). Next, the spectrums (i.e. eigenvalues) $\{\lambda_i\}_{i=1}^m$ and $\{\eta_i\}_{i=1}^m$ of $\tilde{K}_X$ and $\tilde{K}_Y$ are respectively calculated. The empirical null distribution can be simulated by simulating a large enough i.i.d samples from the standard Normal distribution [Zhang et al., 2011] and then generate the test statistic according to (3.10). Finally, the p-value is computed by calculating the proportion of simulated samples that are greater than or equal to the observed $m\Xi_{b, k_{\mathcal{X}}, k_{\mathcal{Y}}}(\mathbf{z})$ value.

Additionally, Zhang et al. [2011] has provided an approximation to the null distribution with a two-parameter Gamma distribution. Despite the computational advantage of such an approach, the permutation and spectral approaches are still preferred since there is no consistency guarantee for the Gamma distribution approach.

3.3 Block-Based HSIC

The quadratic time test statistics are prohibitive for large dataset as it requires $\mathcal{O}(m^2)$ time in terms of storage and computation. Furthermore, one requires an approximation of the asymptotic null distribution in order to compute the p-value. As we discussed in the previous section, this is usually done by randomly permute the Y observations (i.e. the permutation approach) or by performing an eigen-decomposition

of the centred kernel matrices for X and Y (i.e. the spectral approach). Both approaches are expensive in terms of memory and can be computationally infeasible. In this section, we propose a block-based estimator of HSIC which reduce the computational time to linear in the number of samples. The asymptotic null distribution of this estimator will be shown to have a simple form as a result of the Central Limit Theorem (CLT).

3.3.1 The Block HSIC Statistic

Let us consider that the sample is split into blocks of size $B \ll m$: $\{x_i, y_i\}_{i=1}^m \stackrel{i.i.d.}{\sim} P_{XY}$ becomes $\left\{ \left\{ x_i^{(b)}, y_i^{(b)} \right\}_{i=1}^B \right\}_{b=1}^{m/B}$ (where we assumed for simplicity that m is divisible by B). We follow the approach from Zaremba et al. [2013], Sejdinovic et al. [2014] and extend it to independence testing. We compute the unbiased HSIC statistic (Eq. 3.7) on each block $b \in \{1, \dots, \frac{m}{B}\}$:

$$\hat{\eta}_b = \frac{1}{B(B-3)} \left[\text{tr}(\tilde{K}_x^{(b)} \tilde{K}_y^{(b)}) + \frac{\mathbf{1}^T \tilde{K}_x^{(b)} \mathbf{1} \mathbf{1}^T \tilde{K}_y^{(b)} \mathbf{1}}{(B-1)(B-2)} - \frac{2}{B-2} \mathbf{1}^T \tilde{K}_x^{(b)} \tilde{K}_y^{(b)} \mathbf{1} \right] \quad (3.16)$$

and average them over blocks to establish the block-based estimator for HSIC:

$$\hat{\Xi}_{k_X, k_Y} = \frac{B}{m} \sum_{b=1}^{m/B} \hat{\eta}_b. \quad (3.17)$$

3.3.2 Null Distribution of Block-Based HSIC

For the block HSIC statistic, the asymptotic null distribution is a consequence of the Central Limit Theorem (CLT) under the regime where $m \rightarrow \infty$, $B \rightarrow \infty$ and $\frac{m}{B} \rightarrow \infty$ ³. First of all, note that the linear time test statistic $\hat{\Xi}_{k_X, k_Y}$ is an average of block-based statistics $\hat{\eta}_b$ for $b \in \{1, \dots, \frac{m}{B}\}$ which are independent and identically distributed. Secondly, we recall that $\mathbb{E}(\hat{\eta}_b) = 0$ for $\hat{\eta}_b$ is an unbiased estimator of HSIC. Finally, $\text{Var}(\hat{\Xi}_{k_X, k_Y}) = \frac{B}{m} \text{Var}(\hat{\eta}_b) = \frac{B}{m} \frac{1}{B^2} \text{Var}(W)$ with W being the random variable distributed according to $\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \lambda_i \eta_j (N_{i,j}^2 - 1)$. In the limit as $m \rightarrow \infty$, $B \rightarrow \infty$ and $\frac{m}{B} \rightarrow \infty$:

$$\sqrt{mB} \hat{\Xi}_{k_X, k_Y} \xrightarrow{D} \mathcal{N}(0, \sigma_{k,0}^2). \quad (3.18)$$

³For example, $B = m^\delta$ with $\delta \in (0, 1)$

where the variance $\sigma_{k,0}^2$ is the variance of the null distributions in Expression (3.10) and (3.13) i.e. the variance of W and it is given by

$$\sigma_{k,0}^2 = 2 \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \lambda_i^2 \eta_j^2 \quad (3.19)$$

$$= 2\mathbb{E}_{XX'}(\tilde{k}_{P_X}^2(X, X'))\mathbb{E}_{YY'}(\tilde{k}_{P_Y}^2(Y, Y')). \quad (3.20)$$

3.3.3 Linear Time Null Distribution Estimation

Expression (3.18) guarantees the Gaussianity of the null distribution of the block-based statistic and henceforth makes the computation of p-value straight-forward. We simply return the test statistic $\sqrt{mB} \frac{\hat{\eta}_b}{\sqrt{\hat{\sigma}_{k,0}^2}}$ and compare against the corresponding quantile of $\mathcal{N}(0, 1)$ which is the approach taken in Gretton et al. [2012b], Zaremba et al. [2013], Sejdinovic et al. [2014]. Note that the resulting null distribution is actually a t-distribution but with a very large number of degrees of freedom, which can be treated as a Gaussian distribution.

The difficulty of estimating the null distribution lies in estimating $\sigma_{k,0}^2$. We suggest two ways to estimate such variance [Sejdinovic et al., 2014]: within-block permutation and within-block direct estimation. These two approaches are at most quadratic in B within each block which means that the computational cost of estimating the variance is of the same order as that of computing the statistic itself.

Within-block permutation can be done as follows. Within each block, we compute the test statistic using (3.16). At the same time, we track in parallel a sequence $\hat{\eta}_b^*$ obtained using the same formula but with $\{y_i\}_{i=1}^m$ underwent a permutation. The former is used to calculate the overall block statistics and the latter is used to estimate the null variance $\hat{\sigma}_{k,0}^2 = B^2 \text{Var} \left[\{\hat{\eta}_b^*\}_{b=1}^{m/B} \right]$ as the independence between the samples holds by construction.

Within-block direct estimation can be achieved by using (3.20) and the corresponding unbiased estimates of $\mathbb{E}_{XX'}(\tilde{k}_{P_X}^2(X, X'))$ and $\mathbb{E}_{YY'}(\tilde{k}_{P_Y}^2(Y, Y'))$ which can be calculated as follows. For X , the estimate of the variance for each block is given by Song et al. [2012]:

$$(\hat{\sigma}_{k,x}^2)^{(b)} = \frac{2}{B(B-3)} \left[\text{tr}(\tilde{K}_x^{(b)} \tilde{K}_x^{(b)}) + \frac{(\mathbf{1}^T \tilde{K}_x^{(b)} \mathbf{1})^2}{(B-1)(B-2)} - \frac{2}{B-2} \mathbf{1}^T (\tilde{K}_x^{(b)})^2 \mathbf{1} \right] \quad (3.21)$$

Then, we compute

$$\hat{\sigma}_{k,x}^2 = \frac{B}{m} \sum_{b=1}^{m/B} (\hat{\sigma}_{k,x}^2)^{(b)}. \quad (3.22)$$

to obtain an unbiased estimate for $\mathbb{E}_{X,X'}(\tilde{k}_{P_X}^2(X, X'))$. Similarly, replacing all x with y , we obtain an unbiased estimate for $\mathbb{E}_{Y,Y'}(\tilde{k}_{P_Y}^2(Y, Y'))$. The estimate of the variance is therefore:

$$\hat{\sigma}_{k,0}^2 = 2\hat{\sigma}_{k,x}^2\hat{\sigma}_{k,y}^2 \quad (3.23)$$

Such block-based approach is trivially parallelisable: the HSIC statistic defined in (3.16) as well as the variance estimator of (3.21) can be computed separately for each block. This may provide further computational efficiency gains. As remarked in Sejdinovic et al. [2014], we note that under the null hypothesis, the approach undertaken by Zaremba et al. [2013] is to estimate the null variance directly with the empirical variance of $\{\hat{\eta}_b\}_{b=1}^{m/B}$. As the null variance is consistently estimated under the null hypothesis, this ensures the correct level of Type I error. However, without using the variance of the “bootstrap samples”, such an estimate of the null variance will systematically overestimate as B grows with m . Hence, it will result in a reduced statistical power due to inflated p-values.

Regarding the choice of B , Zaremba et al. [2013] discussed that the null distribution is close to that guaranteed by the CLT when B is small and hence the Type I error will be closer to the desired level. However, the disadvantage is the small statistical power for each given sample size. Conversely, Zaremba et al. [2013] pointed out that a larger B results in a lower variance empirical null distribution and hence higher power. Hence, they suggested a sensible family of heuristics is to set $B = \lfloor m^\gamma \rfloor$ for some $0 < \gamma < 1$. As a result, the complexity of the block-based test is $\mathcal{O}(Bm) = \mathcal{O}(m^{1+\gamma})$.

3.4 Approximate HSIC Through Primal Representations

Having discussed how we can construct a linear time HSIC test by processing the dataset in blocks, we now move on to consider how the scaling up can be done through low rank approximations of the Gram matrix. In particular, we will discuss Nyström type approximation (Section 3.4.1) and Random Fourier Features (RFF) type approximation (Section 3.4.2). Both types of approximation act directly on the

primal representation of the kernel hence provide finite representations of the feature maps.

Recall that the definition of HSIC of X and Y in (3.6) can also be written in terms of the cross covariance operator C_{XY} : $\Xi(X, Y) = \|C_{XY}\|_{\mathcal{H}_{k_X} \otimes \mathcal{H}_{k_Y}}^2$ [Gretton et al., 2005a, 2008]. Given a data set $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ with $x_i \in \mathbb{R}^{D_x}$ and $y_i \in \mathbb{R}^{D_y}$ for all i , consider the empirical version of $\Xi_{k_X, k_Y}(X, Y)$ with kernels k_X and k_Y for X and Y respectively:

$$\begin{aligned} \hat{\Xi}_{k_X, k_Y}(X, Y) &= \left\| \frac{1}{m} \sum_{i=1}^m k_X(\cdot, x_i) \otimes k_Y(\cdot, y_i) - \left(\frac{1}{m} \sum_{i=1}^m k_X(\cdot, x_i) \right) \otimes \left(\frac{1}{m} \sum_{i=1}^m k_Y(\cdot, y_i) \right) \right\|^2 \end{aligned} \quad (3.24)$$

$$= \left\| \frac{1}{m} \sum_{i=1}^m \left(k_X(\cdot, x_i) - \frac{1}{m} \sum_{r=1}^m k_X(\cdot, x_r) \right) \otimes \left(k_Y(\cdot, y_i) - \frac{1}{m} \sum_{r=1}^m k_Y(\cdot, y_r) \right) \right\|^2 \quad (3.25)$$

where the Hilbert Schmidt norm is taken in the product space $\mathcal{H}_{k_X} \otimes \mathcal{H}_{k_Y}$. Note, this empirical cross covariance operator is infinite dimensional. However, when approximate feature representations are used, the cross covariance operator is a finite dimensional matrix and hence the Hilbert-Schmidt norm is equivalent to the Frobenius norm (F).

If we let $\bar{\phi}(x_i) = k_X(\cdot, x_i) - \frac{1}{m} \sum_{r=1}^m k_X(\cdot, x_r)$ and $\bar{\psi}(y_i) = k_Y(\cdot, y_i) - \frac{1}{m} \sum_{r=1}^m k_Y(\cdot, y_r)$, the above expression can be further simplified as

$$\begin{aligned} \hat{\Xi}_{k_X, k_Y}(X, Y) &= \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m \langle \bar{\phi}(x_i) \otimes \bar{\psi}(y_i), \bar{\phi}(x_j) \otimes \bar{\psi}(y_j) \rangle \end{aligned} \quad (3.26)$$

$$= \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m \langle \bar{\phi}(x_i), \bar{\phi}(x_j) \rangle \langle \bar{\psi}(y_i), \bar{\psi}(y_j) \rangle \quad (3.27)$$

$$= \frac{1}{m^2} \text{Trace}(H K_x H H K_y H). \quad (3.28)$$

Hence, we obtain the expression in (3.9). If instead, we replace $\bar{\phi}(x_i)$ and $\bar{\psi}(y_i)$ by the corresponding low-rank approximations $\tilde{\phi}(x_i) = \tilde{k}_X(\cdot, x_i) - \frac{1}{m} \sum_{r=1}^m \tilde{k}_X(\cdot, x_r)$ and $\tilde{\psi}(y_i) = \tilde{k}_Y(\cdot, y_i) - \frac{1}{m} \sum_{r=1}^m \tilde{k}_Y(\cdot, y_r)$, we can obtain the approximated HSIC statistic. The details of which are provided in the following sections.

3.4.1 Nyström HSIC

In this section, we use the traditional Nyström approach to provide an approximation that consider the similarities between the so called inducing variables and the given dataset. We will start with a review of Nyström method and then we will provide the explicit feature map representation of the Nyström HSIC estimator. To finish, we will discuss two null distribution estimation approaches that cost at most linear in the number of samples.

The reduced-rank approximation matrix provided by Nyström method [Williams and Seeger, 2001] represents each data point by a vector based on its kernel similarity to the inducing variables and the induced kernel matrix. The approximation is achieved by randomly sample n data points (i.e. inducing variables) from the given m samples and compute the approximate kernel matrix $\hat{K} \approx K$ as:

$$\hat{K} = K_{m,n} K_{n,n}^{-1} K_{n,m} \quad (3.29)$$

where each of $K_{r,s}$ can be think of as the $r \times s$ block of the full Gram matrix K computed using all given samples. Further, we can write (3.29) as:

$$\begin{aligned} \hat{K} &= K_{m,n} K_{n,n}^{-\frac{1}{2}} \left(K_{m,n} K_{n,n}^{-\frac{1}{2}} \right)^T \\ &= \hat{\Phi} \hat{\Phi}^T \end{aligned} \quad (3.30)$$

Hence, an explicit feature representation of $\hat{K}$ is obtained. Note that Snelson and Ghahramani [2006] further relaxed the setting and propose to use inducing points that are not necessarily a subset of the given data but only need to explain the dataset well for a good performance.

3.4.1.1 The Nyström HSIC Statistic

To further reduce computation cost, we propose to approximate this reduced-rank kernel matrix $\hat{K}$ with the $n \times n$ uncentered covariance matrix $\hat{C}$ as the eigenspectrum of the covariance matrix approximates that of the kernel matrix:

$$\begin{aligned} \hat{C} &= \left(K_{m,n} K_{n,n}^{-\frac{1}{2}} \right)^T K_{m,n} K_{n,n}^{-\frac{1}{2}} \\ &= \hat{\Phi}^T \hat{\Phi} \end{aligned} \quad (3.31)$$

Hence the uncentered covariance matrices for X is $\hat{C}_X = \hat{\Phi}_X^T \hat{\Phi}_X$ and the one for Y is $\hat{C}_Y = \hat{\Phi}_Y^T \hat{\Phi}_Y$. In order to approximate the biased HSIC estimator (3.9) using this

explicit feature map representation, the feature maps $\hat{\Phi}_X$ and $\hat{\Phi}_Y$ needed to be centred. We suggest centre each column separately by subtracting its mean for both $\hat{\Phi}_X$ and $\hat{\Phi}_Y$, i.e. denote $\tilde{\Phi}_\cdot = (I_m - \frac{1}{m}\mathbf{1}\mathbf{1}^T)\hat{\Phi}_\cdot = H\hat{\Phi}_\cdot \in \mathcal{R}^{m \times n_\cdot}$ for X and Y respectively.

Using the methods described above, we can substitute approximated kernel functions $\hat{k}_X = \hat{\Phi}_X$ and $\hat{k}_Y = \hat{\Phi}_Y$ into the empirical version of $\Xi_{k_X, k_Y}(X, Y)$ (3.24):

$$\begin{aligned} & \hat{\Xi}_{Ny, \tilde{k}_X, \tilde{k}_Y}(X, Y) \\ &= \left\| \frac{1}{m} \sum_{i=1}^m \hat{\Phi}_X(x_i) \hat{\Phi}_Y(y_i)^T - \left(\frac{1}{m} \sum_{i=1}^m \hat{\Phi}_X(x_i) \right) \left(\frac{1}{m} \sum_{i=1}^m \hat{\Phi}_Y(y_i) \right)^T \right\|_F^2 \end{aligned} \quad (3.32)$$

$$= \left\| \frac{1}{m} \tilde{\Phi}_X^T \tilde{\Phi}_Y \right\|_F^2 \quad (3.33)$$

where $\tilde{\Phi}_X(x_i) \in \mathbb{R}^{n_x}$ and $\tilde{\Phi}_Y(y_i) \in \mathbb{R}^{n_y}$ can both be computed in linear time in m . This is the biased Nyström estimator of HSIC. Essentially, we approximate the cross covariance operator C_{XY} by the Nyström estimator $\tilde{C}_{XY} = \frac{1}{m} \tilde{\Phi}_X^T \tilde{\Phi}_Y \in \mathbb{R}^{n_x \times n_y}$ which only requires $\mathcal{O}(n_x n_y m)$. In essence, the HSIC statistic computed using Nyström as we described here is a HSIC statistic computed using a different kernel. As a remark, we note that it is not immediately clear how one can choose the inducing points optimally. In Section 3.5, we simulate the inducing data from the same distribution as data X for the synthetic experiment and subsample the data randomly for the real data experiment. But, we will leave the more general case as further work.

3.4.1.2 Null Distribution Estimations

Having introduced the biased Nyström HSIC statistic, we will now move on to discuss two null distribution estimation methods, namely the permutation approach and the Nyström spectral approach. The permutation approach is exactly the same as Section 3.2.4.1 with $\Xi_{k_X, k_Y}(\mathbf{z})$ replaced by $\hat{\Xi}_{Ny, \tilde{k}_X, \tilde{k}_Y}(\mathbf{z})$. It is worth noting that for each permutation, we need to simulate a new set of inducing points for X and Y such that $n_x, n_y \ll m$ with m being the number of samples.

Likewise, the Nyström spectral approach is similar to that described in Section 3.2.4.2 where eigen-decompositions of the centred Gram matrices are required to simulate the null distribution. The difference is that we approximate the centred Gram matrices using Nyström method and the HSIC V -statistic is replaced by the Nyström HSIC estimator $\hat{\Xi}_{Ny, \tilde{k}_X, \tilde{k}_Y}(\mathbf{z})$. So, the null distribution is then estimated

using the eigenvalues from the covariance matrices $\tilde{\Phi}_X^T \tilde{\Phi}_X$ and $\tilde{\Phi}_Y^T \tilde{\Phi}_Y$. In such a way, the computational complexity is reduced from the original $\mathcal{O}(m^3)$ to $\mathcal{O}(n_x^3 + n_y^3 + (n_x^2 + n_y^2)m + n_x n_y m)$ i.e. linear in m .

3.4.2 Random Fourier Feature HSIC

So far, we have looked at two large-scale approximation techniques that are applicable to any positive definite kernel. If the corresponding kernel also happens to be translation-invariant with the moment condition in (3.39), however, an additional popular large-scale technique can be applied: random Fourier features of Rahimi and Recht [2007] which is based on Bochner’s representation. Note that many other kernels though not translational invariant, e.g. arc-cosine kernel, are also universal and approximable by random features [Cho and Saul, 2009]. In this section, we will first review Bochner’s theorem and subsequently build up to how random Fourier features can be used to approximate large kernel matrices. Utilising it in the context of independence testing, we propose the RFF HSIC estimator and further consider two null distribution estimation approaches.

3.4.2.1 Bochner’s Theorem

Through the projection of data into lower dimensional randomised feature space, Rahimi and Recht [2007] proposed a method of converting the training and evaluation of any kernel machine into the corresponding operations of a linear machine. In particular, they showed using a randomised feature map $z : \mathbb{R}^d \rightarrow \mathbb{R}^D$ we can obtain

$$k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle \approx z(x)^T z(y) \quad (3.34)$$

where $x, y \in \mathbb{R}^d$. More specifically, Rahimi and Recht [2007] demonstrate the construction of feature spaces that uniformly approximate shift-invariant kernels with $D = O(d\epsilon^{-2} \log \frac{1}{\epsilon^2})$ where ϵ is the accuracy of approximation. However, as we will see later certain moment conditions need to be satisfied.

Bochner’s Theorem provides the key observation behind such approximation. This classical theorem [Wendland, 2005, Theorem 6.6] is useful in several contexts where one deals with translation-invariant kernels k , i.e. $k(x, y) = k(x - y)$. As well as constructing large-scale approximation to kernel methods [Rahimi and Recht, 2007], it can also be used to determine whether a kernel is characteristic i.e. if the Fourier transform of a kernel is supported everywhere then the kernel is characteristic [Sriperumbudur, 2010].

Theorem 3.4.1. Bochner's Theorem [Wendland, 2005] *A continuous transition-invariant kernel k on $\mathbb{R}^d$ is positive definite if and only if $k(x - y)$ is the Fourier transform of a non-negative measure.*

For a properly scaled transition-invariant kernel k , the theorem guarantees that its Fourier transform $\Gamma(w)$ is a non-negative measure on $\mathbb{R}^d$. Without serious loss of generality, Γ is a probability distribution. Since we would like to approximate real-valued kernel matrices, let us consider the approximation which uses only real-valued features. $k(x - y)$ can be written as :

$$k(x - y) = \int_{\mathbb{R}^d} \exp(iw^T(x - y)) d\Gamma(w) \quad (3.35)$$

$$= \int_{\mathbb{R}^d} \cos(w^T(x - y)) + i \sin(w^T(x - y)) d\Gamma(w) \quad (3.36)$$

$$= \int_{\mathbb{R}^d} \cos(w^T(x - y)) d\Gamma(w) \quad (3.37)$$

$$= \int_{\mathbb{R}^d} \{ \cos(w^T x) \cos(w^T y) + \sin(w^T x) \sin(w^T y) \} d\Gamma(w) \quad (3.38)$$

provided that

$$\mathbb{E}_{\Gamma}(w^T w) < \infty. \quad (3.39)$$

Note, (3.37) follows because kernels are real valued and (3.38) uses the double angle formula for cosine. The random features can be computed by first sampling $\{w_j\}_{j=1}^D \stackrel{i.i.d.}{\sim} \Gamma$ and then for $x_j \in \mathbb{R}^d$ with $j \in \{1, \dots, n\}$, setting

$$z(x_j) = \sqrt{\frac{2}{D}} (\cos(w_1^T x_j), \sin(w_1^T x_j), \dots, \cos(w_{\frac{D}{2}}^T x_j), \sin(w_{\frac{D}{2}}^T x_j))$$

for $j \in \{1, \dots, n\}$.

Here, we deal with explicit feature space, and apply linear methods to approximate the Gram matrix through the covariance matrix $Z(x)^T Z(x)$ of dimension $D \times D$ where $Z(x)$ is the matrix of random features. Essentially, (3.39) guarantees that the second moment of the Fourier transform of this translational invariant kernel k to be finite and hence ensure the uniform convergence of $z(x)^T z(y)$ to $\kappa(x - y)$ [Rahimi and Recht, 2007].

3.4.2.2 RFF HSIC Estimator

The derivation of the biased RFF HSIC estimator follows in the same manner as Section 3.4.1.1. However, with the RFF approximations of the kernel matrices, (3.24)

becomes:

$$\begin{aligned} & \hat{\Xi}_{RFF, \tilde{k}_x, \tilde{k}_y}(\mathbf{z}) \\ &= \left\| \frac{1}{m} \sum_{i=1}^m Z_x(x_i) Z_y(y_i)^T - \left(\frac{1}{m} \sum_{i=1}^m Z_x(x_i) \right) \left(\frac{1}{m} \sum_{i=1}^m Z_y(y_i) \right)^T \right\|_F^2 \end{aligned} \quad (3.40)$$

$$= \left\| \frac{1}{m} Z_x^T H Z_y \right\|_F^2 \quad (3.41)$$

where $Z_x \in \mathbb{R}^{m \times D_x}$ and $Z_y \in \mathbb{R}^{m \times D_y}$. Hence, when RFF estimators are substituted, the cross covariance operator is simply a $D_x \times D_y$ matrix. In the same way as the Nyström HSIC estimator, the HSIC statistic computed using RFF is a HSIC statistic computed using a different kernel, i.e. one that is induced by the random features. It is worth noting that the analysis of convergence of such estimator can possibly be done similarly to the analysis by Sutherland and Schneider [2015] for MMD. However, we will leave this for future work.

To use the RFF HSIC statistic in independence testing, the permutation approach and spectral approach in the previous section can be adopted for null distribution estimation with $\hat{\Xi}_{Ny, \tilde{k}_x, \tilde{k}_y}(\mathbf{z})$ replaced by $\hat{\Xi}_{RFF, \tilde{k}_x, \tilde{k}_y}(\mathbf{z})$. Just as the case with inducing points, the $\{w_j\}_{j=1}^D$ should be sampled each time independently for X and Y when the RFF approximations Z_x and Z_y needed to be computed. As a remark, the number of inducing points and the number of w_j s plays a similar role in both methods which controls the trade off between computational complexity and statistical power. In practice, as we will demonstrate in the next section, such number can be much smaller than the size of the dataset without compromising the performance.

3.5 Experiments

In this section, we present three synthetic data experiments and one real data experiment to study the behaviour of our large scale HSIC tests. The main experiment is on a challenging non-linear low signal-to-noise ratio dependence dataset to assess the numerical performance amongst the large scale HSIC tests. To investigate the performance of these test in a small scale, we further conduct linear and sine dependence experiments to compare with currently established methods for independence testing. Throughout this section, we set the significance level of the hypothesis test to be $\alpha = 0.05$. Both Type I and Type II errors are calculated based on 100 trials. The 95% confidence intervals are computed based on normality assumption, i.e.

$\hat{\mu} \pm 1.96\sqrt{\frac{\hat{\mu}(1-\hat{\mu})}{100}}$, where $\hat{\mu}$ is the estimate for the statistical power. Unless otherwise stated, HSIC, RFF and Nyström approaches are all using Gaussian kernel with median-heuristic.

3.5.1 Simple Linear Experiment

We begin with an investigation of the performance of our methods on a toy example with a small number of observations, in order to check the agreements between large scale approximation methods we proposed and the exact methods where they are still feasible. Towards this aim, we consider a simple linear dependence experiment, but where the dependence between the response Y and the input X is in only a single dimension of X . In particular,

$$X \sim \mathcal{N}(0, I_d) \quad \text{and} \quad Y = X_1 + Z$$

where d is the dimensionality of data vector X and X_1 indicate the first dimension of X . The noise Z is independent standard Gaussian noise. We would like to compare methods based on HSIC to a method based on Pearson's Correlation which is explicitly aimed at linear dependence and should give strongest performance. However, as the latter cannot be directly applied to multivariate data, we consider a SubCorr statistic: $\text{SubCorr} = \frac{1}{d} \sum_{i=1}^d \text{Corr}(Y, X_i)^2$ where $\text{Corr}(Y, X_i)$ is the Pearson's correlation between Y and the i^{th} dimension of X . In addition, we will also consider SubHSIC statistic: $\text{SubHSIC} = \frac{1}{d} \sum_{i=1}^d \text{HSIC}(Y, X_i)$. For these two methods, we will use a permutation approach as their distributions are not immediately clear.

In Figure 5.1, the dimension of X is set to be 10. Both the number of random features in RFF and the number of inducing variables in Nyström are set to 10. We do not use the block-based method as the sample sizes are small. From Figure 5.1 (right), we see that SubCorr yields the highest power as expected. HSIC and SubHSIC with Gaussian median-heuristic kernels perform similarly though, with all three giving the power of 1 at the sample size of 100. On the other hand, Figure 5.1 (left) shows that the two large scale methods are still able to detect the dependence at these small sample sizes, even though there is some loss in power in comparison to HSIC and they would require a larger sample size. As we will see, this requirement for a larger sample size will be offset by a much lower computational cost in large-scale examples.

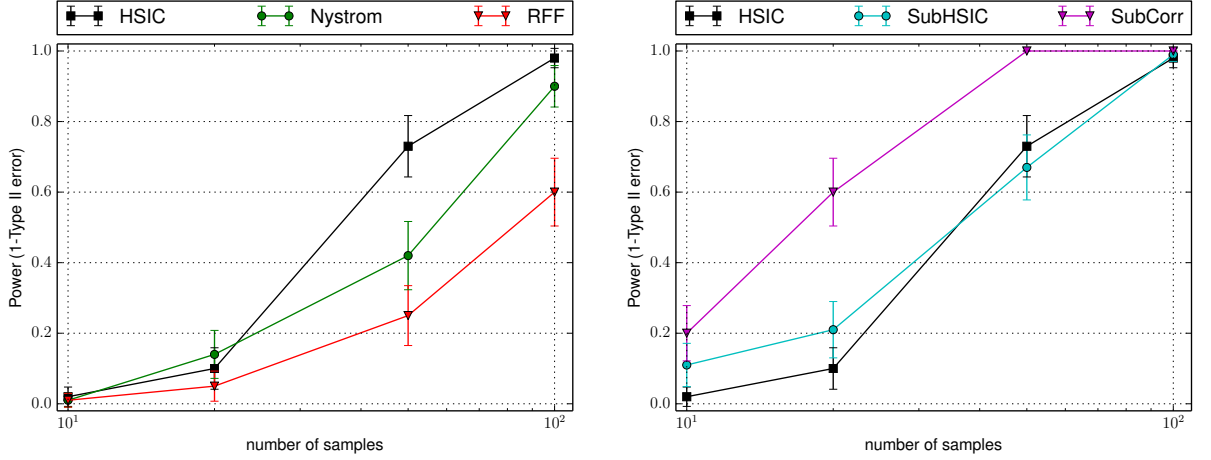


Figure 3.1: Simple linear experiment for $d = 10$. Left: comparing HSIC spectral approach with Nystrom spectral method and RFF spectral method. Right: HSIC spectral approach with SubHSIC and SubCorr.

3.5.2 Nonlinear Experiments

In this section, we consider two nonlinear experiments: one with relatively small sample sizes (up to 4000 samples) and a large scale scenario (with number of samples from 1000 to 10^7). For both experiments, we investigate the power vs. time trade-off of the introduced large-scale approximate tests and compare their performance to quadratic time ones. A summary of the methods investigated here is illustrated in Table 3.1.

In addition to HSIC and its large scale versions, we will also consider a normalisation of HSIC: dCor [Székely et al., 2007, Székely and Rizzo, 2009] which can be formulated in terms of HSIC using a Brownian Kernel with parameter $h = 0.5$ [Sejdinovic et al., 2013b, Appendix A]. For clarity and consistency of comparison, methods using the same kernel will be compared with each other. In particular, we will compare HSIC and its large-scale approximations with GdCor (dCor using Gaussian kernel with median heuristic bandwidth parameter) and its corresponding large-scale approximations. As the asymptotic null distribution for GdCor is unclear, a permutation approach (see Section. 3.2.4.1) will be used for p-value computations. Similar comparison will be done for the Brownian kernel with $h = 0.5$.

Note that Random Fourier Feature approaches cannot be directly applied to the Brownian kernel as it is not translational invariant. In addition, we will not use

Gaussian Kernel		Brownian Kernel	
Exact $\mathcal{O}(m^2)$	Approximate $\mathcal{O}(mn)$	Exact $\mathcal{O}(m^2)$	Approximate $\mathcal{O}(mn)$
(G)HSIC	<i>Block</i> RFF Nyström	BHSIC	<i>Block</i> RFF ? Nyström
GdCor	RFF Nyström	(B)dCor	RFF ? Nyström

Table 3.1: Summary of the methods compared in the nonlinear dependence experiments. Letters in the parentheses indicate the default kernel (G: Gaussian; B: Brownian). Note that the Random Fourier Feature approach is not directly applicable to the Brownian kernel (as it is not shift invariant).

the block-based approach in the small-scale non-linear experiment as the asymptotic normality assumption for the null distribution requires large sample size - i.e., large number of blocks as well as a large number of samples per block.

3.5.2.1 Small-Scale Nonlinear Experiment

Consider a sine dependence experiment to investigate time vs power tradeoffs of large-scale tests. The dataset consists of a sample of size m generated i.i.d. according to:

$$X \sim \mathcal{N}(0, I_d) \quad \text{and} \quad Y = 20 \sin(4\pi(X_1^2 + X_2^2)) + Z$$

where d is the dimensionality of data vector X , X_i indicates the i^{th} dimension of X and $Z \sim \mathcal{N}(0, 1)$. For this experiment, the dependency can be detected with a relatively small number of samples. Figure 3.2 illustrates the power as a function of the sample size when $d = 2$; the number of random Fourier features and inducing variables are here both set to 50.

From Figure 3.2, dCor clearly outperforms all the other methods with its Nyström approximation giving the closest performance in terms of power. Reassuringly, all the six methods using Gaussian kernel give very similar power performance. Although dCor and its large-scale approximation give superior power performance, the performance of HSIC and its large-scale approximations seems to be indifferent to the kernel used. Figure 3.3, however, tells a much more interesting story - the large-scale methods all reach the power of 1 in a test time which is several orders of magnitude smaller, demonstrating the utility of the introduced tests.

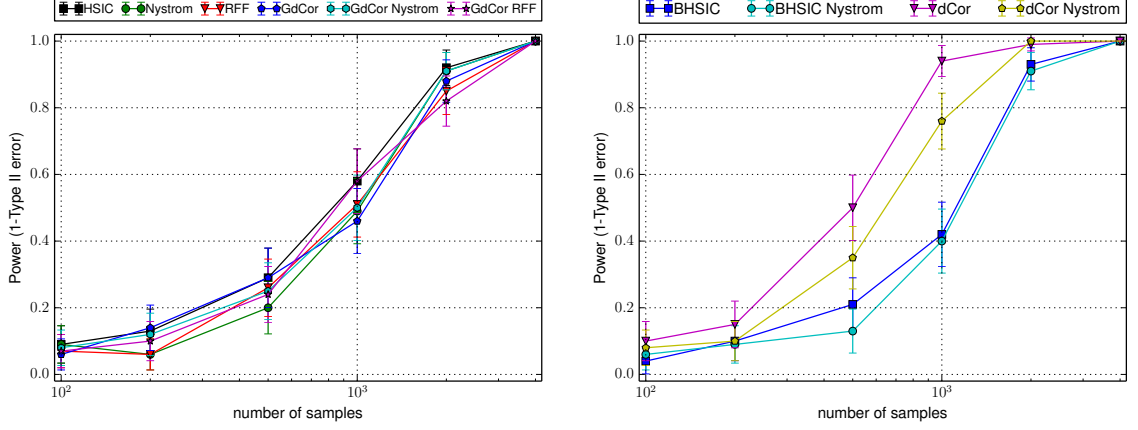


Figure 3.2: Nonlinear sine dependency experiment for $d = 2$. Left: All methods using Gaussian kernel with median heuristic bandwidth. Right: All methods using Brownian kernel with $h = 0.5$.

3.5.2.2 Large-Scale Nonlinear Experiment

Now, we would like to compare the performance of the proposed large scale HSIC tests with each other – at sample sizes where standard HSIC / dCor approaches are no longer feasible. Gaussian kernel with median heuristic bandwidth will be used throughout this section contrasting the block-based spectral approach, RFF spectral approach and the Nyström spectral approach. The other approximate methods listed in Table 3.1 require costly permutation approaches to compute the null distribution, and we therefore do not examine them in this subsection.

We consider a challenging nonlinear and low signal-to-noise ratio experiment, where a sample of size m is generated i.i.d. according to:

$$X \sim \mathcal{N}(0, I_d) \quad \text{and} \quad Y = \sqrt{\frac{2}{d}} \sum_{j=1}^{d/2} \text{sign}(X_{2j-1} X_{2j}) |Z_j| + Z_{\frac{d}{2}+1}$$

where d is the dimensionality of the data set X and $Z \sim \mathcal{N}(0, I_{\frac{d}{2}+1})$. Note that Y is independent of each individual dimension of X and that the dependence is non-linear. For $d = 50$ and 100 , we would like to explore the relationship between the test power across a different number of samples $m = \{10^5, 2 \times 10^5, 5 \times 10^5, 10^6, 2 \times 10^6, 5 \times 10^6, 10^7\}$. The number of random features, inducing variables and block size are all set to 200 so that their computational cost is comparable. For RFF and Nyström methods, we

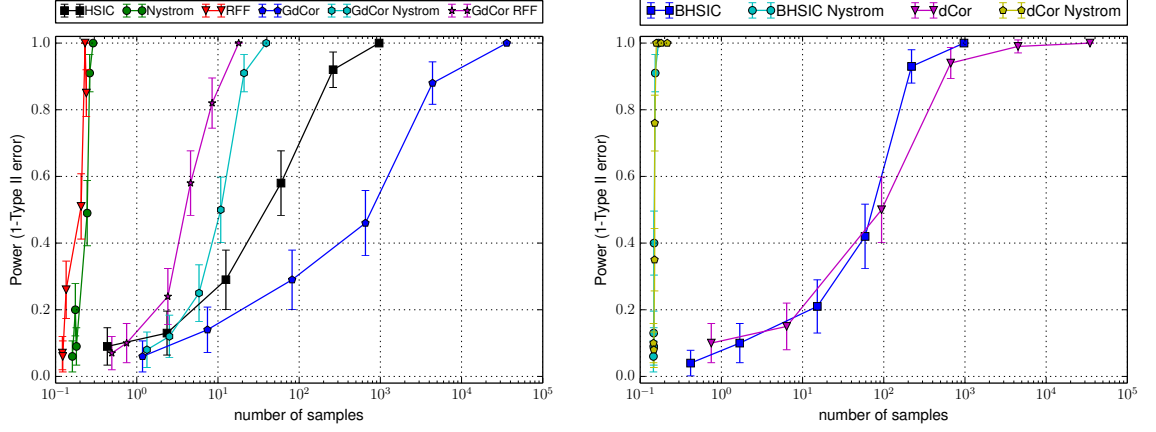


Figure 3.3: The corresponding average testing time plot for the sine dependence experiment for $d = 2$.

used the spectral approach to estimate the null distribution.

Figure 3.4 is a plot of the test power against the number of samples whereas Figure 3.5 is a plot of the test power against average testing time. In the computational time comparison plot, we also present baseline experiments which simply apply the quadratic time HSIC spectral approach on a subset of the available data (with a subset size be in $\{100, 200, 500, 1000, 2000, 4000\}$). For the considered computation budgets, it is clear that this approach is unable to detect the dependence.

It is clear in Figure 3.4 that for both $d = 50$ and $d = 100$, the RFF method gives the best performance in power for a fixed number of samples, followed by the Nyström method and then by the block-based approach. Although parallel computing could be employed for the block-based HSIC method, as we observed, RFF and Nyström methods will be preferred for achieving higher statistical power at any given samples size. The RFF method is able to achieve zero type II error (i.e. no failure to reject a false null) with 5×10^4 samples for $d = 50$ and 5×10^5 samples for $d = 100$, while the Nyström method has a 80% false negative rate at these sample sizes. The power vs time plot in Figure 3.5 gives a similar picture as Figure 3.4 confirming the superiority of the RFF method on this example.

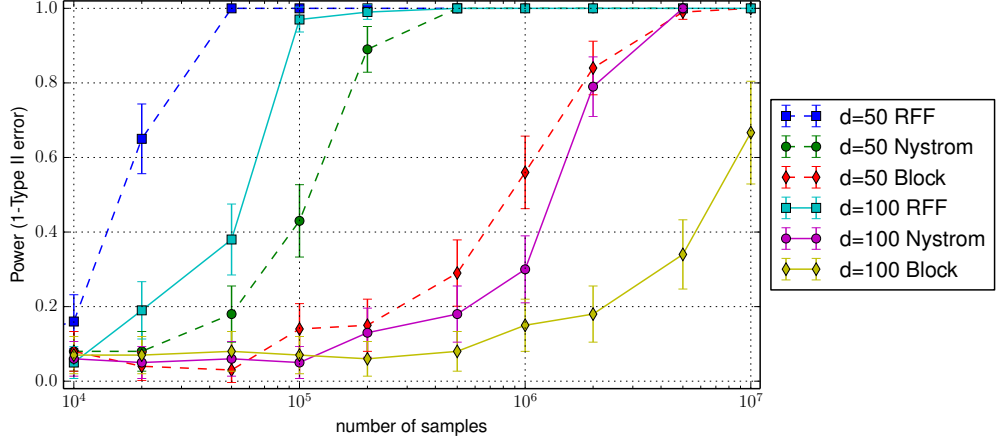


Figure 3.4: Large Scale Experiment: The statistical power comparison between the three large scale independence testing methods based on 100 trials. Dotted line: $d = 50$; solid line: $d = 100$. The 95% confidence intervals are computed based on normality assumption, i.e. $\hat{\mu} \pm 1.96 \sqrt{\frac{\hat{\mu}(1-\hat{\mu})}{100}}$, where $\hat{\mu}$ is the estimate for the statistical power.

3.5.2.3 Real Data Experiment

In this section, we examine the performance of the three proposed large-scale HSIC tests on real data – a subset of the Million Song Dataset (MSD) [Bertin-Mahieux et al., 2011]. The dataset consists of 515,345 observations with each song X represented by 90 features, of which 12 features are timbre average (over all segments) of the song, and 78 features are timbre covariance. The aim is to detect the dependence between X and the year of release Y which ranges from 1922 to 2011. As such dependency is rather easy to detect [Jitkrittum et al., 2017], we jitter each entry of the X matrix with independent Gaussian noise with mean 0 and variance 1000. The significance level is set at $\alpha = 0.05$ and we repeat for 200 trials where each time a subset of the data is randomly sampled. The number of random Fourier features, inducing points and block size are all set to be 10. Figure 3.6 shows the power performance across a range of sample sizes. For this dataset, Nyström spectral approach provides the best power performance followed by RFF spectral approach. Similar to the previous experiment, the Block based method gives the least power for any fixed number of samples. The power vs. average testing time plot gives a similar picture.

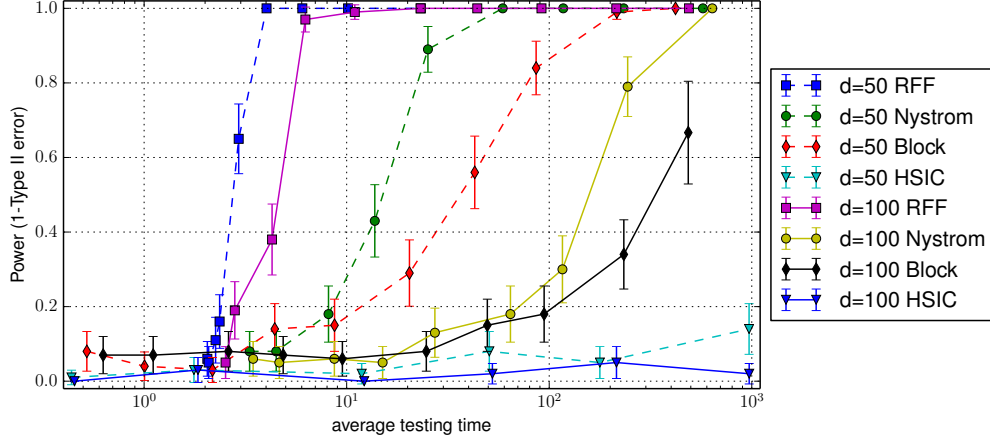


Figure 3.5: Large Scale Experiment: The average testing time comparison between the three large scale independence testing methods for samples size $m = \{10^5, 2 \times 10^5, 5 \times 10^5, 10^6, 2 \times 10^6, 5 \times 10^6, 10^7\}$. Quadratic time HSIC spectral method is also used on subset of data of size $\{100, 200, 500, 1000, 2000, 4000\}$ as baseline comparison for similar testing time. Dotted line: $d = 50$; solid line: $d = 100$.

3.6 Discussion and Conclusion

We have proposed three large scale estimators of HSIC, a kernel-based nonparametric dependence measure – these are the block-based estimator, the Nyström estimator and the RFF estimator. We have subsequently established suitable independence testing procedures for each method – by taking advantage of the normal asymptotic null distribution of the block-based estimator and by employing an approach that directly estimates the eigenvalues appearing in the asymptotic null distribution for the Nyström and RFF methods. All three tests significantly reduce computational complexity in memory and time over the standard HSIC-based test. We have verified the validity of our large-scale testing methods and its favourable tradeoffs between testing power and computational complexity on challenging high-dimensional synthetic data. We have observed that RFF and Nyström approaches have considerable advantages over the block-based test. Several further extensions can be studied: the developed large-scale approximations are readily applicable to three-variable interaction testing [Sejdinovic et al., 2013a], conditional independence testing [Fukumizu et al., 2008] as well as application in causal discovery [Zhang et al., 2011, Flaxman et al., 2015]. Moreover, the RFF HSIC approach can be extended using the additional smoothing of characteristic function representations similarly to the approach of Chwialkowski et al. [2015] in the context of two-sample testing. Furthermore, one can also consider

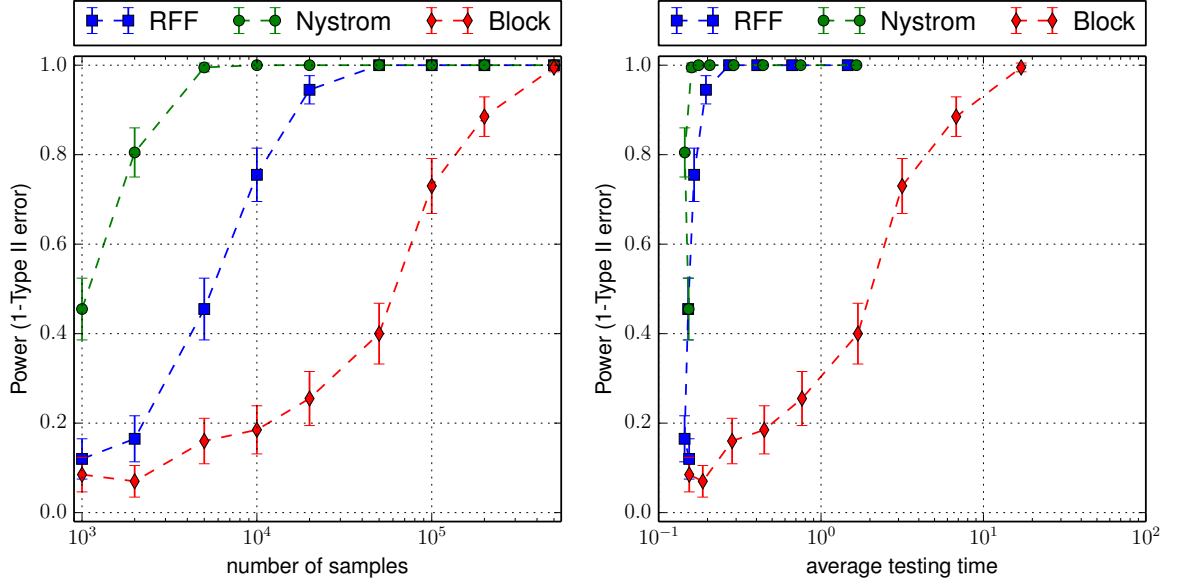


Figure 3.6: Real Data Experiment: (Left) power vs. sample size; (Right) power vs. average testing time. The statistical power comparison between the three large scale independence testing methods based on 200 trials. The 95% confidence intervals are computed based on normality assumption, i.e. $\hat{\mu} \pm 1.96\sqrt{\frac{\hat{\mu}(1-\hat{\mu})}{100}}$, where $\hat{\mu}$ is the estimate for the statistical power.

the robustness of the proposed approaches in the case where heterogeneous datasets are considered, or when some form of within-sample dependence is introduced, e.g. when testing for independence between time series [Chwialkowski and Gretton, 2014].

3.7 Appendix: Proof of Theorem 3.2.1

We note that Lyons [2013] gives a proof to a similar theorem [Lyons, 2013, Theorem 2.7] regarding generalised versions of distance covariance (dCov). We closely follow the steps given in Proposition 2.6 and Theorem 2.7 of Lyons [2013]. However, the proof provided here is tailored to the kernel view of HSIC/dCov duality [Sejdinovic et al., 2013b] and it is slightly more general in that it applies to any semimetric (i.e. a metric that does not satisfies the triangle inequality) of negative type⁴. Unless stated otherwise, the notation follows that used in the main part of this paper.

⁴We follow the Definition 2 of Sejdinovic et al. [2013b]: A semimetric ρ is said to have negative type if $\forall n \geq 2, z_1, \dots, z_n \in \mathcal{Z}$ and $\alpha_1, \dots, \alpha_n \in \mathbb{R}$, with $\sum_{i=1}^n \alpha_i = 0$, $\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \rho(z_i, z_j) \leq 0$.

Recall from the main part of this paper that the existence of the HSIC statistic $\Xi_{k_{\mathcal{X}}, k_{\mathcal{Y}}}(\mathbf{z})$ defined in (3.6) requires the marginal distributions to have finite first moment with respect to the kernels, i.e. $P_X \in \mathcal{M}_{k_{\mathcal{X}}}^1(\mathcal{X})$ and $P_Y \in \mathcal{M}_{k_{\mathcal{Y}}}^1(\mathcal{Y})$. By Proposition 20 of Sejdinovic et al. [2013b], this translates directly into finite first moment conditions with respect to the semimetrics: $P_X \in \mathcal{M}_{d_x}^1(\mathcal{X})$ and $P_Y \in \mathcal{M}_{d_y}^1(\mathcal{Y})$ when $k_{\mathcal{X}}$ generates d_x and $k_{\mathcal{Y}}$ generates d_y . Similarly for the imposed finite second moment condition.

More specifically, a valid semimetric d of negative type on $\mathcal{Z}$ generated by a non-degenerate kernel k on $\mathcal{Z}$ can be written as $d(z, z') = k(z, z) + k(z', z') - 2k(z, z')$ [Sejdinovic et al., 2013b, Corollary 16]. Then, such semimetric d_x centred at the probability measure P_X defined on $\mathcal{X}$ is

$$\begin{aligned} d_{P_X}(x, x') &:= d_x(x, x') - \int d_x(x, x') dP_X(x') \\ &\quad - \int d_x(x', x) dP_X(x) + \int d_x(x, x') dP_X^2(x, x') \end{aligned} \quad (3.42)$$

Similarly, $d_{P_Y}(y, y')$ is the semimetric centred at the probability measure P_Y defined on $\mathcal{Y}$. If we substitute the kernel representation of the semimetric for X and Y respectively into $d_{P_X}(x, x')$ and $d_{P_Y}(y, y')$, we obtain the following:

$$d_{P_X}(x, x') = -2\tilde{k}_{P_X}(x, x') \text{ and } d_{P_Y}(y, y') = -2\tilde{k}_{P_Y}(y, y') \quad (3.43)$$

where $\tilde{k}_{P_X}(x, x')$ and $\tilde{k}_{P_Y}(y, y')$ are defined in (3.12).

Let $(X^i, Y^i) \sim \theta$ be independent for $i \in \{1, 2, \dots, 6\}$, we introduce the “core” defined in Lyons [2013],

$$h((X^1, Y^1), (X^2, Y^2), \dots, (X^6, Y^6)) := f(X^1, X^2, X^3, X^4) f(Y^1, Y^2, Y^5, Y^6) \quad (3.44)$$

where for $z_i \in \mathcal{X}$ or $z_i \in \mathcal{Y}$:

$$f(z_1, z_2, z_3, z_4) := d.(z_1, z_2) - d.(z_1, z_3) - d.(z_2, z_4) + d.(z_3, z_4) \quad (3.45)$$

$$= -2[k.(z_1, z_2) - k.(z_1, z_3) - k.(z_2, z_4) + k.(z_3, z_4)] \quad (3.46)$$

The second line follows from the relationship between d and k where $k_{\mathcal{X}}$ is used for $z_i \in \mathcal{X}$ and $k_{\mathcal{Y}}$ for $z_i \in \mathcal{Y}$.

In fact, we can prove that the expectation of (3.44) is four times the HSIC of X and Y . To see this, we first need to show that such expectation is well defined.

Indeed, note that for a valid semimetric d of negative type on $\mathcal{Z}$, $\sqrt{d(z, z')} = \|k(\cdot, z) - k(\cdot, z')\|_{\mathcal{H}_k}$, then the following inequality holds:

$$d(x, y) \leq d(x, z) + d(y, z) + 2\sqrt{d(x, z)d(y, z)} \quad \forall x, y, z \in \mathcal{Z} \quad (3.47)$$

It then follows that

$$\begin{aligned} & |f(z_1, z_2, z_3, z_4)| \\ & \leq 2d(z_2, z_3) + 2\sqrt{d(z_1, z_3)d(z_2, z_3)} + 2\sqrt{d(z_2, z_3)d(z_2, z_4)} \\ & \leq 2d(z_2, z_3) + 2\max\{d(z_1, z_3), d(z_2, z_3)\} + 2\max\{d(z_2, z_3), d(z_2, z_4)\} \\ & := g_1(z_1, z_2, z_3, z_4) \end{aligned}$$

and that

$$\begin{aligned} & |f(z_1, z_2, z_3, z_4)| \\ & \leq 2d(z_1, z_4) + 2\sqrt{d(z_2, z_4)d(z_1, z_4)} + 2\sqrt{d(z_1, z_3)d(z_1, z_4)} \\ & \leq 2d(z_1, z_4) + 2\max\{d(z_2, z_4), d(z_1, z_4)\} + 2\max\{d(z_1, z_3), d(z_1, z_4)\} \\ & := g_2(z_1, z_2, z_3, z_4) \end{aligned}$$

Hence, replacing the z_i in g_1 with X^i and the z_i in g_2 with Y^i ,

$$|h((X^1, Y^1), (X^2, Y^2), \dots, (X^6, Y^6))| \leq g_1(X^1, X^2, X^3, X^4)g_2(Y^1, Y^2, Y^5, Y^6).$$

Since the marginal distributions have finite first moments with respect to the semimetrics of negative type, then each of the terms in g_1 and g_2 is integrable and hence g_1 and g_2 are integrable. Moreover, since the marginal distributions are assumed to have finite second moments with respect to the kernels, then the joint distribution satisfies $P_{XY} \in \mathcal{M}_{k_X \otimes k_Y}^1(\mathcal{X} \times \mathcal{Y})$. This implies that $P_{XY} \in \mathcal{M}_{\tilde{d}}^1(\mathcal{X} \times \mathcal{Y})$ where $\tilde{d}$ is generated by the product kernel $k_X \otimes k_Y$ [Sejdinovic et al., 2013b, Theorem 24]. Therefore, h is integrable. Subsequently, by taking the expectation and utilising Fubini's theorem, we obtain that

$$\mathbb{E}(h((X^1, Y^1), (X^2, Y^2), \dots, (X^6, Y^6))) = 4\mathbb{E}(\tilde{k}_{P_X}(X^1, X^2)\tilde{k}_{P_Y}(Y^1, Y^2)) \quad (3.48)$$

$$= 4\Xi_{k_X, k_Y}(\mathbf{z}) \quad (3.49)$$

which is 4 times the HSIC.

In order to use the theory of degenerate V -statistics to obtain the asymptotic distribution, we need to consider the symmetrised version of h , which we define as follows

$$\bar{h}((X^1, Y^1), \dots, (X^6, Y^6)) := \frac{1}{6!} \sum_{\sigma \in A} h((X^{\sigma(1)}, Y^{\sigma(1)}), \dots, (X^{\sigma(6)}, Y^{\sigma(6)}))$$

where A denotes the set of all permutations of $\{1, \dots, 6\}$. Then, under the null hypothesis of independence $P_{XY} = P_X \times P_Y$,

$$\begin{aligned} \bar{h}_2((x, y), (x', y')) &:= \mathbb{E}[\bar{h}((x, y), (x', y'), (X^3, Y^3), \dots, (X^6, Y^6))] \\ &= \frac{4}{15} \tilde{k}_{P_X}(x, x') \tilde{k}_{P_Y}(y, y') \end{aligned}$$

If we fix the first two positions to be $\{1, 2\}$ and randomly permute the rest, we obtain 24 different combinations. Similarly if we fix the first two positions to be $\{2, 1\}$, we also obtain 24 different combinations. Some algebraic manipulation shows that these are the combinations that gives the expectation of h to be $\tilde{k}_{P_X}(x, x') \tilde{k}_{P_Y}(y, y')$. In fact, these are the only combinations as when either $\{1\}$ or $\{2\}$ or both are not in the first two positions, the expectation of h is zero and all terms cancel out. For example, consider the case $\mathbb{E}(h((X^3, Y^3), (x, y), (X^4, Y^4), (x', y'), (X^5, Y^5), (X^6, Y^6)))$, in which case $\mathbb{E}(f(Y^3, y, Y^5, Y^6)) = 0$.

Another important condition to check is that $\bar{h}_2$ has finite second moment. It was shown earlier that $h((X^1, Y^1), \dots, (X^6, Y^6))$ of (3.44) is integrable. Equivalently, $\mathbb{E}(f(X^1, X^2, X^3, X^4)f(Y^1, Y^2, Y^5, Y^6)) < \infty$. Replace Y with X and let $X^5 = X^3$ and $X^6 = X^4$ to obtain $\mathbb{E}(f(X, X', X^3, X^4)^2) < \infty$. In other words, $f(X^1, X^2, X^3, X^4)$ has finite second moment. Hence, $h((X^1, Y^1), \dots, (X^6, Y^6))$ has finite second moment under the null hypothesis. Additionally, by Jensen's inequality, $\mathbb{E}(|\bar{h}_2((X, Y), (X, Y))|) \leq \{\mathbb{E}((\bar{h}_2((X, Y), (X, Y)))^2)\}^{1/2} < \infty$.

Hence, by Theorem B in Chapter 6 of Serfling [2002], which says

$$m \left(\frac{1}{m^2} \sum_{i,j} \bar{h}_2((x_i, y_i), (x_j, y_j)) \right) \xrightarrow{D} \sum_{r=1}^{\infty} \bar{\gamma}_r Z_r^2$$

as the sample size $m \rightarrow \infty$, we obtain that

$$m \left(\frac{1}{m^2} \sum_{i,j} \tilde{k}_{P_X}(x_i, x_j) \tilde{k}_{P_Y}(y_i, y_j) \right) \xrightarrow{D} \sum_{r=1}^{\infty} \gamma_r Z_r^2 \quad (3.50)$$

with $Z_r \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1) \forall r$ and $\{\gamma_r\}_{r=1}^\infty$ are the eigenvalues of the operator $S_{\tilde{k}}$: $L_\theta^2(\mathcal{X} \times \mathcal{Y}) \rightarrow L_\theta^2(\mathcal{X} \times \mathcal{Y})$ defined as:

$$S_{\tilde{k}}g(x, y) = \int_{\mathcal{X} \times \mathcal{Y}} \tilde{k}_{P_X}(x, x')\tilde{k}_{P_Y}(y, y')g(x', y')d\theta(x', y')$$

Under the null hypothesis $P_{XY} = P_X \times P_Y$, the above operator is given by the tensor product of $S_{\tilde{k}_{P_X}}$ and $S_{\tilde{k}_{P_Y}}$ [Lyons, 2013, Remark 2.9]. Therefore $\{\gamma_r\}_{r=1}^\infty$ are the products of the eigenvalues of these two operators. Recall that $k_{\mathcal{X}}$ is equivalent to $\tilde{k}_{P_X}$ and $k_{\mathcal{Y}}$ is equivalent to $\tilde{k}_{P_Y}$ in the sense of Definition 17 of Sejdinovic et al. [2013b], i.e. they generate the same semimetric of negative type. By noting that $\Xi_{b, k_{\mathcal{X}}, k_{\mathcal{Y}}}(\mathbf{Z})$ is invariant under the change of kernel within the class of equivalent kernels [Sejdinovic et al., 2013b, Theorem 24], $\frac{1}{m^2} \sum_{i,j} \tilde{k}_{P_X}(x_i, x_j)\tilde{k}_{P_Y}(y_i, y_j)$ is the highest order term in the V -statistic and thus determine its asymptotic null distribution.

Chapter 4

Large-Scale Normalised Kernel Independence Tests

So far, we have considered dependence measure based on the squared Hilbert-Schmidt norm of a cross-covariance operator. It is natural to consider if a RKHS equivalent of cross-correlation can also be defined. Indeed, Fukumizu et al. [2008] proposed, as a normalised version of HSIC, the squared Hilbert Schmidt norm of the NOrmalised Cross Covariance Operator (NOCCO). In the same work, Fukumizu et al. [2008] showed that it provides better performance than HSIC on a number of synthetic and real data experiments. However it experiences the same prohibitive computational cost when the sample size is large. In this chapter, we mimic the approach from Chapter 3 [Zhang et al., 2018] and apply random Fourier feature (RFF) approximation to the NOCCO statistic. Utilising machineries from the literature on Hotelling's T^2 -statistic [Hotelling, 1931], we derive the asymptotic null distribution of the RFF NOCCO statistic. This results in a distribution free test simplifying the p-value computation and hence leads to an attractive alternative to the large scale independence tests based on HSIC. On a variety of synthetic experiments, we illustrate its robustness to kernel bandwidth selection and the computational benefits of a distribution free test.

4.1 Background

Given random variables X and Y defined on some non-empty topological spaces $\mathcal{X}$ and $\mathcal{Y}$, let $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ be any kernels defined on the respective domains. We use $\mathcal{H}_{k_{\mathcal{X}}}$ and $\mathcal{H}_{k_{\mathcal{Y}}}$ to denote the induced reproducing kernel Hilbert spaces (RKHSs) by $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ respectively. Following the setting of [Fukumizu et al., 2008], we enforce a

finite first moment condition on the marginal probability measures P_X and P_Y , i.e.

$$P_X \in \mathcal{M}_{k_X}^1(\mathcal{X}) \quad \text{and} \quad P_Y \in \mathcal{M}_{k_Y}^1(\mathcal{Y}), \quad (4.1)$$

where $\mathcal{M}_k^\theta(\mathcal{Z}) = \{\nu \in \mathcal{M}(\mathcal{Z}) : \int k^\theta(z, z) d\nu(z) < \infty\}$ as defined in Chapter 3.

The cross-covariance operator $\Sigma_{XY} : \mathcal{H}_{k_Y} \rightarrow \mathcal{H}_{k_X}$ is the unique bounded operator such that

$$\langle f, \Sigma_{XY} g \rangle_{\mathcal{H}_{k_X}} = \text{Cov}[f(X), g(Y)] \quad \forall f \in \mathcal{H}_{k_X}, g \in \mathcal{H}_{k_Y}. \quad (4.2)$$

By replacing Y with X , the covariance operator $\Sigma_{XX} : \mathcal{H}_{k_X} \rightarrow \mathcal{H}_{k_X}$ can be defined and similarly Σ_{YY} . The cross-covariance operator is an extension of cross-covariance matrix, while covariance operator is an extension of covariance matrix – these notions coincide where RKHSs are finite-dimensional and isomorphic to Euclidean spaces. Through the use of nonlinear kernels k_X and k_Y , it is capable of representing higher order moments [Fukumizu et al., 2008]. Baker [1973] showed that the cross-covariance operator can be decomposed into a product of the marginal covariance operators and another operator which can be thought of as capturing cross-correlation, i.e.

$$\Sigma_{XY} = \Sigma_{XX}^{1/2} V_{XY} \Sigma_{YY}^{1/2}. \quad (4.3)$$

Such V_{XY} is unique and its operator norm is upper-bounded by 1. In the literature, it is known as the Normalised Cross-Covariance Operator (NOCCO) [Fukumizu et al., 2008, 2007].

NOCCO encodes the same information regarding the dependency between X and Y as the cross-covariance operator Σ_{XY} but it removes the effects of the marginals. This relationship between V_{XY} and Σ_{XY} can be understood analogously with that between covariance $\text{Cov}[X, Y]$ and correlation $\text{Corr}[X, Y] = \text{Cov}[X, Y] / \sqrt{\text{Var}[X]\text{Var}[Y]}$. The squared Hilbert Schmidt norm of NOCCO: $I^{NOCCO} := \|V_{XY}\|_{\text{HS}}^2$ can be used as a dependence measure [Fukumizu et al., 2008]. Since HSIC is defined as $\Xi_{k_X, k_Y}(X, Y) = \|\Sigma_{XY}\|_{\text{HS}}^2$ [Gretton et al., 2005a, 2008], in finite dimensions, the NOCCO statistic can be understood as HSIC with whitened features, i.e. $I^{NOCCO} = \|\Sigma_{XX}^{-1/2} \phi(x) \phi(y)^\top \Sigma_{YY}^{-1/2}\|_2^2 = \|\tilde{\phi}(x) \tilde{\phi}(y)^\top\|_2^2 = \|\tilde{\Sigma}_{XY}\|_2^2$ where $\phi(x)$ and $\phi(y)$ are the corresponding feature maps of kernels k_X and k_Y .

4.1.1 Relationship Between Kernel Dependence Measures

In fact, many important kernel dependence measures can be expressed in terms of the cross-covariance operator and its normalised version. Here, we provide a short review of the most relevant kernel dependence measures. For the interested readers, see for example, [Fukumizu et al., 2007, Section 2.3] and [Lopez-Paz, 2016, Section 5.2.2].

Kernel Canonical Correlation Analysis (KCCA)

The Kernel Canonical Correlation Analysis (KCCA) [Akaho, 2001, Melzer et al., 2001, Bach and Jordan, 2002] was originally proposed as a nonlinear extension of the classical canonical correlation analysis [Hotelling, 1936]. It seeks to find $f \in \mathcal{H}_X$ and $g \in \mathcal{H}_Y$ that solve the following problem:

$$\max_{f \in \mathcal{H}_X, g \in \mathcal{H}_Y, f \neq 0, g \neq 0} \frac{\text{Cov}[f(X), g(Y)]}{\text{Var}[f(X)]^{1/2} \text{Var}[g(Y)]^{1/2}}. \quad (4.4)$$

Under certain assumptions on the cross-covariance operator, the above can equivalently be formulated as finding $\tilde{f} \in \mathcal{H}_X$ and $\tilde{g} \in \mathcal{H}_Y$ such that

$$\max_{\substack{\tilde{f} \in \mathcal{H}_X, \tilde{g} \in \mathcal{H}_Y \\ \|\tilde{f}\|_{\mathcal{H}_X} = \|\tilde{g}\|_{\mathcal{H}_Y} = 1}} \langle \tilde{g}, V_{YX} \tilde{f} \rangle_{\mathcal{H}_Y}. \quad (4.5)$$

The Kernel Canonical Correlation (KCC) is the maximum value attained.

Constrained Covariance (COCO)

Related to KCCA, the COstrained COvariance (COCO) was proposed by Gretton et al. [2005b], which can also be seen as a nonlinear extension of classical linear canonical correlation (which relies on CCA). Unlike KCCA, it uses the unit eigenfunction of the cross-covariance operator Σ_{YX} instead of its normalised counterpart. The solution of COCO is given by the largest singular value of the cross-covariance operator by solving

$$\max_{\substack{f \in \mathcal{H}_X, g \in \mathcal{H}_Y \\ \|f\|_{\mathcal{H}_X} = \|g\|_{\mathcal{H}_Y} = 1}} \langle g, \Sigma_{YX} f \rangle = \max_{\substack{f \in \mathcal{H}_X, g \in \mathcal{H}_Y \\ \|f\|_{\mathcal{H}_X} = \|g\|_{\mathcal{H}_Y} = 1}} \text{Cov}[f(X), g(Y)] \quad (4.6)$$

When one is interested in finding the functions f and g , it is a less direct nonlinear extension of CCA since it normalises the covariance by the RKHS norms of f and g and it tends to find functions with large variance for $f(X)$ and $g(Y)$ which may not correspond to the most correlated features [Fukumizu et al., 2007, Section 2.3].

NOCCO and HSIC

While COCO and KCC consider the unit eigenfunction of the cross-covariance operator and the normalised cross-covariance operator respectively, HSIC and NOCCO consider the full spectrum of these operators through the square Hilbert Schmidt norm. The relationship between KCC and COCO is comparable to that between NOCCO and HSIC, in the sense that the former is based on the normalised cross-covariance operator and the latter is based on the cross-covariance operator. We summarise the relationships in Table 4.1.

	Cross Covariance Operator Σ_{XY}	Normalised Cross Covariance Operator V_{XY}
Hilbert Schmidt Norm (i.e. $\sum_i \lambda_i^2$)	HSIC [Gretton et al., 2005a]	NOCCO [Fukumizu et al., 2008]
Operator Norm (i.e. $\sup_i \lambda_i$)	COCO [Gretton et al., 2005b]	KCCA [Bach and Jordan, 2002]

Table 4.1: Summary of relationship between kernel based measures of dependence. λ_i denotes the singular values of the relevant operators.

4.1.2 Related Work

Kernel dependence measures require prohibitive computational cost due to the computation (and inversion) of the kernel matrices. In the previous chapter, we have introduced three large-scale kernel independence tests based on HSIC. Similar efforts have been made in establishing computational efficient methods for KCCA. Lopez-Paz et al. [2014] proposed the Randomised CCA (RCCA) which projects the given observations through a set of random features (e.g. Nyström or Mercer random features [Lopez-Paz, 2016, Section 3.2.2]) before carrying out the CCA procedure on the random feature dataset. Furthermore, to ensure that the dependence measure is invariant with respect to transformations on the marginal distributions, Lopez-Paz et al. [2013] proposed the Randomized Dependence Coefficient (RDC) which maps the two input samples to their respective empirical copula transformations before carrying out the RCCA procedure to obtain the largest canonical correlation. Large sample consistency results of RDC to KCCA were also established in the same work. Under certain normality assumptions, Barlett’s approximation [Mardia et al., 1979] can be used to provide closed form expression for the asymptotic null distribution of

the RDC statistic, which results in a distribution free test.

While the measures in Table 4.1 consider specific functions of the eigenspectrums of the (normalised) cross-covariance operator in the RKHS, Dauxois and Nkiet [1998] introduced a larger class of functions defined on the correlation operator spectrum in a general Hilbert space. More specifically, Dauxois and Nkiet [1998] considered dependence measures that are symmetric nondecreasing functions of the spectrum of appropriate Hilbert space operators. Since they did not use formalism of reproducing kernel Hilbert spaces and hence resort to projections onto n orthonormal basis functions of L^2 on the domains of X and Y (e.g. B-Splines basis [Dauxois and Nkiet, 1998, Example 4.2]) to obtain an approximation of the population dependence measure. Such approximated measure requires further finite sample approximations when only m i.i.d. samples are observed. Consistency results of the approximated measure of dependence to its population counterpart are established in [Dauxois and Nkiet, 1998, Proposition 4.1 and 5.1]. They showed that empirical centering is needed to avoid the trivial eigenvector associated with the eigenvalue 1 [Dauxois and Nkiet, 1998, Lemma 5.1]. The dependency measure is asymptotically normal under the alternative hypothesis [Dauxois and Nkiet, 1998, Proposition 5.2] and is asymptotically distributed as a chi-square random variable under the null hypothesis [Dauxois and Nkiet, 1998, Proposition 5.3]. Such asymptotic null distribution only depends on the dimensions of orthonormal basis functions and it is hence distribution free.

HSIC has been shown to have better statistical power comparing to COCO in Gretton et al. [2005a], while Fukumizu et al. [2008] showed statistical tests based on the Hilbert Schmidt norm of NOCCO have superior performance comparing to HSIC. This motivates our study in this chapter to establish a computationally efficient normalised kernel dependence test. In this chapter, we propose a large-scale approximation of the squared Hilbert Schmidt norm of NOCCO: I^{NOOCO} utilising random Fourier features approximation of the kernel matrices. For convenience, we will refer to I^{NOOCO} as the NOCCO statistic and its large-scale approximation as the RFF NOCCO statistic. The derived RFF NOCCO statistic reassembles that of a Hotelling's T statistic and subsequently enables the derivation of the asymptotic null distribution of the RFF NOCCO statistic. This results in a distribution free test and hence provides significant speed-ups comparing to the kernel dependence test based on NOCCO statistic.

4.2 Empirical NOCCO Statistic

Given a paired independently identically distributed sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ with each $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, denote the chosen kernels $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ respectively. Fukumizu et al. [2008] provided an empirical estimate for NOCCO:

$$\hat{V}_{XY}^{(m)} = (\hat{\Sigma}_{XX}^{(m)} + \epsilon_n I)^{-1/2} \hat{\Sigma}_{XY}^{(m)} (\hat{\Sigma}_{YY}^{(m)} + \epsilon_n I)^{-1/2} \quad (4.7)$$

where $\hat{\Sigma}_{XY}^{(m)}$ is estimated as

$$\hat{\Sigma}_{XY}^{(m)} = \frac{1}{m} \sum_{i=1}^m \left(k_{\mathcal{X}}(\cdot, x_i) - \frac{1}{m} \sum_{j=1}^m k_{\mathcal{X}}(\cdot, x_j) \right) \left(k_{\mathcal{Y}}(\cdot, y_i) - \frac{1}{m} \sum_{j=1}^m k_{\mathcal{Y}}(\cdot, y_j) \right)^T. \quad (4.8)$$

$\hat{\Sigma}_{XX}^{(m)}$ and $\hat{\Sigma}_{YY}^{(m)}$ are similarly defined. $\epsilon_n > 0$ is a sequence such that $\epsilon_n \rightarrow 0$ as $n \rightarrow \infty$. With this empirical estimation of the (cross-) covariance operators, the NOCCO statistic can be written as

$$\hat{I}^{NOCCO}(X, Y) := \|\hat{V}_{XY}^{(m)}\|_{HS}^2 \quad (4.9)$$

$$= \text{Tr}(R_X R_Y) \quad (4.10)$$

$$= \text{Tr}(G_X (G_X + m\epsilon_m I_m)^{-1} G_Y (G_Y + m\epsilon_m I_m)^{-1}) \quad (4.11)$$

where G_X is the centred gram matrices such that

$$G_{X,ij} = \left\langle k_{\mathcal{X}}(\cdot, X_i) - \frac{1}{m} \sum_{l=1}^m k_{\mathcal{X}}(\cdot, x_l), k_{\mathcal{X}}(\cdot, X_j) - \frac{1}{m} \sum_{l=1}^m k_{\mathcal{X}}(\cdot, x_l) \right\rangle_{\mathcal{H}_{\mathcal{X}}} \quad (4.12)$$

and G_Y is defined similarly. Direct computation of the NOCCO test statistic will lead to a heavy computational cost of $\mathcal{O}(m^3)$ due to inversion of kernel matrices. To use such statistic in independence testing, one needs to further appeal to permutation approach in order to obtain samples from the null distribution. Such test becomes prohibitive to use in practise since it scales as $\mathcal{O}(\text{number of shuffles} \times m^3)$. This prohibitive computational cost motivates our work to scale up the NOCCO statistic and propose a distribution free independence test to enable its use in the large data regime.

4.3 Random Fourier Features (RFF) Approximation

We have seen in Section 3.4 from the previous chapter, HSIC can be scaled up through low rank approximations of the kernel matrices. When the chosen kernels are shift

invariant (e.g. the Gaussian RBF kernel), Random Fourier Features (RFF) type of approximation can be used. For more general types of kernels, Nyström type of approximations can be used. In this chapter, we focus on RFF approximation of the NOCCO statistic and note that the Nyström approximation of the NOCCO statistic follows similarly as the derivations presented here.

Suppose X and Y are mapped to finite dimensional features, the random Fourier feature maps $\phi(X) \in \mathbb{R}^{m_x}$ and $\psi(Y) \in \mathbb{R}^{m_y}$ respectively, we will consider the following population quantities. Let $\mu_X := \mathbb{E}(\phi(X))$ and $\mu_Y := \mathbb{E}(\psi(Y))$ denote the respective kernel mean embeddings. Similarly, let $\mu_{XY} := \mathbb{E}[\phi(X)\psi(Y)^T]$ denote the kernel mean embedding of the product of the features. Further, we denote the centred feature maps as: $\phi^c(X) := \phi(X) - \mu_X$ and $\psi^c(Y) := \psi(Y) - \mu_Y$. With this notation, the cross-covariance matrix can be written as $\Sigma_{XY} = \mathbb{E}[\phi^c(X)\psi^c(Y)^T]$ and the covariance matrices can be written as $\Sigma_{XX} = \mathbb{E}[\phi^c(X)\phi^c(X)^T]$ and $\Sigma_{YY} = \mathbb{E}[\psi^c(Y)\psi^c(Y)^T]$.

Define the population-centred “joint features”:

$$\xi^c(X, Y) := \text{vec}(\phi^c(X)\psi^c(Y)^T) \in \mathbb{R}^{m_x m_y}. \quad (4.13)$$

where $\text{vec}(\cdot)$ denote the vectorisation of a matrix¹. Then,

$$\mathbb{E}[\xi^c(X, Y)] = \text{vec}(\mu_{XY} - \mu_X \mu_Y^T) =: \Delta_{XY} \quad (4.14)$$

$$\mathbb{V}\text{ar}[\xi^c(X, Y)] = \mathbb{E}[\xi^c(X, Y)\xi^c(X, Y)^T] - \Delta_{XY}\Delta_{XY}^T =: \Sigma_{XY, XY}^\xi. \quad (4.15)$$

Under the null hypothesis of $P_{XY} = P_X P_Y$, $\Delta_{XY} = 0$ and

$$\begin{aligned} \Sigma_{XY, XY} &= \mathbb{E}[\text{vec}(\phi^c(X)\psi^c(Y)^T)\text{vec}(\phi^c(X)\psi^c(Y)^T)^T] \\ &= \mathbb{E}[\psi^c(Y)\psi^c(Y)^T \otimes \phi^c(X)\phi^c(X)^T] \\ &= \Sigma_{YY} \otimes \Sigma_{XX}. \end{aligned} \quad (4.16)$$

By multidimensional Central Limit Theorem [van der Vaart, 1998],

$$\frac{1}{\sqrt{m}} \sum_{i=1}^m \xi^c(x_i, y_i) \xrightarrow{D} \mathcal{N}(0, \Sigma_{YY} \otimes \Sigma_{XX}). \quad (4.17)$$

By Lemma 4.6.1, (4.17) becomes

$$\frac{1}{\sqrt{m}} \sum_{i=1}^m (\Sigma_{YY}^{-1/2} \otimes \Sigma_{XX}^{-1/2}) \xi^c(x_i, y_i) \xrightarrow{D} \mathcal{N}(0, I_{m_x m_y}). \quad (4.18)$$

¹The vectorisation of a $m \times n$ matrix A is a linear transformation that convert the matrix into a $mn \times 1$ vector.

Now, let's consider the empirical version of the above. Let the empirical estimates of the kernel mean embeddings be $\hat{\mu}_x := \frac{1}{m} \sum_{i=1}^m \phi(x_i)$ and $\hat{\mu}_y := \frac{1}{m} \sum_{i=1}^m \psi(y_i)$. As before, we can then define the empirically centred feature maps $\hat{\phi}^c(x) := \phi(x) - \hat{\mu}_x$ and $\hat{\psi}^c(y) := \psi(y) - \hat{\mu}_y$. The empirical covariance operators are hence $\hat{\Sigma}_{XX} = \frac{1}{m} \sum_{i=1}^m \hat{\phi}^c(x_i) \hat{\phi}^c(x_i)^T$ and $\hat{\Sigma}_{YY} = \frac{1}{m} \sum_{i=1}^m \hat{\psi}^c(y_i) \hat{\psi}^c(y_i)^T$. Similarly, we can define the empirical cross-covariance operator $\hat{\Sigma}_{XY} = \frac{1}{m} \sum_{i=1}^m \hat{\phi}^c(x_i) \hat{\psi}^c(y_i)^T$. The empirically centred "joint features" is $\hat{\xi}^c(X, Y) := \text{vec}(\hat{\phi}^c(X) \hat{\psi}^c(Y)^T) \in \mathbb{R}^{m_x m_y}$. We now establish the asymptotic null distribution of the RFF approximated NOCCO statistic in the following Theorem.

Theorem 4.3.1. (Asymptotic Null Distribution of Approximated NOCCO)

Under the null hypothesis, let the dataset $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \stackrel{i.i.d.}{\sim} P_{XY} = P_X P_Y$, with $P_X \in \mathcal{M}_{k_X}^1(\mathcal{X})$ and $P_Y \in \mathcal{M}_{k_Y}^1(\mathcal{Y})$. Provided the regularisation constant $\epsilon_n \rightarrow 0$ as $n \rightarrow \infty$, then

$$\tilde{I}^{NOCCO}(X, Y) = m \left\| \left(\hat{\Sigma}_{XX} + \epsilon_n I \right)^{-1/2} \hat{\Sigma}_{XY} \left(\hat{\Sigma}_{YY} + \epsilon_n I \right)^{-1/2} \right\|_F^2 \xrightarrow{D} \chi_{m_x m_y}^2$$

where m_x and m_y are the number of random features for X and Y respectively.

The proof of the above theorem will be presented as follows. For the ease of notation, we denote $\tilde{\Sigma}_{XX} := \hat{\Sigma}_{XX} + \epsilon_n I$ and $\tilde{\Sigma}_{YY} := \hat{\Sigma}_{YY} + \epsilon_n I$. Since all the empirical operators are finite dimensional, the Hilbert Schmidt norm is equivalent to the Frobenius norm, hence the empirically estimated NOCCO statistic is

$$\begin{aligned} \tilde{I}^{NOCCO}(X, Y) &= m \left\| \tilde{\Sigma}_{XX}^{-1/2} \hat{\Sigma}_{XY} \tilde{\Sigma}_{YY}^{-1/2} \right\|_F^2 \end{aligned} \quad (4.19)$$

$$= m \left\| \tilde{\Sigma}_{XX}^{-1/2} \left[\frac{1}{m} \sum_{i=1}^m \phi(x_i) \psi(y_i)^T - \frac{1}{m^2} \left(\sum_{j=1}^m \phi(x_j) \right) \left(\sum_{l=1}^m \psi(y_l) \right)^T \right] \tilde{\Sigma}_{YY}^{-1/2} \right\|_F^2 \quad (4.20)$$

Multiplying the empirical covariance operators from both sides, the vectorised first term inside the norm becomes:

$$\frac{1}{m} \sum_{i=1}^m \text{vec} \left(\tilde{\Sigma}_{XX}^{-1/2} \phi(x_i) \psi(y_i)^T \tilde{\Sigma}_{YY}^{-1/2} \right) \quad (4.21)$$

$$= \frac{1}{m} \sum_{i=1}^m \left(\tilde{\Sigma}_{YY}^{-1/2} \otimes \tilde{\Sigma}_{XX}^{-1/2} \right) \text{vec}(\phi(x_i) \psi(y_i)^T) \quad (4.22)$$

using the relationship between Kronecker product and vectorisation² [Schäcke, 2013]. Similarly, the vectorised second term in the norm becomes

$$-\frac{1}{m} \sum_{i=1}^m \text{vec} \left(\tilde{\Sigma}_{XX}^{-1/2} \hat{\mu}_x \hat{\mu}_y^T \tilde{\Sigma}_{YY}^{-1/2} \right) \quad (4.23)$$

$$= -\frac{1}{m} \sum_{i=1}^m \left(\tilde{\Sigma}_{YY}^{-1/2} \otimes \tilde{\Sigma}_{XX}^{-1/2} \right) \text{vec} \left(\hat{\mu}_x \hat{\mu}_y^T \right). \quad (4.24)$$

Since taking the Frobenius norm of the matrix is the same as that of the vectorised matrix, the expression of NOCCO in (4.20) becomes

$$m \left\| \left(\tilde{\Sigma}_{YY}^{-1/2} \otimes \tilde{\Sigma}_{XX}^{-1/2} \right) \frac{1}{m} \sum_{i=1}^m \left\{ \text{vec}(\phi(x_i) \psi(y_i)^T) - \text{vec}(\hat{\mu}_x \hat{\mu}_y^T) \right\} \right\|_F^2 \quad (4.25)$$

$$= m \left\| \left(\tilde{\Sigma}_{YY}^{-1/2} \otimes \tilde{\Sigma}_{XX}^{-1/2} \right) \text{vec}(\hat{\Sigma}_{XY}) \right\|_F^2 \quad (4.26)$$

$$= m \left\| \left(\tilde{\Sigma}_{YY}^{-1/2} \otimes \tilde{\Sigma}_{XX}^{-1/2} \right) \left(\frac{1}{m} \sum_{i=1}^m \hat{\xi}^c(x_i, y_i) \right) \right\|_F^2. \quad (4.27)$$

Note, the empirical covariance matrices $\hat{\Sigma}_{XX}$ and $\hat{\Sigma}_{YY}$ as we introduced here are the maximum likelihood estimators (MLE) of the population covariance matrices Σ_{XX} and Σ_{YY} . Hence, the consistency of MLEs implies $\tilde{\Sigma}_{XX} \xrightarrow{a.s.} \Sigma_{XX}$ and $\tilde{\Sigma}_{YY} \xrightarrow{a.s.} \Sigma_{YY}$ [Young and Smith, 2005] as $\epsilon_n \rightarrow 0$. Since $\hat{\mu}_x$ and $\hat{\mu}_y$ are also the MLEs, $\hat{\mu}_x \xrightarrow{a.s.} \mu_x$ and $\hat{\mu}_y \xrightarrow{a.s.} \mu_y$. As a consequence of the continuous mapping theorem (CMT), $\hat{\xi}^c(x_i, y_i) \xrightarrow{a.s.} \xi^c(x_i, y_i)$ and subsequently $\frac{1}{m} \sum_{i=1}^m \hat{\xi}^c(x_i, y_i) \xrightarrow{a.s.} \frac{1}{m} \sum_{i=1}^m \xi^c(x_i, y_i)$ [Shao, 2007].

Since random matrices are random vectors with elements arranged in columns, hence by CMT we obtain $\tilde{\Sigma}_{YY}^{-1/2} \otimes \tilde{\Sigma}_{XX}^{-1/2} \xrightarrow{a.s.} \Sigma_{YY}^{-1/2} \otimes \Sigma_{XX}^{-1/2}$ and

$$\tilde{f}^{\text{NOCCO}}(X, Y) \xrightarrow{a.s.} \left\| \sqrt{m} \left(\Sigma_{YY}^{-1/2} \otimes \Sigma_{XX}^{-1/2} \right) \left(\frac{1}{m} \sum_{i=1}^m \xi^c(x_i, y_i) \right) \right\|_2^2. \quad (4.28)$$

From the analysis of population quantities, recall from (4.17), $\frac{1}{\sqrt{m}} \sum_{i=1}^m \xi^c(x_i, y_i)$ is a normally distributed, zero mean Gaussian vector with covariance matrix $\Sigma_{XY, XY} = \Sigma_{YY} \otimes \Sigma_{XX}$. Hence, the Hotelling's T^2 statistic [Hotelling, 1931] of such Gaussian vector is $T^2 = m \left(\frac{1}{m} \sum_{i=1}^m \xi^c(x_i, y_i) \right)^T \Sigma_{XY, XY}^{-1} \left(\frac{1}{m} \sum_{i=1}^m \xi^c(x_i, y_i) \right)$ which converges in

²For matrices $A(m \times n)$, $B(p \times q)$ and vectors $a(n \times 1)$, $b(p \times 1)$, then $\text{vec}(Aab^T B) = (B^\top \otimes A) \text{vec}(ab^T)$.

distribution to a χ^2 distribution with $m_x m_y$ degrees of freedom. Note that

$$\begin{aligned} & \left\| \sqrt{m} \left(\Sigma_{YY}^{-1/2} \otimes \Sigma_{XX}^{-1/2} \right) \left(\frac{1}{m} \sum_{i=1}^m \xi^c(x_i, y_i) \right) \right\|_2^2 \\ &= \text{Tr} \left[m \left(\frac{1}{m} \sum_{i=1}^m \xi^c(x_i, y_i) \right)^T \left(\Sigma_{YY}^{-1/2} \otimes \Sigma_{XX}^{-1/2} \right)^T \left(\Sigma_{YY}^{-1/2} \otimes \Sigma_{XX}^{-1/2} \right) \left(\frac{1}{m} \sum_{i=1}^m \xi^c(x_i, y_i) \right) \right] \end{aligned} \quad (4.29)$$

$$= \text{Tr} \left[m \left(\frac{1}{m} \sum_{i=1}^m \xi^c(x_i, y_i) \right)^T (\Sigma_{YY} \otimes \Sigma_{XX})^{-1} \left(\frac{1}{m} \sum_{i=1}^m \xi^c(x_i, y_i) \right) \right] \quad (4.30)$$

$$= m \left(\frac{1}{m} \sum_{i=1}^m \xi^c(x_i, y_i) \right)^T \Sigma_{XY, XY}^{-1} \left(\frac{1}{m} \sum_{i=1}^m \xi^c(x_i, y_i) \right) \quad (4.31)$$

$$= T^2 \sim \chi_{m_x m_y}^2 \quad (4.32)$$

where (4.30) follows due to Lemma 4.2.10 and Corollary 4.2.11 in [Horn and Johnson, 1991]. Hence, $\tilde{I}^{NOCCO}(X, Y) \xrightarrow{D} \chi_{m_x m_y}^2$ [van der Vaart, 1998].

Theorem 4.3.1 shows that the asymptotic null distribution of the RFF NOCCO statistic is independent of the kernel used and naturally give rise to a distribution free independence test. The RFF NOCCO statistic can be directly compared with the $(1-\alpha)\%$ quantile of $\chi_{m_x m_y}^2$ distribution, where α is the significance level, which avoids the eigendecomposition required in spectral approach and the repetitive computation of the test statistic in the permutation approach. This results in a fast independence test for large samples.

4.3.1 Relationship with Dauxois and Nkiet [1998]

As discussed in Section 4.1.2, the framework of Dauxois and Nkiet [1998] considers general symmetric nondecreasing functions of the spectrum of appropriate Hilbert space operators. Although no explicit RKHS perspective was considered, the function considered in their Example 4.3 is in fact a function of the proposed RFF NOCCO statistic. Given the notations from Dauxois and Nkiet [1998], we will show that the Proposition 5.3 of their work is a more general version of Theorem 4.3.1 as we derived above. However, the proof provided by them is more complicated due to its generality.

Given an iid sample $(X_i, Y_i)_{1 \leq i \leq m}$ and n orthonormal basis functions, $r_\Phi^{n,m}(X, Y)$ denotes the measure of dependence. It is equivalently written as $\Phi_n(\lambda_{n,m})$, the sym-

metric nondecreasing function on the spectrum $\lambda_{n,m}$ of the operator. Note, the subscripts n, m are not the indices of the singular values, n indicates the number of basis functions and m represents the number of samples. The operator is the usual operator for linear canonical correlation analysis

$$R_{n,m} = (V_1^{n,m})^{-1/2} V_{12}^{n,m} (V_2^{n,m})^{-1} V_{21}^{n,m} (V_1^{n,m})^{-1/2} \quad (4.33)$$

$$\text{with } V_1^{n,m} = \frac{1}{m} \sum_{i=1}^m \phi^n(X_i) \otimes \phi^n(X_i) \quad (4.34)$$

$$V_2^{n,m} = \frac{1}{m} \sum_{i=1}^m \psi^n(Y_i) \otimes \psi^n(Y_i) \quad (4.35)$$

$$V_{12}^{n,m} = \frac{1}{m} \sum_{i=1}^m \psi^n(Y_i) \otimes \phi^n(X_i) = (V_{21}^{n,m})^* \quad (4.36)$$

where $(\phi_i^n)_{1 \leq i \leq p_n}$ is a linearly independent system that span a subspace $\mathcal{V}_X^n \subset L^2(P_X)$ for $n \in \mathbb{N}$ and $\phi^n(X) = (\phi_1^n(X), \dots, \phi_{p_n}^n(X))^\top$. $(\psi_i^n)_{1 \leq i \leq q_n}$ and $\psi^n(Y)$ are similarly defined. Additionally, the systems $(\phi_i^n)_{1 \leq i \leq p_n}$ and $(\psi_i^n)_{1 \leq i \leq q_n}$ are orthonormal in $L^2(P_X)$ and $L^2(P_Y)$ respectively with $p_n < q_n$. K_{Φ_n} is the real value denoting the derivative of Φ_n evaluated at 0.

Proposition 5.3 of Dauxois and Nkiet [1998] shows that $mK_{\Phi_n}^{-1} r_{\Phi}^{n,m}(X, Y)$ converges in distribution to $\chi_{(p_n-1)(q_n-1)}^2$ as the number of samples $m \rightarrow \infty$. They argue that the operator $R_{n,m}$ contains the trivial eigenvalue of 1, so instead they work with the centered operator:

$$S_{n,m} = R_{n,m} - [(V_1^{n,m})^{1/2} \alpha^n] \otimes [(V_1^{n,m})^{1/2} \alpha^n] \quad (4.37)$$

where $(V_1^{n,m})^{1/2} \alpha^n$ is a unit eigenvector of $R_{n,m}$. They showed in the proof of Proposition 5.3, that for a general symmetric nondecreasing function $\Phi_n(\cdot)$, under the null hypothesis

$$mK_{\Phi_n}^{-1} \Phi_n(\lambda_{n,m}) = \text{Tr}(mS_{n,m}). \quad (4.38)$$

Recall, the empirical NOCCO is defined as $m\|\tilde{\Sigma}_{XX}^{-1/2} \hat{\Sigma}_{XY} \tilde{\Sigma}_{YY}^{-1/2}\|_F^2$. By the definition of Frobenius norm, this is exactly the trace of the scaled operator $R_{n,m}$ in the limit $\epsilon_n \rightarrow 0$:

$$\hat{I}^{NOCCO} = \text{Tr} \left(\tilde{\Sigma}_{XX}^{-1/2} \hat{\Sigma}_{XY} \tilde{\Sigma}_{YY}^{-1/2} \tilde{\Sigma}_{YY}^{-1/2} \hat{\Sigma}_{YX} \tilde{\Sigma}_{XX}^{-1/2} \right) \quad (4.39)$$

$$= \text{Tr}(mR_{n,m}). \quad (4.40)$$

While our derivation for the asymptotic null distribution [Theorem 4.3.1] shows that $\text{Tr}(mR_{n,m})$ converges to $\chi^2_{(p_n)(q_n)}$ with $p_n = m_x$ and $q_n = m_y$ the number of random features for X and Y respectively, their results says if we remove the trivial eigenvector, a $\chi^2_{(p_n-1)(q_n-1)}$ is obtained as an asymptotic null distribution. Essentially, we provided a simpler proof of the specific case, where Φ is the sum of squared singular values, within their general framework utilising machineries from Hotelling’s T statistic. Although our random Fourier features is not an orthonormal basis, it seems for fixed finite basis, the test of Dauxois and Nkiet [1998] always takes the exact same form as the RFF NOCCO test we proposed here. For the random Fourier features considered in our work, it could be computationally inefficient to explicitly compute the centering of the squared Frobenius norm of empirical NOCCO and the asymptotic distribution may no longer hold.

4.4 Experiments

In a series of synthetic data experiments, we examine the statistical performance of the NOCCO test and the proposed RFF NOCCO test vs. HSIC test and RFF HSIC test. Through the rotational experiment from Fukumizu et al. [2008] and its variance, we investigate the robustness of NOCCO based tests to the choice of kernel length-scale parameter. In addition, we repeat the experiments from the previous chapter to investigate the computational saving of the proposed RFF NOCCO test.

The significance level is set to 5% and Gaussian kernels are used in all experiments. Unless otherwise stated, median heuristic is used to estimate the bandwidth of the kernel. The regularisation parameter ϵ_n is fixed at 10^{-6} . The 95% confidence intervals of the rejection rates are computed based on normality assumption: $\hat{\mu} \pm 1.96 \times \sqrt{\frac{\hat{\mu}(1-\hat{\mu})}{100}}$ with $\hat{\mu}$ the number of rejections out of 100 repetition.

4.4.1 Robustness to Kernel Length-Scale

We consider the rotational experiment from Fukumizu et al. [2008]. The random variables $X^{(0)}$ and $Y^{(0)}$ are independent and uniformly distributed on $[-2, 2]$ and $[0.5, 1.5] \cup [-1.5, -0.5]$ respectively. $(X^{(\theta)}, Y^{(\theta)})$ is the rotation of $(X^{(0)}, Y^{(0)})$ by an angle $\theta \in [0, \pi/4]$ as illustrated in Figure 4.1. $X^{(\theta)}$ and $Y^{(\theta)}$ are dependent when $\theta \neq 0$. The dependence becomes easier to detect as the angle of rotation increases. In addition to this 1-dimensional experiment, we consider the same experiment in 3 dimensions. In the 3-dimensional case, denote $X^{(\theta)} = [X_1^{(\theta)}, X_2, X_3]$ and $Y^{(\theta)} = [Y_1^{(\theta)}, Y_2, Y_3]$,

$X_1^{(\theta)}$ and $Y_1^{(\theta)}$ are distributed as in the 1-dimensional experiment while the second and third dimensions are unrotated and distributed as $X_2, X_3 \stackrel{i.i.d.}{\sim} \text{Uniform}[-2, 2]$ and $Y_2, Y_3 \stackrel{i.i.d.}{\sim} \text{Uniform}[0.5, 1.5] \cup [-1.5, -0.5]$. The number of random Fourier features is fixed at 10 and 100 sets of 200 data points are generated for each experiment. Permutation tests are used for NOCCO statistic $\hat{I}^{NOCCO}$ and HSIC $\hat{\Xi}$, while spectral test is used for RFF HSIC and distribution free test is used for RFF NOCCO statistic.

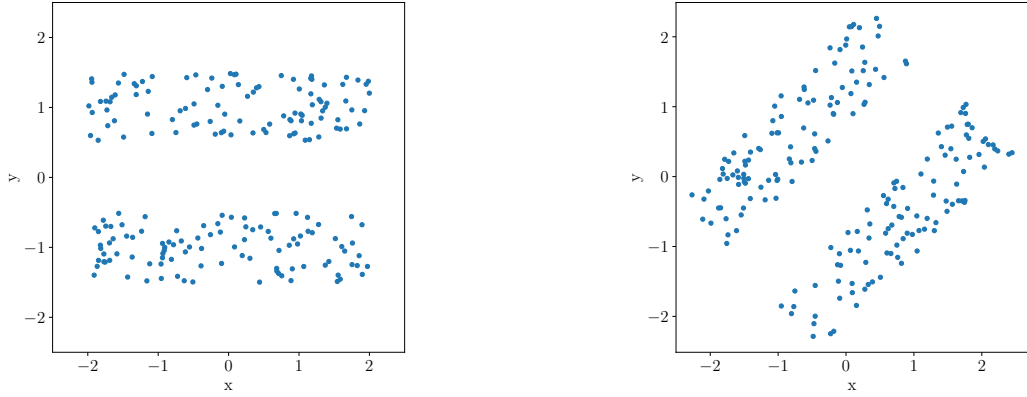


Figure 4.1: Visualisation of data from rotational experiment in Section 4.4.1. Left: $\theta = 0$. Right: $\theta = \pi/4$.

Figure 4.2 shows when $\theta = 0$, i.e. $X^{(0)}$ is independent of $Y^{(0)}$, all methods have correct Type I control. As the rotation angle increases, the problem becomes easier and the statistical power is consistently increasing for all methods. In both dimensions, we see that the exact NOCCO statistic gives the best statistical performance followed by the RFF NOCCO and HSIC and its RFF approximations.

Using median heuristic, the dependence measure based on NOCCO seems to have superior performance comparing to HSIC. We would like to investigate further the effect of the bandwidth on the performance of the four methods. When median heuristic is used, the estimated bandwidth is approximately 0.8 for both X and Y in 1D and approximately 1.7 in 3D. Figure 4.3 shows the same experiment in 1 dimension with the kernel lengthscale being smaller (left) and larger (right) than the estimation provided by median heuristic. When the lengthscale is 0.1, HSIC has superior performance comparing to NOCCO based methods, while the RFF HSIC gives the worse performance. This suggests that NOCCO statistic suffers less from the RFF approximation. When the lengthscale is much larger than that estimated by median

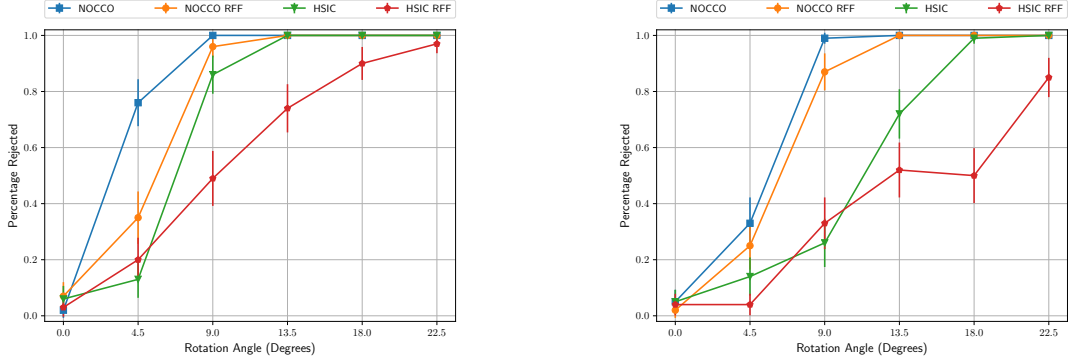


Figure 4.2: Rotational experiment comparing HSIC permutation approach, RFF HSIC spectral method, NOCCO permutation approach, RFF NOCCO with distribution free test. Left: Rotational experiment in 1D. Right: Rotational experiment in 3D.

heuristic, NOCCO and NOCCO RFF perform similarly to when the lengthscale is estimated by median heuristic. However, HSIC and RFF HSIC are struggling to detect the dependence.

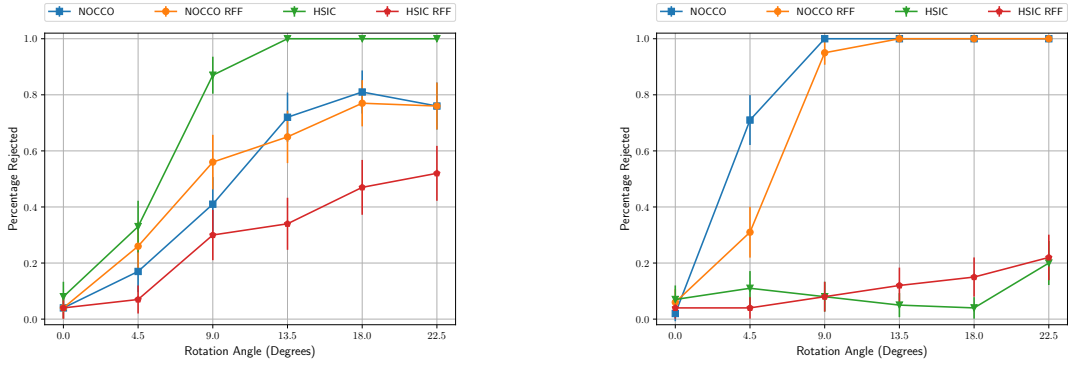


Figure 4.3: Rotational experiment in 1 dimension. Gaussian kernel with length-scale: 0.1 (Left) and 2 (Right).

In 3 dimensions, Figure 4.4 shows the rejection rate using 0.5 and 5 as lengthscale. When the lengthscale is smaller than that estimated by the median heuristic, HSIC again provides the best performance while HSIC RFF and NOCCO based methods all perform similarly. When the lengthscale is 5, the performance of all methods are similar to that observed in Figure 4.2 [Right]. However, we can still notice the slight decrease in performance of HSIC at angles $\theta = \{4.5, 9, 13.5, 18\}$.

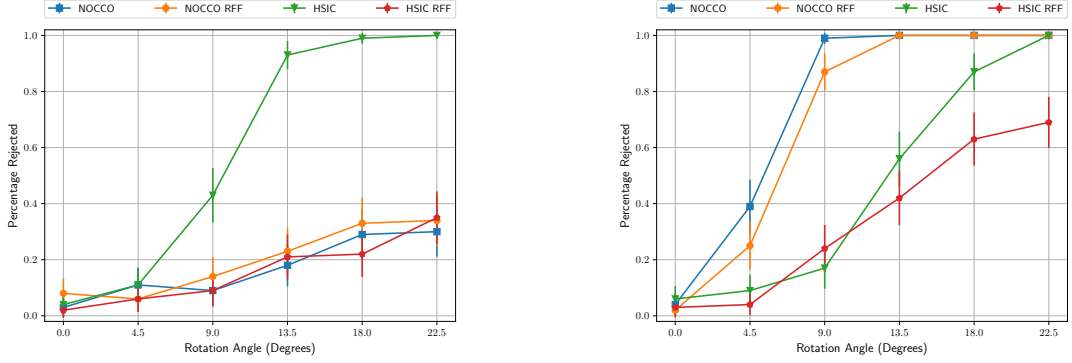


Figure 4.4: Rotational experiment in 3 dimensions. Gaussian kernel with length-scale: 0.5 (Left) and 5 (Right).

Overall, NOCCO based tests appear more robust to the misspecification of the bandwidth parameter. Fukumizu et al. [2008] showed in Theorem 4 that for sufficiently expressive kernels like Gaussian and Laplacian kernels, the population I^{NOCCO} has the remarkable property that it is independent of the kernels. More specifically, it can be written in terms of integral expressions containing only the probability density functions, this recovers the form of the mean square contingency. Although any empirical finite samples estimates depends on the choice of kernels, Fukumizu et al. [2008] showed that such empirical dependence measure converges to I^{NOCCO} in Hilbert Schmidt norm. It makes intuitive sense that NOCCO based tests are more robust to kernel choice. Even though HSIC seems to be particularly sensitive to the choice of the lengthscale parameter, it is able to outperform NOCCO based tests when the correct lengthscale parameter is chosen. Intuitively, this is to be expected as the NOCCO statistic is effectively adding a variance regulariser that could have some crude smoothing effect. Despite this loss of power, its robustness to lengthscale choice can be advantageous in real life applications.

4.4.2 Computational Saving from Distribution Free Test

We repeat the synthetic experiments from Section 5.4 to investigate the performance of NOCCO based dependence tests from the perspective of computational efficiency. We include the more interesting results in the main text while the results from simple linear experiment are included in Appendix 4.6.2. Recall, the bandwidth parameter was set using median heuristic. With this choice of bandwidth parameter, HSIC based methods provide better performance comparing to NOCCO based methods in

all three experiments.

In the sine dependency experiment, due to the available computational resources, we have set a maximum running time of 2×10^3 seconds. NOCCO permutation test with a samples size $\{2, 4\} \times 10^3$ was unable to complete within this time limit and hence the broken line in Figure 4.5. While the proposed RFF NOCCO test requires more samples to achieve the same power performance as HSIC based tests, it is able to provide better power performance for a given computation time. This illustrates the significant advantage of having a distribution free test.

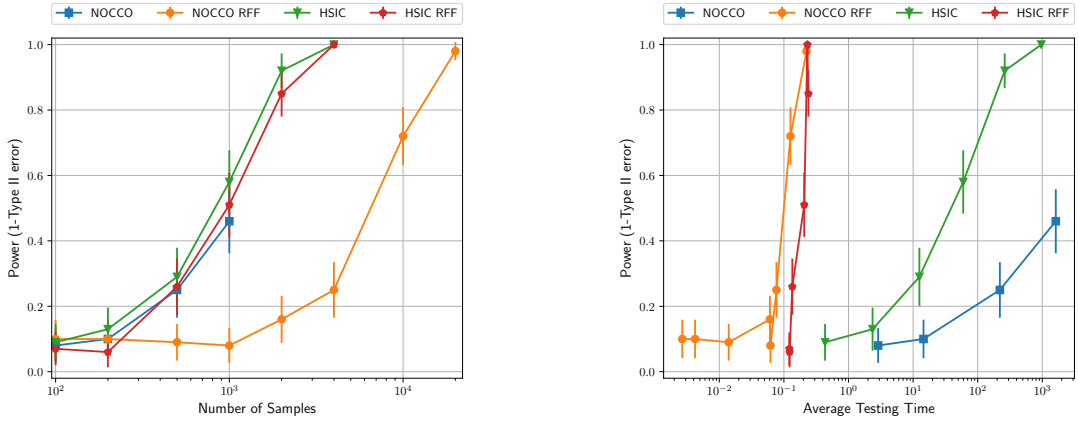


Figure 4.5: Sine Dependency Experiment for $d = 2$ comparing HSIC spectral approach, RFF HSIC spectral approach, NOCCO permutation approach and RFF NOCCO distribution free test. 50 random Fourier features were used. Left: The test power as a function of the number of samples. Right: The test power as a function of the average testing time in seconds.

However, in the large scale experiment, we observe that RFF HSIC consistently achieve better performance per number of samples and per average testing time. The most significant cost in computational time that differentiates NOCCO and HSIC comes from the inversions of the covariance matrices in the NOCCO statistic. To ensure numerical stability, our implementation of the RFF NOCCO statistic requires explicit eigendecomposition. While computing the asymptotic null distribution for the spectral test of HSIC RFF requires the eigendecomposition of the same matrices, we note that only eigenvalues are required in the HSIC RFF case, however all the eigenpairs are required for the NOCCO RFF case. Our implementation in python

used `numpy.linalg.eigvalsh`³ for computing the eigenvalues in the asymptotic null distribution of RFF HSIC estimator and `numpy.linalg.eigh`⁴ for computing the eigenpairs in the NOCCO RFF statistic. When computing only the eigenvalues, `numpy.linalg.eigvalsh` uses the Pal-Walker-Kahan variant of the QR algorithm [Francis, 1961, 1962, Kublanovskaya, 1961], which scales as $\mathcal{O}(m_x^2)$ where $m_x = m_y$ is the number of random Fourier features used. When computing all the eigenpairs, `numpy.linalg.eigh` uses a divide and conquer algorithm [Cuppen, 1981], which scales as $4/3m_x^3 + \mathcal{O}(m_x^2)$. Such operation becomes significantly costly in terms of computational time when the number of features becomes larger, e.g. 200, as required in this experiment. We argue that more efficient implementation methods can be explored and we will leave it as future work since it is outside the scope of this work.

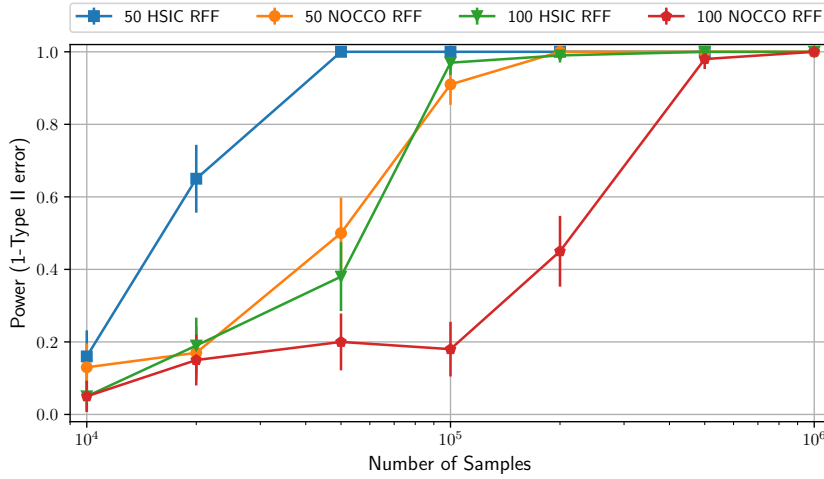


Figure 4.6: Large Scale Experiment: the statistical power as a function of samples size comparing the RFF HSIC spectral method and RFF NOCCO distribution free test. 50 stands for 50 dimensions and 100 stands for 100 dimensions. 200 random Fourier features were used.

From these comparison experiments, we see that although RFF NOCCO test is able to achieve higher power per computational time when the number of random features are small, the eigendecompositions required in computing the RFF NOCCO statistic become significantly more costly when larger number of random features is required (e.g. 200 as in this experiment). This outweighs the benefit of the distribution free test associated with RFF NOCCO and makes it less efficient than RFF

³<https://docs.scipy.org/doc/numpy-1.10.4/reference/generated/numpy.linalg.eigvalsh.html>

⁴<https://docs.scipy.org/doc/numpy/reference/generated/numpy.linalg.eigh.html>

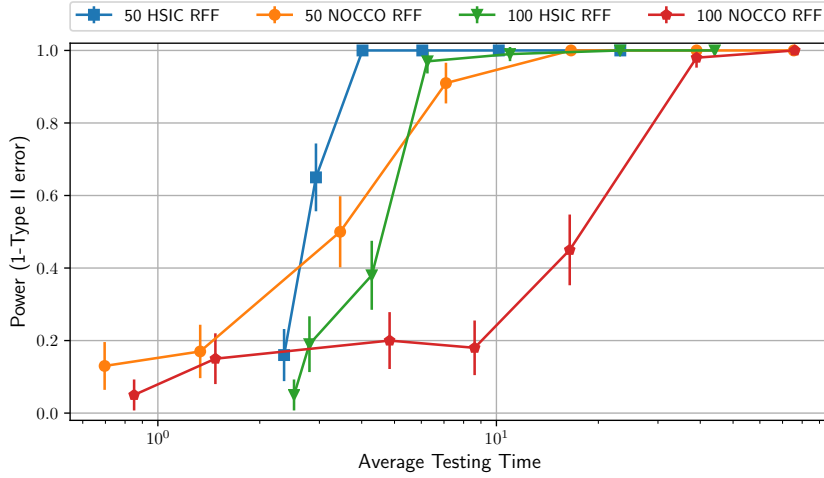


Figure 4.7: Large Scale Experiment: the statistical power comparison as a function of the average testing time in seconds.

HSIC in terms of power-time tradeoff.

4.5 Conclusion

We have proposed an RFF estimator of the NOCCO dependence measure and subsequently established a distribution free dependency test. The asymptotic null distribution derived for the RFF NOCCO statistic closely assembles that of Dauxois and Nkiet [1998]. However, we have provided a much simpler derivation utilising the form of Hotelling’s T^2 statistic which does not rely on an orthonormal basis. Nonetheless, their framework is general and can be adopted to derive the asymptotic distribution of the primal representation of the NOCCO statistic when certain assumptions on the feature maps are satisfied. Although asymptotic consistency to the population quantity was considered in Dauxois and Nkiet [1998], the authors never considered the consistency when the number of random Fourier feature increases (i.e. when p_n and q_n increases). A first step was made by Lopez-Paz et al. [2013] and this can be a nice direction for future work. Another interesting question to ask would be, if there is a RKHS equivalent to Dauxois and Nkiet [1998]’s system of deterministic orthonormal basis?

The proposed test significantly reduced computational complexity in memory and time over the standard NOCCO permutation test. We have verified the validity of the proposed large-scale method on a series of synthetic data experiments. We have

demonstrated the robustness of NOCCO based dependence test to bandwidth selection on the rotational experiment from Fukumizu et al. [2008]. Further, on the challenging high dimensional datasets from Chapter 3, we have illustrated the favourable trade-offs between testing power and computational complexity in the regime of a small number of Fourier features. However, when the number of Fourier features becomes large, the computation complexity of computing the RFF NOCCO statistic becomes significant which renders it less efficient than the RFF HSIC test. We argue different implementation methods can be investigated to ensure numerical stability while avoiding the explicit eigendecomposition in computing the RFF NOCCO statistic.

4.6 Appendix

4.6.1 Supporting Results

Lemma 4.6.1. *If $C = A \otimes B$, then $C^{-1/2} = A^{-1/2} \otimes B^{-1/2}$.*

Proof. Consider the singular value decomposition of A and B :

$$\begin{aligned} A &= U_A \Sigma_A V_A^* \\ B &= U_B \Sigma_B V_B^* \end{aligned}$$

This implies that

$$\begin{aligned} A^{-1/2} &= U_A \Sigma_A^{-1/2} V_A^* \\ B^{-1/2} &= U_B \Sigma_B^{-1/2} V_B^* \end{aligned}$$

Then by Theorem 2.15 in [Horn and Johnson, 1991],

$$\begin{aligned} A^{-1/2} \otimes B^{-1/2} &= (U_A \Sigma_A^{-1/2} V_A^*) \otimes (U_B \Sigma_B^{-1/2} V_B^*) \\ &= (U_A \otimes U_B) (\Sigma_A^{-1/2} \otimes \Sigma_B^{-1/2}) (V_A^* \otimes V_B^*) \end{aligned} \tag{4.41}$$

$$\begin{aligned} &= (U_A \otimes U_B) (\Sigma_A \otimes \Sigma_B)^{-1/2} (V_A \otimes V_B)^* \\ &= (A \otimes B)^{-1/2} = C^{-1/2} \end{aligned} \tag{4.42}$$

(4.42) follows since both Σ_A and Σ_B are diagonal matrices. □

4.6.2 Simple Linear Experiment

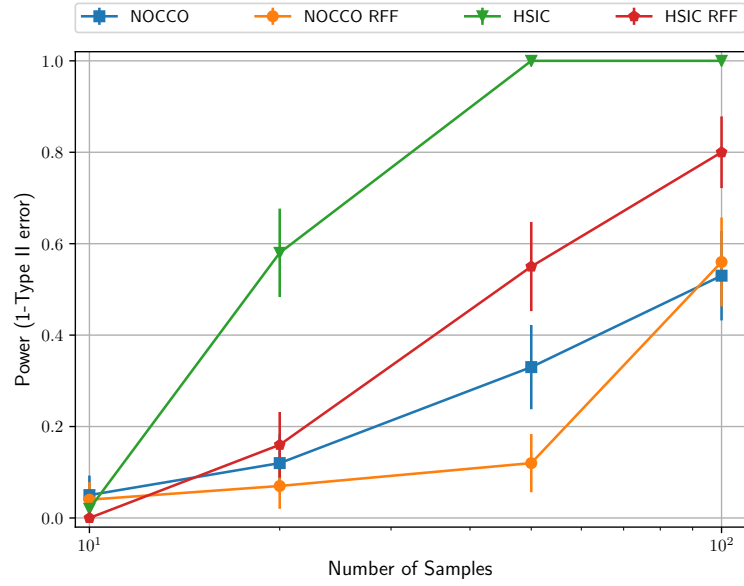


Figure 4.8: Simple linear experiment for $d = 10$ comparing NOCCO permutation approach and RFF NOCCO approach with HSIC spectral approach and RFF HSIC approach. 10 random Fourier features were used.

Chapter 5

Feature-to-Feature Regression for a Two-Step Conditional Independence Test

This chapter consists of a self-contained paper:

Q. Zhang, S. Filippi, S. Flaxman, and D. Sejdinovic, Feature-to-Feature Regression for a Two-Step Conditional Independence Test, in Uncertainty in Artificial Intelligence (UAI), 2017.

Abstract

The algorithms for causal discovery and more broadly for learning the structure of graphical models require well calibrated and consistent conditional independence (CI) tests. We revisit the CI tests which are based on two-step procedures and involve regression with subsequent (unconditional) independence test (RESIT) on regression residuals and investigate the assumptions under which these tests operate. In particular, we demonstrate that when going beyond simple functional relationships with additive noise, such tests can lead to an inflated number of false discoveries. We study the relationship of these tests with those based on dependence measures using reproducing kernel Hilbert spaces (RKHS) and propose an extension of RESIT which uses RKHS-valued regression. The resulting test inherits the simple two-step testing procedure of RESIT, while giving correct Type I control and competitive power. When used as a component of the PC algorithm, the proposed test is more robust to the case where hidden variables induce a switching behaviour in the associations present in the data.

5.1 Introduction

Conditional independence tests are an important component of causal discovery (cf. [Shalizi, 2016, Chapter 28]). For example, the popular PC algorithm [Spirtes et al., 2000] for recovering dependence structure among a set of variables starts from a complete undirected graph and recursively removes the edges between variables based on conditional independence testing. With normally distributed variables and linear relationships, conditional independence testing can be performed using estimates of partial correlation. Namely, when testing the hypothesis that $X \perp\!\!\!\perp Y|Z$, partial correlation is the correlation of the residuals of X and Y after linearly regressing each of them on Z separately. However, in the presence of nonlinearities and when one needs to condition on a random vector Z of larger dimensions, conditional independence testing is a challenging problem [Bergsma, 2004]. A popular approach for conditional independence testing within PC algorithm is RESIT (REgression with Subsequent Independence Test) [Hoyer et al., 2009, Peters et al., 2014], which can incorporate nonlinearities by extending the partial correlation approach in the following two ways: (1) it uses a flexible nonparametric regression of X and Y on Z , and (2) the subsequent test is based on a nonlinear dependence measure between the resulting residuals. While RESIT greatly broadens the class of models in which causal discovery with conditional independence tests is possible, as we will see, it is sensitive to departures from its modelling assumption, which itself is not straightforward to verify. In particular, it is likely to give an inflated number of false positives in the presence of associations which do not directly conform to functional relationships (as illustrated in Figure 5.1).

In the last decade, kernel embeddings of probability measures [Smola et al., 2007, Sriperumbudur et al., 2010] have been widely used to construct nonparametric hypothesis tests, including tests for the two-sample problem [Gretton et al., 2007, 2012a], independence [Gretton et al., 2008, Chwialkowski et al., 2015], conditional independence [Fukumizu et al., 2008, Zhang et al., 2011, Doran et al., 2014], three-variable interaction [Sejdinovic et al., 2013a], joint independence [Pfister et al., 2017], and goodness of fit [Chwialkowski et al., 2016]. The various kernel-based tests for conditional independence [Fukumizu et al., 2008, Zhang et al., 2011, Doran et al., 2014] allow to measure more general forms of conditional dependence than RESIT. However, these approaches typically involve complex statistics and computationally expensive procedures to estimate their distributions under the null hypothesis. In this paper

we propose a kernel-based conditional independence approach which aims to strike a balance between the simplicity of the RESIT approach and the robustness to the functional association and additive noise assumptions made by the RESIT.

The procedure we introduce is a test for *weak conditional independence* as defined by [Daudin, 1980] since, like RESIT, it focuses on individual effects¹ of the conditioning variable Z on X and Y . As such, it does not fully characterise conditional independence, but does benefit from a simpler testing procedure and an improved power in comparison to the “strong” tests which do characterise conditional independence. When these individual effects are not present and as such weak CI tests are insufficient to detect a complex joint effect on (X, Y) , they can be combined with tests for multivariate interaction [Sejdinovic et al., 2013a].

An independent recent work, Strobl et al. [2018], considers a very similar approach to ours – indeed, the method we propose is essentially equivalent to the RCoT approach of Strobl et al. [2018], which in addition operates on explicit primal representations of feature maps leading to a decreased computational cost. However, Strobl et al. [2018] do not comment on the connections with the two-step procedures for CI testing and RESIT. Hence, our contribution is to provide the unifying framework which places the proposed method as a generalisation of RESIT, pointing the deficiencies of RESIT and the interplay between its structural assumptions and nonlinear dependencies. In addition, while RCoT uses approximations to the null distributions with parametric families, we simply employ a permutation-based approach using the two-step interpretation of the proposed test, and do not observe any deviations from the desired significance level in the experiments. The test maintains correct Type I control even under the challenging switching associations induced by hidden variables, where RESIT is inappropriate.

In Section 5.2, we overview the key notion of weak conditional independence and the existing approach for testing it as well as three related tests for general conditional independence using kernel methods. In Section 5.3 we develop our novel procedure for weak conditional independence testing. Section 5.4 evaluates the performance of our approach and compares it to RESIT and other conditional independence tests on both synthetic and real-world data.

¹Since regressions of X and Y on Z are done separately and hence cannot capture the joint dependence of (X, Y) .

5.2 Background

We start by introducing notation and reviewing the existing kernel-based conditional independence tests. Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ be non-empty measurable spaces, with Borel σ -algebras $\mathcal{B}_{\mathcal{X}}$, $\mathcal{B}_{\mathcal{Y}}$ and $\mathcal{B}_{\mathcal{Z}}$ respectively. Let k , l and m be measurable positive definite kernels on these respective domains with the corresponding reproducing kernel Hilbert spaces (RKHSs) $\mathcal{H}_{\mathcal{X}}$, $\mathcal{H}_{\mathcal{Y}}$, $\mathcal{H}_{\mathcal{Z}}$. Let (X, Y, Z) be a triple of random variables with the joint probability law P_{XYZ} on $(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}, \mathcal{B}_{\mathcal{X}} \times \mathcal{B}_{\mathcal{Y}} \times \mathcal{B}_{\mathcal{Z}})$. We will assume $\mathbb{E}_{X \sim P_X} [k(X, X)] < \infty$ and similarly for l and m , which will ensure that $\mathcal{H}_{\mathcal{X}} \subset L^2_{\mathcal{X}}(P_X) = \{f : \mathbb{E}[f^2(X)] < \infty\}$, and similarly $\mathcal{H}_{\mathcal{Y}} \subset L^2_{\mathcal{Y}}(P_Y)$ and $\mathcal{H}_{\mathcal{Z}} \subset L^2_{\mathcal{Z}}(P_Z)$ [Steinwart and Christmann, 2008, Section 4.3]. We will also assume that the kernels k , l and m are characteristic, such that their RKHSs are dense in the corresponding L^2 -spaces [Sriperumbudur et al., 2010].

5.2.1 Weak Conditional Independence

Conditional independence $X \perp\!\!\!\perp Y | Z$ is equivalent to

$$\mathbb{P}(X \in A | Z) \mathbb{P}(Y \in B | Z) = \mathbb{P}(X \in A, Y \in B | Z) \quad (5.1)$$

as random variables, for all measurable sets $A \in \mathcal{B}_{\mathcal{X}}$ and $B \in \mathcal{B}_{\mathcal{Y}}$, which in turn can be written as

$$\begin{aligned} & \mathbb{E}_Z [\mathbf{1}\{Z \in C\} \mathbb{P}(X \in A | Z) \mathbb{P}(Y \in B | Z)] \\ &= \mathbb{E}_Z [\mathbf{1}\{Z \in C\} \mathbb{P}(X \in A, Y \in B | Z)] \\ &= \mathbb{P}(X \in A, Y \in B, Z \in C) \end{aligned}$$

for all measurable sets $A \in \mathcal{B}_{\mathcal{X}}$, $B \in \mathcal{B}_{\mathcal{Y}}$ and $C \in \mathcal{B}_{\mathcal{Z}}$. Relaxation of this property which only requires it to hold for $C = \mathcal{Z}$ gives rise to the notion of *weak conditional independence*, i.e. that $\forall A \in \mathcal{B}_{\mathcal{X}}, B \in \mathcal{B}_{\mathcal{Y}}$

$$\mathbb{E}_Z [\mathbb{P}(X \in A | Z) \mathbb{P}(Y \in B | Z)] = \mathbb{P}(X \in A, Y \in B), \quad (5.2)$$

which we can write as $\mathbb{E}_Z [P_{X|Z} \otimes P_{Y|Z}] = P_{XY}$, understood as the equality of probability measures defined on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_{\mathcal{X}} \times \mathcal{B}_{\mathcal{Y}})$. Weak conditional independence is studied by Daudin [1980], who interprets it in terms of zero expected conditional covariances of square integrable functions, i.e.

$$\mathbb{E}_Z [\text{Cov}[f(X), g(Y) | Z]] = 0 \quad (5.3)$$

for all $f \in L^2_{\mathcal{X}}(P_X)$ and $g \in L^2_{\mathcal{Y}}(P_Y)$. In other words, the residuals in all square integrable functions of X and Y , given Z , are uncorrelated. Fukumizu et al. [2004a] give another characterisation of weak conditional independence using conditional cross-covariance operators between RKHSs. Namely, they define the operator $\Sigma_{YX|Z}: \mathcal{H}_X \rightarrow \mathcal{H}_Y$, such that $\forall f \in \mathcal{H}_X, g \in \mathcal{H}_Y$

$$\langle g, \Sigma_{YX|Z} f \rangle_{\mathcal{H}_Y} = \mathbb{E}_Z [\text{Cov}[f(X), g(Y)|Z]], \quad (5.4)$$

under the additional smoothness assumptions on conditioning variables, i.e. that $\forall f \in \mathcal{H}_X, g \in \mathcal{H}_Y$, $\mathbb{E}[f(X)|Z = \cdot]$ and $\mathbb{E}[g(Y)|Z = \cdot]$ both belong to $\mathcal{H}_Z$ [Fukumizu et al., 2004a, Proposition 4], [Alpay, 2001]. Theorem 8 of Fukumizu et al. [2004a] then shows that, for characteristic kernels, this operator vanishes if and only if weak conditional independence holds. i.e.

$$\Sigma_{YX|Z} = 0 \iff \mathbb{E}[P_{X|Z} \otimes P_{Y|Z}] = P_{XY}. \quad (5.5)$$

The conditional independence clearly implies weak conditional independence but the converse is not true. A simple counterexample is where X, Y and Z are all pairwise independent, but are jointly dependent due to a three-variable interaction. For a concrete case, similar to the one described in Sejdinovic et al. [2013a], let $X, Y, W \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ and define $Z = \text{sign}(XY)|W|$. In this case, $\mathbb{E}(f(X)|Z) = \mathbb{E}(f(X))$ and $\mathbb{E}(g(Y)|Z) = \mathbb{E}(g(Y))$ implying that LHS in (5.3) vanishes. However, given Z , the pair (X, Y) can only take values in two out of four quadrants, so that X and Y are clearly conditionally dependent given Z . Essentially, whenever the variables are pairwise independent but jointly dependent, any two variables are weakly conditionally independent given the third and need not be “strongly” conditionally independent. We note however that nonparametric tests exist which are directly aimed at such multivariate interaction: Sejdinovic et al. [2013a] reports stronger power than the conditional independence tests in these cases. Thus, we challenge the paradigm by which all departures from conditional independence should be tested with the same procedure and instead choose to focus on weak conditional independence testing. When multivariate interaction is present and the individual effects of Z on X and on Y are weak or non-existing and only the joint dependence of (X, Y) on Z is possible to detect, then arguably (X, Y) should be considered a single random vector.

5.2.2 RESIT: The Two-Step Approach

RESIT approach [Hoyer et al., 2009, Peters et al., 2014, Flaxman et al., 2015] converts a conditional independence testing problem into an unconditional one by removing the

effect of a confounder Z through a flexible nonparametric regression. More specifically, it assumes an additive noise model where X and Y are expressed as some deterministic functions of Z plus an additive zero-mean noise term, i.e.

$$X = f(Z) + n_x \quad (5.6)$$

$$Y = g(Z) + n_y \quad (5.7)$$

where n_x and n_y are zero-mean random variables independent of Z . Under the assumptions (5.6) and (5.7), the conditional independence can be characterised as:

$$X \perp\!\!\!\perp Y | Z \iff n_x \perp\!\!\!\perp n_y. \quad (5.8)$$

Thus, the test proceeds by first regressing X on Z and Y on Z , and then testing for the (unconditional) independence between the fitted residuals of these regressions $\hat{\epsilon}_x = X - \hat{\mathbb{E}}[X|Z]$ and $\hat{\epsilon}_y = Y - \hat{\mathbb{E}}[Y|Z]$. This approach obviously crucially depends on the regression procedure in order to remove the effect of Z by using an appropriate model of regression functions f and g . Gaussian Process regression and kernel ridge regression are often used for the first step, whereas Hilbert-Schmidt Independence Criterion [Gretton et al., 2008] can be used in the second step in order to capture potentially *nonlinear dependence* between residuals. Because of its focus on individual effects of the conditioning variable Z on X and on Y , it is clear that RESIT only tests for weak conditional independence.

As we will demonstrate below, RESIT can result in a substantial inflation of false discoveries when assumptions (5.6) and (5.7) are violated. Indeed, RESIT is unable to handle non-functional dependence between the variables X and Y and the conditioning variable Z – the regression will not remove the dependence on Z and the fitted residuals will be dependent even in the cases where X and Y are independent given Z . We see this as an undesirable property for many real data applications, especially when conditioning on a multivariate Z , as these assumptions are difficult to verify and may be violated. In particular, there could be hidden categorical confounders which introduce a switching behaviour in the way X and Y depend on Z , i.e. X and Y are functionally dependent on Z and those confounders, but not on Z itself. An example of such a relationship is illustrated in Figure 5.2 which plots the total expenditure on health per capita vs the gross national income per capita for 178 countries: there is a superposition of two functional relationships corresponding to two subsets of countries depending on whether their economy relies on oil [WHO,

2009].

For a simple example of how brittle the RESIT approach is, consider the dataset presented in Figure 5.1. Here, the dependence of both X and Y on Z is a mixture of two functional relationships (linear for X , quadratic for Y), arguably very simple – but also indicating the presence of some latent switching mechanism for both X and Y . However, the conditional expectations in both cases are constant and thus independent of Z . Hence, the fitted regression functions do not capture any dependence on Z and as a result, the residuals (rightmost plot) are clearly dependent, regardless of whether in fact $X \perp\!\!\!\perp Y|Z$ (indeed, in this case, as described in Section 5.4.2, the conditional independence does hold because switching variables are independent). RESIT with a nonlinear independence test on these residuals will therefore falsely reject the null hypothesis. This is a highly undesirable property as it leads to an inflated number of false positives, and when the same method is used inside PC algorithm, it can potentially result in reporting spurious causal links as illustrated in Section 5.4.3. This problem is exacerbated for Z being a random vector, even of a small dimension, since such cases which go beyond simple functional relationships become difficult to notice. One way to avoid the inflation of false positives in these cases is to only test for *linear dependence* between the residuals. However, this clearly results in a reduction of power as it misses many alternative models where (5.6) and (5.7) are satisfied and n_x and n_y are nonlinearly dependent.

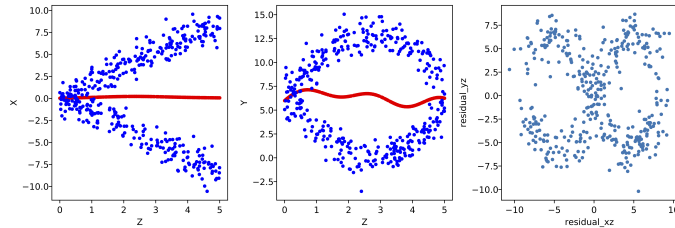


Figure 5.1: Data generated from the null model of Section 5.4.2 and kernel ridge regressions are used to remove dependency on Z . Left: X against Z . Middle: Y against Z . Simulated data (blue) and fitted values (red). Right: residuals of the regression for Y against residuals of the regression for X which clearly exhibits non-linear dependency.

5.2.3 Kernel Tests for “Strong” Conditional Independence

In contrast to the RESIT test, there have been several approaches in the literature that aim to measure all forms of conditional dependence, typically at the expense of

more complex statistics which have the distribution under the null hypothesis that is more difficult to estimate.

Conditional Cross-Covariance Operators Fukumizu et al. [2008] proposed a general nonparametric characterisation of conditional independence based on the conditional cross-covariance operators between RKHSs. The cross-covariance operator $\Sigma_{YX} : \mathcal{H}_X \rightarrow \mathcal{H}_Y$ is defined through

$$\langle g, \Sigma_{YX} f \rangle = \text{Cov}(f(X), g(Y)), \quad f \in \mathcal{H}_X, \quad g \in \mathcal{H}_Y,$$

i.e., it is a nonlinear extension of the cross-covariance matrix (cf. Baker [1973], Fukumizu et al. [2004a]). The conditional cross covariance operator $\Sigma_{YX|Z}$ is then defined as

$$\Sigma_{YX|Z} = \Sigma_{YX} - \Sigma_{YZ} \Sigma_{ZZ}^{-1} \Sigma_{ZX} \quad (5.9)$$

in analogy to the conditional cross-covariance matrix $C_{YX|Z} = C_{YX} - C_{YZ} C_{ZZ}^{-1} C_{ZX}$ formula for jointly Gaussian random vectors.

The conditional dependence measure can then be based on estimating $\|\Sigma_{\ddot{Y}\ddot{X}|Z}\|_{HS}^2$, where $\ddot{Y} = (Y, Z)$ and $\ddot{X} = (X, Z)$, as it can be shown [Fukumizu et al., 2008] that $\Sigma_{\ddot{Y}\ddot{X}|Z} = 0$ if and only if $X \perp\!\!\!\perp Y|Z$ when the product of the three positive definite kernels, klm , results in a characteristic kernel on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. In addition, Fukumizu et al. [2008] considers the conditional cross-correlation operators, which can be expressed as $V_{YX|Z} = \Sigma_{Y\ddot{Y}}^{-1} \Sigma_{YX|Z} \Sigma_{\ddot{X}X}^{-1}$.

The test statistics with these approaches however do not have clear asymptotic null distributions. Consequently, one needs to adopt a *local permutation-based approach* which needs to take into account an artificial discretisation of condition Z (e.g. through clustering) so that the marginal structures are preserved. As remarked by Zhang et al. [2011], this usually requires large sample size and the results become unreliable when the dimension of Z increases.

Kernel Conditional Independence Test (KCI-Test²) Following similar reasoning, Zhang et al. [2011] characterise conditional independence based on partial association [Daudin, 1980] where it was shown that

$$X \perp\!\!\!\perp Y|Z \iff \mathbb{E}(\tilde{f}\tilde{g}) = 0 \quad (5.10)$$

²Termed by Zhang et al. [2011]

for all $\tilde{f} \in \mathcal{E}_{XZ} := \{\tilde{f} \in L^2_{\mathcal{X} \times \mathcal{Z}}(P_{XZ}) | \mathbb{E}(\tilde{f}|Z) = 0\}$ and $\tilde{f}(\ddot{X}) = f(\ddot{X}) - \mathbb{E}(f|Z)$ for $f \in L^2_{\mathcal{X} \times \mathcal{Z}}(P_{XZ})$. The notations for $\tilde{g}$, g and $\mathcal{E}_{YZ}$ are defined similarly. As noted, if the functions f and g are restricted to the spaces $\mathcal{H}_{\mathcal{X} \times \mathcal{Z}}$ and $\mathcal{H}_{\mathcal{Y} \times \mathcal{Z}}$, then such characterisation is exactly the same as Fukumizu et al. [2008]. Zhang et al. [2011] derive an explicit formulae for the asymptotic null distribution and propose an approach for approximating it based on eigen-decompositions of kernel matrices or fitting a parametric family. However, this null distribution becomes harder to accurately approximate in practice as the dimension of the conditioning variable increases.

Permutation-based Maximum Mean Discrepancy (MMD) Following the difficulties faced by the approaches of Fukumizu et al. [2008] and Zhang et al. [2011] in estimating the null distributions, Doran et al. [2014] introduce a reduction of the conditional independence testing problem into a two-sample testing one, by considering a carefully selected permutation which mimics a sample from $P_Z P_{X|Z} P_{Y|Z}$ for which the null hypothesis of conditional independence holds. While casting the problem as a simple two-sample test is attractive, promising a better calibration and improved performance in the cases where dimensionality of the conditioning variable is high, this approach does require a costly optimization procedure over the space of doubly stochastic matrices in order to select the required permutation.

5.3 Proposed Method

We have seen that RESIT, although a simple and effective conditional independence test for additive noise models, can lead to undesirable properties when the additive noise modelling assumption is violated. On the other hand, strong kernel-based independence tests require no restrictive assumptions on the type of the underlying relationships between the variables, but need complex and difficult to tune testing procedures, often requiring to solve difficult side-problems in order to have an estimate of the distribution under the null hypothesis (clustering the conditioning variable or optimisation over the space of permutations). Therefore, our goal is to identify an approach which aims to strike a balance between these two approaches, and construct a test which is more robust to the departures from the modelling assumption in (5.6) and (5.7). At the same time, we seek an approach which can be cast as an unconditional independence test on some form of regression residuals – allowing us to employ the well established existing methodology for unconditional tests, including straightforward permutation-based approaches for estimation of the null distribution. Using

heuristic arguments, we derive in this section an intuitive test based on well known empirical mean RKHS quantities. The rigorous statistical analysis of this approach is left to further work.

We first note that the assumptions (5.6) and (5.7) specify that the conditional distributions $P_{X|Z=z}$ and $P_{Y|Z=z}$ depend on z only through the expectations $\mathbb{E}[X|Z=z]$ and $\mathbb{E}[Y|Z=z]$. Therefore, any additional dependence on z , e.g. in the conditional variance: $\text{Var}[X|Z=z] \neq \text{Var}[X]$ will not be captured by the regression approaches, implying that the residuals remain dependent by construction, leading to spurious rejections when nonlinear dependence tests on residuals are employed. Therefore, we are interested in also modelling how higher-order moments of these distributions depend on z , leading us to consider *conditional expectations of feature maps* $\mathbb{E}[\phi(X)|Z=z]$ and $\mathbb{E}[\psi(Y)|Z=z]$ for some $\phi: \mathcal{X} \rightarrow \mathcal{H}_X$, $\psi: \mathcal{Y} \rightarrow \mathcal{H}_Y$ and Hilbert spaces $\mathcal{H}_X$, $\mathcal{H}_Y$, i.e. the mean embeddings of the corresponding conditional distributions. For characteristic kernels [Sriperumbudur et al., 2010], these embeddings fully characterise the corresponding distributions (and thus, their dependence on z).

We thus propose the following two-step approach: first construct feature representations of X and Y and then perform the *vector-valued regression* of these feature representations of each of them on Z separately. Just like in RESIT, the second step is a test of independence between the resulting residuals, but note that the residuals are themselves elements of the corresponding (potentially infinite-dimensional) feature spaces. Fortunately, as we show in the remainder of this section, using the kernel trick allows to perform each of these steps without ever explicitly computing the feature maps. The final result is a simple test statistic, very similar to HSIC of Gretton et al. [2008], which lends itself to direct computation. Moreover, its null distribution can be straightforwardly estimated using a permutation-based approach. Thus, the key step of the procedure is the feature transformation of the responses in regression – we will see that this greatly relaxes the modelling assumptions and in particular allows to model complex dependencies between the responses and the conditioning variable which cannot be expressed in functional forms with additive noise.

Consider now k , l and m to be measurable positive definite kernels with the corresponding RKHSs $\mathcal{H}_X$, $\mathcal{H}_Y$, $\mathcal{H}_Z$. We denote by $\phi(x) \in \mathcal{H}_X$ and $\psi(y) \in \mathcal{H}_Y$ the feature map representations of k and l . Note that Daudin’s condition (5.3) for weak

conditional independence can be written as

$$\begin{aligned}
0 &= \mathbb{E}_Z [\text{Cov}[f(X), g(Y)|Z]] \\
&= \mathbb{E}_{XYZ} [(f(X) - \mathbb{E}f(X|Z))(g(Y) - \mathbb{E}g(Y|Z))] \\
&= \mathbb{E}_{XYZ} [n_f(X, Z)n_g(Y, Z)]
\end{aligned} \tag{5.11}$$

where we denoted $n_f(X, Z) = f(X) - \mathbb{E}f(X|Z)$, $n_g(Y, Z) = g(Y) - \mathbb{E}g(Y|Z)$ to be the regression residuals for X and Y transformed through square integrable functions $f \in L^2_{\mathcal{X}}(P_X)$ and $g \in L^2_{\mathcal{Y}}(P_Y)$. Note that these residuals would in general depend on Z and that they have mean zero by construction. Thus, weak conditional independence is equivalent to *uncorrelatedness* of residuals $n_f(X, Z)$ and $n_g(Y, Z)$ for all $f \in L^2_{\mathcal{X}}(P_X)$ and $g \in L^2_{\mathcal{Y}}(P_Y)$. Moreover, when $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ are dense in $L^2_{\mathcal{X}}(P_X)$ and $L^2_{\mathcal{Y}}(P_Y)$, it suffices to consider $f \in \mathcal{H}_{\mathcal{X}}$ and $g \in \mathcal{H}_{\mathcal{Y}}$. But then

$$\begin{aligned}
&\mathbb{E}_{XYZ} [(f(X) - \mathbb{E}f(X|Z))(g(Y) - \mathbb{E}g(Y|Z))] \\
&= \mathbb{E}_{XYZ} \langle f, \phi(X) - \mathbb{E}\phi(X|Z) \rangle_{\mathcal{H}_{\mathcal{X}}} \langle g, \psi(Y) - \mathbb{E}\psi(Y|Z) \rangle_{\mathcal{H}_{\mathcal{Y}}} \\
&= \langle f, \mathbb{E}_{XYZ} (\phi(X) - \mathbb{E}\phi(X|Z)) \otimes (\psi(Y) - \mathbb{E}\psi(Y|Z)) g \rangle_{\mathcal{H}_{\mathcal{X}}} \\
&= \langle f, \mathbb{E}_{XYZ} [n_{\phi}(X, Z) \otimes n_{\psi}(Y, Z)] g \rangle_{\mathcal{H}_{\mathcal{X}}}.
\end{aligned}$$

This suggests the method which regresses $\phi(X)$ and $\psi(Y)$ on Z using vector-valued kernel ridge regression and tests for unconditional linear independence of the residuals $n_{\phi}(X, Z)$ and $n_{\psi}(Y, Z)$ which are elements of RKHSs.

The resulting regression functions are now simply the empirical mean embeddings of the conditional distributions $P_{X|Z=z}$ and $P_{Y|Z=z}$ [Fukumizu et al., 2011, Theorem 1] and can be expressed as follows [Song et al., 2009, Theorem 5]:

$$\hat{\mathbb{E}}(\phi(X)|Z = z) = \sum_{i=1}^n \beta_i^{\mathbf{x}}(z) \phi(x_i) = \Phi_{\mathbf{x}}^{\top} \beta^{\mathbf{x}}(z) \tag{5.12}$$

$$\hat{\mathbb{E}}(\psi(Y)|Z = z) = \sum_{i=1}^n \beta_i^{\mathbf{y}}(z) \psi(y_i) = \Psi_{\mathbf{y}}^{\top} \beta^{\mathbf{y}}(z) \tag{5.13}$$

with $\beta^{\mathbf{x}}(z) = (M + \lambda^{\mathbf{x}} I_n)^{-1} (m(z, z_1), \dots, m(z, z_n))^{\top}$, M is the $n \times n$ matrix such that $M_{ij} = m(z_i, z_j)$, $\Phi_{\mathbf{x}} = (\phi(x_1), \dots, \phi(x_n))^{\top}$ and $\Psi_{\mathbf{y}} = (\psi(y_1), \dots, \psi(y_n))^{\top}$. The notation $\beta^{\mathbf{x}}$ emphasises the fact that the coefficient vector may differ for $\mathbf{x}$ and $\mathbf{y}$; this difference would only be due to different choice of the regularisation parameters $\lambda^{\mathbf{x}}$ and $\lambda^{\mathbf{y}}$. Note that the procedure can straightforwardly be amended to use different kernels $m^{\mathbf{x}}$ and $m^{\mathbf{y}}$ for the two responses, but we do not pursue this further for simplicity of

exposition. We have enclosed two parameter optimisation schemes in the Appendix (Section 5.6).

The residual of the kernel ridge regression of X on Z for the i -th data point is given by

$$\begin{aligned}\hat{\epsilon}_{x,i} &= \phi(x_i) - \hat{\mathbb{E}}(\phi(X)|Z = z_i) \\ &= \Phi_{\mathbf{x}}^{\top} \mathbf{e}_i - \Phi_{\mathbf{x}}^{\top} \beta^{\mathbf{x}}(z_i) \in \mathcal{H}_k\end{aligned}\quad (5.14)$$

so that the whole set of residuals $\hat{\epsilon}_x = (\epsilon_{x,1}, \dots, \epsilon_{x,n})^{\top}$ can be written as:

$$\hat{\epsilon}_x = (I_n - B_{\mathbf{x}})\Phi_{\mathbf{x}} = \left(\frac{1}{\lambda^{\mathbf{x}}}M + I_n\right)^{-1}\Phi_{\mathbf{x}} \quad (5.15)$$

with $B_{\mathbf{x}} = M(M + \lambda^{\mathbf{x}}I_n)^{-1}$ and $\lambda^{\mathbf{x}} > 0$, so that $I_n - B_{\mathbf{x}} = (\frac{1}{\lambda^{\mathbf{x}}}M + I_n)^{-1}$. Similarly the residual of the kernel ridge regression of Y on Z is given by

$$\hat{\epsilon}_y = \left(\frac{1}{\lambda^{\mathbf{y}}}M + I_n\right)^{-1}\Psi_{\mathbf{y}}. \quad (5.16)$$

The problem of testing weak conditional independence between X and Y given Z then translates into testing if $\hat{\epsilon}_x$ and $\hat{\epsilon}_y$ are correlated.

Since they are elements of RKHSs, we can use the Hilbert-Schmidt independence criterion (HSIC) of Gretton et al. [2008] with linear (inner product) kernels on these residuals, i.e. $\kappa_x(\epsilon_{x,i}, \epsilon_{x,j}) = \langle \epsilon_{x,i}, \epsilon_{x,j} \rangle_{\mathcal{H}_X}$ and $\kappa_y(\epsilon_{y,i}, \epsilon_{y,j}) = \langle \epsilon_{y,i}, \epsilon_{y,j} \rangle_{\mathcal{H}_Y}$. Let the empirically centred residuals be $\hat{\epsilon}_{x_i}^c := \hat{\epsilon}_{x_i} - \frac{1}{n} \sum_{j=1}^n \hat{\epsilon}_{x_j}$ and $\hat{\epsilon}_{y_i}^c := \hat{\epsilon}_{y_i} - \frac{1}{n} \sum_{j=1}^n \hat{\epsilon}_{y_j}$. Essentially, we build an empirical cross covariance operator between the residuals $\hat{\epsilon}_x$ and $\hat{\epsilon}_y$, and HSIC is then the squared Hilbert-Schmidt norm of such operator:

$$\Xi(\hat{\epsilon}_x, \hat{\epsilon}_y) = \left\| \frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_{x_i}^c \otimes \hat{\epsilon}_{y_i}^c \right\|_{HS}^2 \quad (5.17)$$

$$= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \langle \hat{\epsilon}_{x_i}^c, \hat{\epsilon}_{x_j}^c \rangle \langle \hat{\epsilon}_{y_i}^c, \hat{\epsilon}_{y_j}^c \rangle \quad (5.18)$$

$$= \frac{1}{n^2} \text{Trace}(H \hat{\epsilon}_x \hat{\epsilon}_x^{\top} H H \hat{\epsilon}_y \hat{\epsilon}_y^{\top} H) \quad (5.19)$$

where $H := I_n - \frac{1}{n} \mathbf{1}\mathbf{1}^{\top}$ is the centering matrix. By definition of $\hat{\epsilon}_x$, we have

$$\hat{\epsilon}_x \hat{\epsilon}_x^{\top} = \left(\frac{1}{\lambda^{\mathbf{x}}}M + I_n\right)^{-1} \Phi_{\mathbf{x}} \Phi_{\mathbf{x}}^{\top} \left(\frac{1}{\lambda^{\mathbf{x}}}M + I_n\right)^{-1} \quad (5.20)$$

and similarly

$$\hat{\epsilon}_y \hat{\epsilon}_y^\top = \left(\frac{1}{\lambda_y} M + I_n\right)^{-1} \Psi_y \Psi_y^\top \left(\frac{1}{\lambda_y} M + I_n\right)^{-1} \quad (5.21)$$

where $\Phi_x \Phi_x^\top = K$ and $\Psi_y \Psi_y^\top = L$ are the $n \times n$ matrices such as $K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i, y_j)$. Combining equations (5.19), (5.20) and (5.21), we obtain

$$\Xi(\hat{\epsilon}_x, \hat{\epsilon}_y) = \frac{1}{n^2} \text{Trace}(\tilde{H}_{z,x} K \tilde{H}_{z,x}^\top \tilde{H}_{z,y} L \tilde{H}_{z,y}^\top)$$

where $\tilde{H}_{z,x} = H(\frac{1}{\lambda_x} M + I_n)^{-1}$ and $\tilde{H}_{z,y} = H(\frac{1}{\lambda_y} M + I_n)^{-1}$. Note that the HSIC statistic between the two residuals can be seen as the inner product between transformations of the kernel matrices K and L through the centering terms $\tilde{H}_{z,x}$ and $\tilde{H}_{z,y}$ which encompass information regarding the impact of Z on respectively X and Y . To obtain the test threshold, standard permutation approach can be used to estimate the null distribution. It is tempting to also consider the nonlinear tests on RKHS residuals, i.e. one may specify some kernel which only depends on RKHS distances (e.g. Gaussian RBF kernels), however, this results in inflated false positive rates for the same reasons as RESIT.

5.3.1 Related Work

Our proposed conditional independence test which we term KRESIT³ (Kernel RESIT) is a generalisation of the RESIT approach in which the regressions on Z are done after feature transforming both responses X and Y . KRESIT is closely related to the method of Zhang et al. [2011]. More precisely, Zhang et al. [2011] can be understood as performing ridge regressions of the feature transformations of pairs (X, Z) and (Y, Z) on Z in order to achieve full characterisation of conditional independence. This makes the method more difficult to interpret in terms of two-step procedures. Recent extension by Strobl et al. [2018] of Zhang et al. [2011] presents the test called RCoT which essentially computes the same statistic as KRESIT, but uses a different estimation procedure of the asymptotic null distribution. RCoT can be viewed as a large scale approximation of KRESIT through the use of random Fourier features (RFF), similarly RCIT of Strobl et al. [2018] provides an approximation of the KCI-test via RFF. Strobl et al. [2018] has shown that RCIT and RCoT give similar performance as KCI-test but they are orders of magnitude faster in large-scale settings. While the large-scale approximations are not the focus of the present work,

³Code available at <https://github.com/oxmlcs/kerpy>.

methods such as RFF or Nyström approximation can readily be employed in KRESIT (for comparisons of these approximation methods in unconditional independence testing, cf. Zhang et al. [2017]).

5.4 Experiments

We apply the proposed method, KRESIT, to both synthetic and real data to evaluate its performance in terms of Type I error and statistical power. We start with a motivating real data example comparing KRESIT and RESIT on two socio-economic and health indicators from the World Health Organisation (WHO) data [WHO, 2009, Rosling, 2008]. Following this, we compare the performance of the proposed approach against RESIT and the three “strong” conditional independence tests described in Section 5.2.3 on a synthetic experiment with increasing dimensionality of the conditioning variable Z and sample sizes. We then apply the two “weak” conditional independence tests (RESIT and KRESIT) to unravel causal relationships in a synthetic dataset, the Boston Housing dataset [Pace and Gilley, 1997, Harrison Jr. and Rubinfeld, 1978] and the Ozone dataset [Breiman and Friedman, 1985] (presented in the Appendix Section 5.6.3).

Unless otherwise stated, the significance level is kept at $\alpha = 0.05$. For KRESIT, Gaussian kernels with median heuristic are used for all three random variables in regression, but the parameters λ^x and λ^y are optimised using grid search to find the minimum total 5-fold cross validation error over a grid of 30 evenly spaced values in the interval $(10^{-6}, 10^1)$. For RESIT, Gaussian kernels with median heuristic are used for kernel ridge regression and on the residuals. The regularisation parameters are tuned in the same way as in KRESIT.

5.4.1 Simple Motivating Example Based on WHO Data

As a motivating example, we apply the proposed approach to two variables in the WHO data set used in Reshef et al. [2011]. We denote by Z the log transformed gross national income per capita and by Y the total expenditure on health per capita. Figure 5.2 (left) can be thought of as the superposition of two relationships where the small minority curve consists of countries whose economies rely largely on oil [WHO, 2009]. We have taken the log transform of the gross national income per capita so that the values are more evenly distributed. After removing all missing values, we have

178 data points. We then construct a synthetic X with non-functional dependence on Z as follows

$$x_i = \begin{cases} (z_i - 10)^2 + n_i^x & \text{if } c_i^x = 1 \\ -(z_i - 10)^2 + 35 + n_i^x & \text{if } c_i^x = 0 \end{cases} \quad (5.22)$$

where $c_i^x \stackrel{i.i.d.}{\sim} \text{Bernoulli}(0.5)$ and $n_i^x \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$. X and Y are conditionally independent given Z by construction.

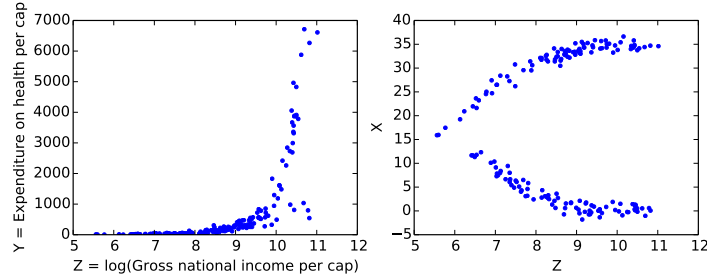


Figure 5.2: Left: Expenditure on health per cap against the logarithm of gross national income per cap (Z). Right: Synthetic data X against Z .

RESIT incorrectly rejected the null hypothesis at 5% significance level with a p-value of 0.0025 owing to the fact that kernel ridge regression is unable to remove the dependency of Z from X and Y and HSIC using Gaussian kernels is able to detect such nonlinear dependency between the residuals. KRESIT, on the other hand, gives a p-value of 0.91 and hence does not reject the null hypothesis that X and Y are weakly conditionally independent. By relaxing the modelling assumption of additive noise and allowing more complex dependencies between responses and the conditioning variable, KRESIT is able to provide a better calibrated test.

5.4.2 Effect of Dimensionality of Z and Sample Size

We examine the probability of Type I and Type II errors of the “strong” conditional independence tests vs. the “weak” conditional independence tests when the dimensionality of the conditioning set Z is increasing ($d = 1, 2, 3, 4, 5, 6, 7$) and the sample sizes take values in $\{40, 80, 120, 160, 200\}$.

Let us consider the following non-functional dependence as the null model:

$$\begin{aligned}
Z_i^j &\in \overset{i.i.d.}{\sim} \text{Uniform}(0, 5), \quad j = 1, \dots, d \\
C_i^x, C_i^y &\overset{i.i.d.}{\sim} \text{Bernoulli}(0.5) \\
n_i^x, n_i^y &\overset{i.i.d.}{\sim} \mathcal{N}(0, 1) \\
X_i \in \mathbb{R} &= \begin{cases} 1.7Z_i^1 + n_i^x & \text{if } c_i^x = 1 \\ -1.7Z_i^1 + n_i^x & \text{if } c_i^x = 0 \end{cases} \\
Y_i \in \mathbb{R} &= \begin{cases} (Z_i^1 - 2.7)^2 + n_i^y & \text{if } c_i^y = 1 \\ -(Z_i^1 - 2.7)^2 + 13 + n_i^y & \text{if } c_i^y = 0 \end{cases}
\end{aligned}$$

Note, X and Y only depend on the first dimension of the conditioning vector Z . To obtain the alternative model, we couple the latent variables C^x and C^y :

$$\begin{aligned}
u_i &\overset{i.i.d.}{\sim} \text{Uniform}(0, 1) \\
C_i^x &\overset{i.i.d.}{\sim} \text{Bernoulli}(0.5) \\
C_i^y &= \begin{cases} C_i^x & \text{if } u_i < 0.3 \\ \overset{i.i.d.}{\sim} \text{Bernoulli}(0.5) & \text{if } u_i \geq 0.3 \end{cases}
\end{aligned}$$

We compare the “weak” conditional independence tests: the proposed KRESIT, RESIT and LRESIT (RESIT with linear kernels on residuals) with the “strong” conditional independence tests: CIperm by Fukumizu et al. [2008], KCIPT by Doran et al. [2014] and KCI-test by Zhang et al. [2011]. The rejection rates are calculated out of 100 trials. Each rejection rate is associated with a 95% Wald confidence interval.

For a given dimension, we observe in Figure 5.3 that the Type I error of RESIT is increasing with the number of samples. As a simple kernel ridge regression cannot remove the dependency of Z from X and Y , the nonlinear dependency pattern in the residuals are more visible as sample size increases. As expected, the proposed approach, KRESIT, has correct Type I control together with LRESIT and the other “strong” conditional independence tests.

To establish sensible comparison in the alternative model (Figure 5.4), we only compare those methods with correct Type I control. When Z is of dimension 1, KRESIT performs similarly to KCI test where the probability of Type II error (1-power) for both decreases to zero as the number of samples increases to 200, while the other tests give weaker power. In dimension 7, KRESIT again gives strong performance,

outperforming other tests. Although LRESIT also tests for weak conditional independence, as expected, it has a reduced power comparing to KRESIT in both cases. While KCI test and KCIPT give similar power performance, CIperm struggles to detect the conditional dependence across all sample sizes.

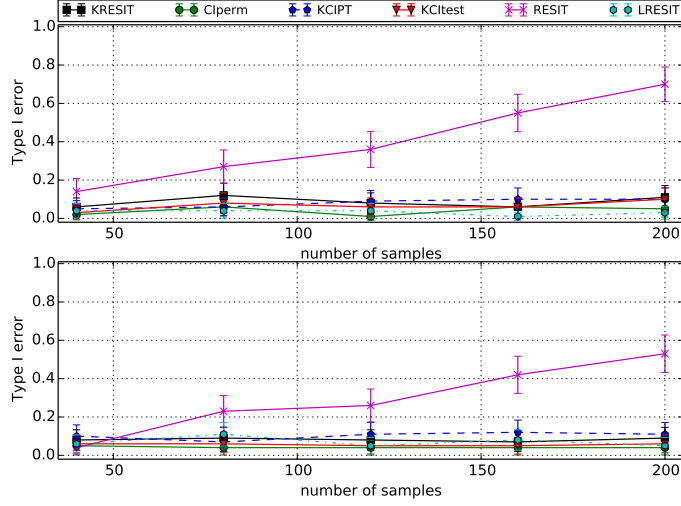


Figure 5.3: Null model experiments for $d = 1$ (Top) and $d = 7$ (Bottom).

5.4.3 Application in Causal Discovery

Conditional independence tests are frequently used in causal discovery to recover the dependence relationships among a set of variables represented as a directed acyclic graph (DAG). Assuming causal Markov condition (i.e., any variable/node is conditionally independent of its nondescendents given its parents [Hausman and Woodward, 1999]) and faithfulness (i.e., the conditional independences of the distribution can be inferred from the d-separation in the graph and vice-versa [Kalisch and Buhlmann, 2007]), the popular constraint-based PC-algorithm [Spirtes et al., 2000] is often used to estimate a Markov equivalence class of DAGs (i.e. a set of graphs that impose the same independences and conditional independences). The estimation starts from a complete undirected graph and recursively deletes edges based on conditional independence decisions. The result is an undirected graph (skeleton) which is then partially directed and further extended to represent the underlying equivalence class of DAG [Kalisch and Buhlmann, 2007]. The quality of the conditional independence tests are hence crucial for the performance of the PC algorithm.

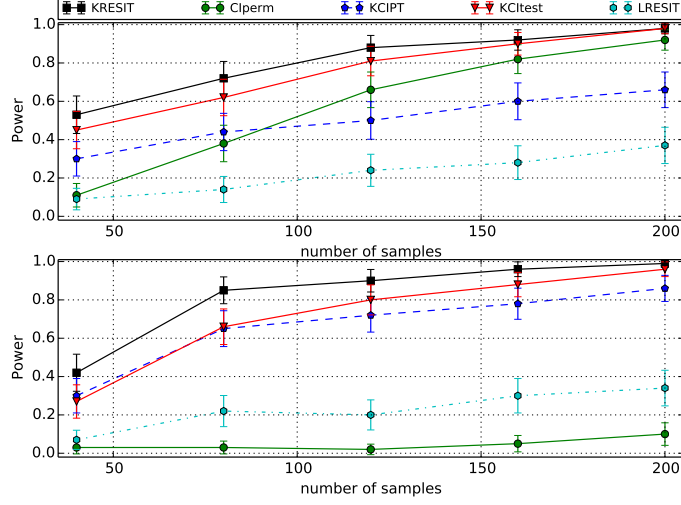


Figure 5.4: Alternative model experiments for $d = 1$ (Top) and $d = 7$ (Bottom).

The PC algorithm used in this section is a modified version of the pcalg implementation in python⁴ with conditional independence test using KRESIT/RESIT and independence test using HSIC with Gaussian kernel using median heuristic. Completed partially acyclic graph (CPDAG) are used to visualise the equivalence classes of DAGs. Whenever an undirected edge $i - j$ is shown, there exists a DAG with $i \rightarrow j$ and a DAG with $i \leftarrow j$ in the equivalence class.

5.4.3.1 Synthetic Data

Consider variables of the null model from Section 5.4.2 and additionally let $A_i = (Y_i - 5)^2/3 + 5 + n_i^A$ and $B_i = 5.5 \tanh(Y_i) + n_i^B$ where $n_i^A, n_i^B \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$. We compare the results obtained from the PC algorithm with the variables $\{X, Y, Z, C^x, C^y, A, B\}$ using KRESIT and RESIT. Using a sample of size 800, we obtain Figure 5.5 (Left). Both KRESIT and RESIT recover the correct structure of the model.

We now turn to investigating the robustness of the method upon removing the variables C^x and C^y . PC algorithm with KRESIT is still able to discover the correct skeleton as shown in Figure 5.5 (Right). An incorrect graph is obtained with RESIT, reporting a spurious causal link $X - Y$. We note that when the switching variables C^x, C^y are present, all associations are functional, but this is no longer the case when C^x, C^y are unobserved.

⁴<https://github.com/keiichishima/pcalg>

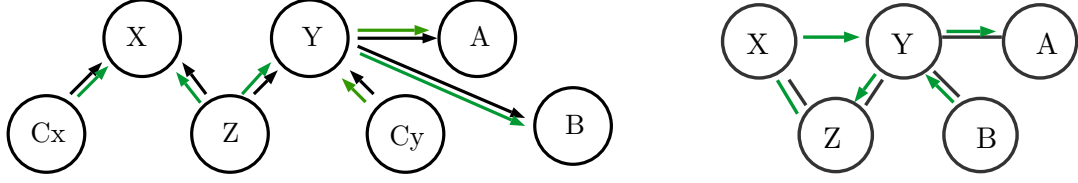


Figure 5.5: Left: The DAG obtained by PC algorithm using KRESIT and RESIT with C^x and C^y being latent variables. Right: The DAG obtained by using KRESIT (black) fully recovers the causal structure, but the one using RESIT (green) does not.

5.4.3.2 Boston Housing Data

We consider the corrected Boston Housing dataset of Pace and Gilley [1997] and pre-whiten each variable using the spatial coordinates with a GP regression as in Flaxman et al. [2015] since there is significant spatial clustering in every single variable in the dataset. There are 14 variables in total and each with 506 observations. For a detailed explanation of each variable, we refer to the original paper [Harrison Jr. and Rubinfeld, 1978]. To correct for multiple testing, we set the significance level to $\alpha = 0.001$.

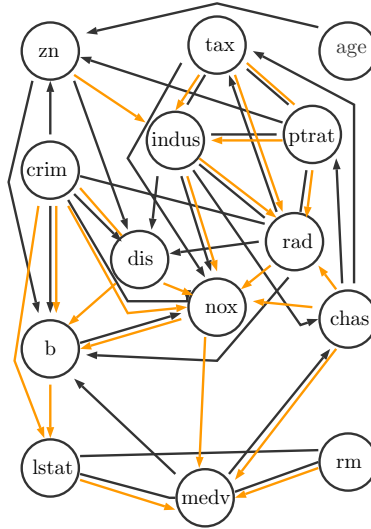


Figure 5.6: Results of PC algorithm with KRESIT (black) and RESIT [Flaxman et al., 2015] (orange) using pre-whitened Boston Housing dataset.

Overall, the CPDAG obtained by our proposed method is substantially different from previous results in the literature [Zhang et al., 2011, Flaxman et al., 2015]. We argue that many potentially unobserved variables could be driving such differences.

The details and some possible interpretations are given in Section 5.6.2. Therefore, the links discovered using RESIT and KRESIT within PC algorithm may need further investigation and one should analyse the results with potential hidden variables in mind and be cautious about any conclusion drawn due to potential violation of structural assumptions.

5.5 Conclusion

We have proposed a weak conditional independence test that extends the popular two-step approach RESIT, which combines regression with an unconditional independence test. We have considered RKHS-valued ridge regressions and subsequently used a test for linear independence on RKHS-valued residuals. While maintaining simple and effective testing procedures of RESIT, the resulting test has a correct Type I control under more challenging scenarios where the modelling assumptions required by RESIT are violated. It also yields competitive or improved power performance to that of the other conditional independence tests. When used in the PC algorithm, the proposed method is more robust than RESIT to hidden variables inducing associations more complex than functional ones with additive noise.

5.6 Appendix

5.6.1 Parameters Optimisation

For the proposed method where Gaussian radial basis function (RBF) kernels are used on all three random variables of interest, there are two groups of hyper-parameters. The first one includes the kernel bandwidths for the kernel matrices K , L and M . The bandwidth for K and L are fixed by using median heuristic. Although we also fixed the bandwidth for M by median heuristic in the main text, here we include a discussion on the optimisation of such parameter when one has a different kernel bandwidth for the regression on the feature space of X and on that of Y . Let's call them σ_{xz}^2 and σ_{yz}^2 respectively. The second group contains the regularisation parameters λ^x and λ^y for the kernel ridge regressions. These parameters can be optimised through grid search or gradient descent on the total cross-validation error as follows.

For a K -fold cross-validation, we aim to minimise the sum of the square of the residuals of each unseen fold given the model trained on the other $K - 1$ folds over

some (λ, σ^2) values. Here, we present detailed derivation for X and that of Y follows similarly. For a pair of fixed $(\lambda^x, \sigma_{xz}^2)$, the cross-validation error of the k^{th} fold is:

$$CVerror_\lambda^k = \frac{1}{u} \sum_{i \in tst} \|\phi(x_i^{tst}) - \Phi_{tr}^T \beta^x(z_i^{tst})\|^2 \quad (5.23)$$

where $\{x_i^{tr}, z_i^{tr}\}_i$ and $\{x_i^{tst}, z_i^{tst}\}_i$ denote respectively the training and testing sets for the k^{th} fold, $\Phi_{tr} = (\phi(x_1^{tr}), \phi(x_2^{tr}), \dots)^\top$ and u is the size of the testing set. Then, the overall cross-validation error is

$$CVerror_\lambda = \frac{1}{K} \sum_{k=1}^K CVerror_\lambda^k. \quad (5.24)$$

To simplify (5.23), we note that

$$\sum_{i \in tst} \|\phi(x_i^{tst}) - \Phi_{tr}^T \beta^x(z_i^{tst})\|^2 \quad (5.25)$$

$$= \sum_{i \in tst} \|\phi(x_i^{tst}) - \Phi_{tr}^T (M_{tr, tr} + \lambda^x I)^{-1} (m(z_i^{tst}, z_1^{tr}), m(z_i^{tst}, z_2^{tr}), \dots, m(z_i^{tst}, z_m^{tr}))^T\|^2 \quad (5.26)$$

$$= Tr([\Phi_{tst} - \mathcal{M}_z][\Phi_{tst} - \mathcal{M}_z]^T) \quad (5.27)$$

$$= Tr(\Phi_{tst} \Phi_{tst}^T - \Phi_{tst} \mathcal{M}_z^T - \mathcal{M}_z \Phi_{tst}^T + \mathcal{M}_z \mathcal{M}_z^T) \quad (5.28)$$

where $\mathcal{M}_z = M_{tst, tr} (M_{tr, tr} + \lambda^x I)^{-1} \Phi_{tr}$. Such prediction error for the k^{th} fold (i.e. (5.28)) can be explicitly formulated in terms of $(\lambda^x, \sigma_{xz}^2)$ by writing the gram matrices in terms of square distance matrices:

$$M_{*, \cdot} = \exp(-0.5 D_{*, \cdot} / \sigma_{xz}^2) \quad (5.29)$$

where $D_{*, \cdot}$ is the square distance matrix computed on the respective sets, i.e. $(D_{*, \cdot})_{ij} = \|z_i - z_j\|_2^2$. Note, no explicit feature map representation is required.

Since both parameters are constrained to be positive, we can rewrite $\lambda^x = \exp(\eta)$ and $\sigma_{xz}^2 = \exp(\gamma)$ and take first order derivate with respect to η and γ to obtain a gradient descent algorithm. Given the expression in (5.24), taking the first order derivative with respect to η (Appendix D of Dattorro [2016]):

$$\begin{aligned} \frac{d}{d\eta} Error_{(\lambda^x, \sigma_{xz}^2)} &= 2 Tr\{K_{tst, tr} \mathcal{M}_\eta M_{tr, tst} \\ &\quad - M_{tst, tr} \mathcal{M}_\eta K_{tr, tr} \mathcal{M}_{inv} M_{tr, tst}\} \end{aligned}$$

where $\mathcal{M}_\eta := (M_{tr,tr} + \lambda^x I)^{-1} [\exp(\eta) I] (M_{tr,tr} + \lambda^x I)^{-1}$ and $\mathcal{M}_{inv} := (M_{tr,tr} + \lambda^x I)^{-1}$. Similarly, take the first order derivative with respect to γ :

$$\begin{aligned}
& \frac{d}{d\gamma} Error_{(\lambda^x, \sigma_{xz}^2)} \\
&= Tr \{ 2 K_{tst,tr} \mathcal{M}_{inv} \mathcal{D}_{tr,tr} \mathcal{M}_{inv} M_{tr,tst} \\
&\quad - 2 K_{tst,tr} \mathcal{M}_{inv} \mathcal{D}_{tr,tst} \\
&\quad + \mathcal{D}_{tst,tr} \mathcal{M}_{inv} K_{tr,tr} \mathcal{M}_{inv} M_{tr,tst} \\
&\quad - M_{tst,tr} \mathcal{M}_{inv} \mathcal{D}_{tr,tr} \mathcal{M}_{inv} K_{tr,tr} \mathcal{M}_{inv} M_{tr,tst} \\
&\quad - M_{tst,tr} \mathcal{M}_{inv} K_{tr,tr} \mathcal{M}_{inv} \mathcal{D}_{tr,tr} \mathcal{M}_{inv} M_{tr,tst} \\
&\quad + M_{tst,tr} \mathcal{M}_{inv} K_{tr,tr} \mathcal{M}_{inv} \mathcal{D}_{tr,tst} \} \tag{5.30}
\end{aligned}$$

where

$$\mathcal{D}_{tr,\cdot} := [0.5 D_{tr,\cdot} \times \exp(-0.5 D_{tr,\cdot} \exp(-\gamma)) \exp(-\gamma) I].$$

At iteration i , we update η and γ according to

$$\eta_i = \eta_{i-1} - \alpha \frac{d}{d\eta} Error_{(\lambda^x, \sigma_{xz}^2)} \Big|_{\eta_{i-1}, \gamma_{i-1}} \tag{5.31}$$

$$\gamma_i = \gamma_{i-1} - \alpha \frac{d}{d\gamma} Error_{(\lambda^x, \sigma_{xz}^2)} \Big|_{\eta_{i-1}, \gamma_{i-1}} \tag{5.32}$$

with learning rate α until convergence. Various adaptive learning rate schemes can be used for faster convergence.

5.6.2 Boston Housing Data

The main variable of interest is the median value of occupied homes (medv). Based on the result using KRESIT (Figure 5.6), the proportion of lower status people (lstat) and (medv) are conditionally dependent given the rest of the variables, so is average number of rooms per unit (rm) and (medv). However, the estimated CPDAG cannot be more specific on the direction of the causation. Additionally, (medv) has been identified to cause the percentage of black population (b). We observe that such link is indeed plausible. In the US around the time of the 1970 census, residential segregation was common place, and neighbourhoods with a higher percentage of blacks were regarded as less desirable to live in. Banks and the federal government itself refused to give loans to blacks in general, and specifically refused to give them loans

when they wanted to buy homes in neighbourhoods that had historically been white.

The variable (medv) has also been identified to cause whether the house is next to the Charles River (chas). However, this edge seems to be misdirected. The result of Flaxman et al. [2015] (using RESIT) has shown the opposite (chas) $\rightarrow$ (medv) which agrees with authors' belief in the original paper [Harrison Jr. and Rubinfeld, 1978]. Besides this, they also showed that (nox), (lstat) and (rm) are the causes for (medv). While the strong independence test of Zhang et al. [2011] also identified (lstat) and (rm) as the causes, it additionally concludes crime rates (crim) and proportion of unit built before 1940 (age) as causes. The strong CI test of Fukumizu et al. [2008] agrees with Zhang et al. [2011] on (lstat) and (age).

5.6.3 Ozone Data

We consider the ozone dataset used in Breiman and Friedman [1985] and pre-whitened each variable using time with a GP regression [Flaxman et al., 2015] as the dataset consists of 330 daily observations with clear temporal autocorrelation. We refer to [Breiman and Friedman, 1985] for the details of each variable and the main variable of interest is the ozone concentration (Ozone). We set the significance level to be $\alpha = 0.05$ as the number of variables is only 9.

Figure 5.7 illustrates that though some edges are present in both our result and the one obtained by Flaxman et al. [2015], the two graphs are quite different. The CPDAG obtained using KRESIT shows that there is no cause of (Ozone) among the variables under consideration, but (Ozone) causes temperature (Temp), humidity (Hum) and inversion base height (InvHt). Flaxman et al. [2015] also concludes that there is no cause of (Ozone) among the variables under consideration, but (Ozone) is one of the causes for visibility (Vis).

The correlations between the concentration of ozone, temperature, inversion base height and humidity in the atmosphere have been shown to exist in many research on meteorology, public health and environmental science. However, the causal directions are less clear as the interaction between them are complex and there are additional factors that plays a part which have not been considered in this dataset. For example, the ratio of volatile organic compounds (VOC) and nitrogen oxides (Nox), the concentration of other green house gases, if heavy cloud is present, etc. We regard our method as promising as it discovers the links between these variables.

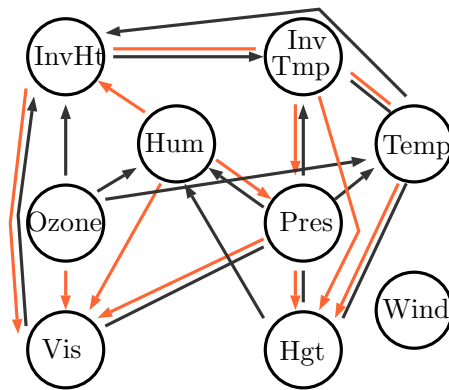


Figure 5.7: Results of PC algorithm with KRESIT (black) and RESIT [Flaxman et al., 2015] (orange) using pre-whitened Ozone dataset.

Chapter 6

Bayesian Two-Sample Testing

6.1 Introduction

Nonparametric two-sample testing is an important branch of hypothesis testing with a wide range of applications. The two-sample testing problem can be described as follows: given a set of samples $\{x_i\}_{i=1}^{n_x} \stackrel{i.i.d.}{\sim} P$ and $\{y_i\}_{i=1}^{n_y} \stackrel{i.i.d.}{\sim} Q$, we wish to evaluate the evidence for the competing hypotheses

$$H_0 : P = Q \text{ v.s. } H_1 : P \neq Q \quad (6.1)$$

with the probability distributions P and Q unknown. In this chapter, we will pursue a Bayesian perspective to this problem. In this perspective, hypotheses can be formulated as models and hypothesis testing can therefore be viewed as a form of model selection, i.e. to identify which model is strongly supported by the data [Jefferys and Berger, 1992].

The classical Bayesian formulation of the two-sample testing problem is in terms of the Bayes factor [Jeffreys, 1935, 1961, Kass and Raftery, 1995]. For a given data set $\mathcal{D}$ and two competing models/hypotheses M_0 (or H_0) and M_1 (or H_1) parameterised by the vectors θ_0 and θ_1 respectively, the Bayes factor is represented as the likelihood ratio of the samples given that they were generated from the same distribution (null hypothesis) to that they were generated from different distributions (alternative hypothesis):

$$BF = \frac{P(\mathcal{D}|M_0)}{P(\mathcal{D}|M_1)} \quad (6.2)$$

$$= \frac{\int P(\theta_0|M_0)P(\mathcal{D}|\theta_0, M_0)d\theta_0}{\int P(\theta_1|M_1)P(\mathcal{D}|\theta_1, M_1)d\theta_1} \quad (6.3)$$

Note, in our nonparametric two-sample testing regime, θ_0 and θ_1 are the hyperparameters. The Bayes factor can be interpreted as the posterior odds on the null distribution when the prior probability on the null distribution is $\frac{1}{2}$ [Kass and Raftery, 1995]. For the two-sample testing problem, the data set under consideration is $\mathcal{D} = \{\{x_i\}_{i=1}^{n_x}, \{y_j\}_{j=1}^{n_y}\}$. Assuming X and Y are independent, under the null hypothesis H_0 , the two samples come from the same distribution parameterised by θ_0 (i.e. a distribution for the pooled samples) which we denote as $P_{x,y} = P = Q$. This implies

$$P(\mathcal{D}|\theta_0, H_0) = \prod_{i=1}^{n_x} P_{x,y}(x_i|\theta_0) \prod_{j=1}^{n_y} P_{x,y}(y_j|\theta_0). \quad (6.4)$$

Under the alternative hypothesis H_1 , the two samples come from different distributions P and Q parameterised respectively by the vectors $\theta_{1,1}$ and $\theta_{1,2}$. These parameter vectors are collectively defined as $\theta_1 := (\theta_{1,1}^\top, \theta_{1,2}^\top)^\top$. This implies

$$P(\mathcal{D}|\theta_1, H_1) = \prod_{i=1}^{n_x} P(x_i|\theta_{1,1}) \prod_{j=1}^{n_y} Q(y_j|\theta_{1,2}). \quad (6.5)$$

If the posterior probability of the model given the data is of interest, it can be easily written in terms of the Bayes factor:

$$P(M_0|\mathcal{D}) = \frac{BF}{1 + BF} \quad (6.6)$$

$$P(M_1|\mathcal{D}) = 1 - P(M_0|\mathcal{D}) = \frac{1}{1 + BF} \quad (6.7)$$

When the prior probabilities on the models are not equal, the posterior probability of the null model can be written as:

$$P(M_0|\mathcal{D}) = \frac{P(M_0) \int P(\theta_0|M_0)P(\mathcal{D}|\theta_0, M_0)d\theta_0}{P(M_0) \int P(\theta_0|M_0)P(\mathcal{D}|\theta_0, M_0)d\theta_0 + P(M_1) \int P(\theta_1|M_1)P(\mathcal{D}|\theta_1, M_1)d\theta_1} \quad (6.8)$$

$$= \frac{BF}{BF + \frac{P(M_1)}{P(M_0)}}. \quad (6.9)$$

As discussed by Borgwardt and Ghahramani [2009], the crucial question is to define the hypotheses/likelihood models such that the resultant Bayesian two-sample test is general and applicable to all data. Ideally, we would like the model to cover as large a class of probability distributions as possible and is also applicable to multivariate distributions and other domains. Bayesian parametric hypothesis testing, i.e. when the probability distributions P and Q are of known form, is well developed

and we refer the readers to Bernardo and Smith [2000]. Most nonparametric work has been focusing on testing a parametric model versus a nonparametric one and a detailed summary has been provided by [Holmes et al., 2015]. More recently, Holmes et al. [2015] proposed to use a Pólya tree prior centered on some distribution G and model the distributions of the pooled samples $P_{x,y}$ ¹ under the null and the individual distributions P and Q under the alternative as i.i.d. draws from such Pólya tree prior. The extension to independence testing was done by Filippi and Holmes [2016]. While this approach indeed gives a flexible model of the underlying probability distributions, it can only be used for 1-dimensional data.

Borgwardt and Ghahramani [2009] proposed two Bayesian two-sample tests based on Bayes factor: a parametric test using exponential families to describe the marginal likelihood distribution for the pooled samples and the individual samples; a nonparametric test using Dirichlet process mixture models (DPMM) with exponential families. For the proposed nonparametric test, the prior on probability distributions allows the applications to multivariate data. Also utilising the Bayes factor as a model comparison tool, Stegle et al. [2010] used Gaussian processes (GP) to model the probability of the observed data under each model in the problem of testing whether a gene is differentially expressed. The values of the hyperparameters (i.e. the kernel hyperparameters and the variance of the noise distribution of the GP model) were set to those that maximise the log posterior distribution of the hyperparameters. While this approach is for detecting the genes that are differentially expressed, they proposed a mixture type of approach for detecting the intervals of the time series such that the effect is present. A binary switch variable was introduced at every observation time point to determine the model that describes the expression level at that time point. Posterior inference of such variable is achieved through variational approximation.

In the independence testing literature, Filippi et al. [2016] proposed to model the probability distributions using DPMM with Gaussian distributions for pairwise dependency detection in large multivariate datasets. Though in theory the Bayes factor with DPM on the unknown densities can be computed via the marginal likelihood, this requires integrating over infinite dimensional parameter space which results in an intractable form [Filippi et al., 2016]. Hence the problem was reformulated using a mixture modelling approach proposed recently by Kamary et al. [2014] where

¹Under the null hypothesis, the samples X and Y are assumed to follow the same distribution, i.e. $P = Q$. Here we denote this distribution as $P_{x,y}$.

hypotheses are the components of a mixture model and the posterior distribution of the mixing proportion is the outcome of the test. More specifically, they propose to model the observations $\mathcal{D}$ by an encompassing mixture model

$$\mathcal{D} \sim \alpha P(\mathcal{D}|\theta_0, M_0) + (1 - \alpha)P(\mathcal{D}|\theta_1, M_1) \quad (6.10)$$

with $0 \leq \alpha \leq 1$ the mixing proportion also to be inferred. Kamary et al. [2014] argues the posterior distribution on α allows for a more thorough assessment of the strength of the support of one model against the other comparing to the single numerical value given by the Bayes factor. It removes the artificial prior probabilities on model indices $P(M_0)$ and $P(M_1)$ required by Bayes factor computations. Most notably, the approach allows the use of noninformative priors when the two competing hypotheses share the same set of parameters. This was prohibitive in the classical Bayesian two-sample testing approach like Bayes factor [DeGroot, 1973].

In this work, we will adapt the classical Bayes factor approach to evaluate the evidence in favour of the null hypothesis. The mixture modelling formulation of the problem will be left as future work. The two-sample testing problem in the most general form as presented above need not have $n_x = n_y$. For simplicity, we restrict our attention to the paired two-sample testing problem, i.e. $\{(x_i, y_i)\}_{i=1}^n$. Instead of considering the observations directly, we propose to work with the witness function (Equation 2.17) of the MMD (Definition 2.3.1), which is proportional to the difference between the two distributions' mean embeddings in the RKHS. Inspired by the work of Flaxman et al. [2016] where the distribution mean embedding is modelled with a Gaussian process prior and a normal likelihood, we use a similar model for the witness function. Intuitively, to model the kernel mean embedding for X (or for Y) directly with a GP prior is not ideal as kernel mean embeddings for a non-negative kernel (like Gaussian kernel and Matérn kernel) are never negative, but the draws from any GP prior can be negative. Hence, it seems more suitable to place a prior directly on the difference between the mean embeddings as we do in this contribution. A further advantage of modelling the difference is that we will no longer require the independence assumption between the random variables X and Y . Such assumption is common in two-sample testing literature both in the frequentist and the Bayesian setting, in particular, a two-sample test based on MMD requires such assumption.

6.2 Proposed Method

Consider a paired data set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ with $x_i, y_i \in \mathcal{X}$ for some generic domains $\mathcal{X}$. Further, let $x_i \stackrel{i.i.d.}{\sim} P$ and $y_i \stackrel{i.i.d.}{\sim} Q$ for some unknown distributions P and Q . Let $k_\theta(\cdot, \cdot)$ be a positive definite kernel parameterised by θ , with the corresponding reproducing kernel Hilbert space $\mathcal{H}$.

In this chapter, we develop a Bayesian two-sample test based on the witness function corresponding to the integral probability metric formulation of the MMD (Section 2.3.2). We consider the empirical estimate of the witness function evaluated at a set of locations and propose a Bayesian inference scheme so that the relative evidence in favour of H_0 and H_1 is quantified. The proposed test is conditional on the choice of the family of kernels parameterised by θ . To ensure the injectivity of the kernel mean maps, we restrict our attention to the set of characteristic kernels, and in particular we work with the Gaussian RBF kernel in this Chapter, but other kernels are readily applicable to the framework. To emphasize the dependence of the kernel function on the lengthscale parameter θ , we write $k_\theta(\cdot, \cdot)$. Denote the respective kernel mean embeddings for X and Y as

$$\mu_X = \mathbb{E}_X(k_\theta(\cdot, X)) = \int_{\mathcal{X}} k_\theta(\cdot, x)P(dx), \quad (6.11)$$

$$\mu_Y = \mathbb{E}_Y(k_\theta(\cdot, Y)) = \int_{\mathcal{Y}} k_\theta(\cdot, y)Q(dy), \quad (6.12)$$

with the empirical estimators and the corresponding estimates denoted by

$$\hat{\mu}_X = \frac{1}{n} \sum_{i=1}^n k_\theta(\cdot, X_i) \quad \text{and} \quad \hat{\mu}_x = \frac{1}{n} \sum_{i=1}^n k_\theta(\cdot, x_i), \quad (6.13)$$

$$\hat{\mu}_Y = \frac{1}{n} \sum_{i=1}^n k_\theta(\cdot, Y_i) \quad \text{and} \quad \hat{\mu}_y = \frac{1}{n} \sum_{i=1}^n k_\theta(\cdot, y_i). \quad (6.14)$$

We denote the witness function up to proportionality as $\delta = \mu_X - \mu_Y$, which is simply the difference between the kernel mean embeddings. Under the null hypothesis, the two distributions P and Q are the same and with the use of characteristic kernels, all information of the probability distribution is preserved through the kernel mean embeddings μ_X and μ_Y . Hence, the null hypothesis is equivalent to $\delta = 0$ and the alternative is equivalent to $\delta \neq 0$.

Given the set of evaluation points $\mathbf{z} = \{z_i\}_{i=1}^p \in \mathcal{X}$, we define the evaluation of δ at $\mathbf{z}$ as

$$\delta(z_i) = \mu_X(z_i) - \mu_Y(z_i) \quad (6.15)$$

$$= \mathbb{E}_X(k_\theta(z_i, X)) - \mathbb{E}_Y(k_\theta(z_i, Y)), \quad \forall i = 1, \dots, p \quad (6.16)$$

$$\delta(\mathbf{z}) = (\delta(z_1), \dots, \delta(z_p))^\top \in \mathbb{R}^p. \quad (6.17)$$

Such evaluations $\delta(\mathbf{z})$ will act as the quantity of interest of our proposed model, while the empirical estimate of $\delta(\mathbf{z})$ on a given set of data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ will be regarded as the observations. This will be made precise in the following sections. Ideally, half of the evaluation points $\mathbf{z} = \{z_i\}_{i=1}^p$ are sampled from P , while the other half are sampled from Q . When direct sampling is not possible (e.g. when we have access to the distributions only through samples), the evaluation points are subsampled from the given data set. We define the p -dimensional vector Δ as the empirical estimate of $\delta(\mathbf{z})$:

$$\Delta = \{\hat{\mu}_x(z_j) - \hat{\mu}_y(z_j)\}_{j=1}^p \quad (6.18)$$

$$= \left\{ \frac{1}{n} \sum_{i=1}^n k_\theta(x_i, z_j) - \frac{1}{n} \sum_{i=1}^n k_\theta(y_i, z_j) \right\}_{j=1}^p \quad (6.19)$$

$$= \left\{ \frac{1}{n} \sum_{i=1}^n (k_\theta(x_i, z_j) - k_\theta(y_i, z_j)) \right\}_{j=1}^p. \quad (6.20)$$

and the corresponding random variable as $\Delta_{XY} := \{\hat{\mu}_X(z_j) - \hat{\mu}_Y(z_j)\}_{j=1}^p$.

Following the classical Bayesian two-sample testing framework, we will quantify the evidence in favour of the two-sample coming from the same distribution vs different distributions through Bayes factor. However, in contrast to (6.4) and (6.5), the independence assumption between random variables X and Y is no longer assumed and hence we compute the Bayes factor as

$$BF = \frac{P(\Delta|H_0)}{P(\Delta|H_1)} = \frac{P(H_0|\Delta)P(H_1)}{P(H_1|\Delta)P(H_0)}. \quad (6.21)$$

We introduce the parameter $M = \{0, 1\}$, such that $M = 0$ indicates the null hypothesis and $M = 1$ indicates the alternative hypothesis. (6.21) can be written as

$$BF = \frac{P(\Delta|M=0)}{P(\Delta|M=1)} = \frac{P(M=0|\Delta)P(M=1)}{P(M=1|\Delta)P(M=0)}. \quad (6.22)$$

Following the Bayesian kernel embeddings approach of Flaxman et al. [2016], we propose to model δ with a Gaussian process (GP) prior under the alternative model in Section 6.3. Assuming a Gaussian noise model, we derive the marginal likelihood of Δ for fixed kernel lengthscale parameter θ . When such a parameter is unknown, the framework of Flaxman et al. [2016] enables the derivation of the posterior distribution of the parameter given the observations. In Section 6.3.1, we detail the computation of the covariance matrix of the Gaussian process prior following the approach by Flaxman et al. [2016]. In Section 6.3.2, we propose two estimation methods of the covariance matrix of the noise model different from the isotropic model used in Flaxman et al. [2016]. To alleviate the heavy computation burden of computing the marginal likelihood of the dataset $\{(x_i, y_i)\}_{i=1}^n | \theta$, Section 6.3.3 proposes alternative formulation of the likelihood utilising the Kronecker product structure of our problem. In Section 6.3.4, we derive the formulas for the Jacobian computation.

Under the null hypothesis, the model simplifies significantly due to the zeroing of δ , i.e. we pose a Gaussian noise model for $\Delta | \theta$ directly. The noise covariance estimation and the Jacobian computation follows that of the alternative model in Section 6.3, while the fast computation of the marginal likelihood mimics the derivation in Section 6.3.3. Section 6.5 proposes a Metropolis Hasting within Gibbs type of approach for inferring the posterior distribution of θ and that of the model given the observed data. Section 6.6 studies the performance of the proposed method on 5 synthetic data experiments.

6.3 Alternative Model: $p(\theta | M = 1, \mathcal{D})$

Under the alternative hypothesis, $\delta = \mu_X - \mu_Y \neq 0$. We propose to model the unknown quantity δ using a Gaussian Process (GP) prior. Draws from a naively defined prior $\mathcal{GP}(0, k_\theta(\cdot, \cdot))$ would almost surely fall outside of the RKHS $\mathcal{H}$ that corresponds to $k_\theta(\cdot, \cdot)$ [Wahba, 1990]. Hence, Flaxman et al. [2016] proposed to define the GP prior as

$$\delta | \theta \sim \mathcal{GP}(0, r_\theta(\cdot, \cdot)) \quad (6.23)$$

with the covariance operator $r_\theta(\cdot, \cdot)$ defined as

$$r_\theta(z, z') := \int k_\theta(z, u) k_\theta(u, z') \nu(du) \quad (6.24)$$

where ν is any finite measure on $\mathcal{X}$. Using results from Lukić and Beder [2001] and Theorem 4.27 of Steinwart and Christmann [2008], Flaxman et al. [2016] show that such choice of r_θ ensures that $\delta \in \mathcal{H}$ with probability 1 by the nuclear dominance for k_θ over r_θ for any stationary kernel k_θ and more generally $\int k_\theta(x, x)\nu(dx) < \infty$. Intuitively, the new covariance operator r_θ is a smoother version of k_θ since it is the convolution of k_θ with itself with respect to a finite measure ν [Flaxman et al., 2016]. For our particular choice of k_θ being a Gaussian RBF kernel on $\mathcal{X} = \mathbb{R}^D$, we will derive the form of r_θ following Flaxman et al. [2016] in Section 6.3.1.

Given the set of evaluation points $\{z_j\}_{j=1}^p$, the prior translates into a p -dimensional multivariate Gaussian distribution

$$\delta(\mathbf{z})|\theta \sim N(0, R_\theta) \quad (6.25)$$

with $[R_\theta]_{ij} = r_\theta(z_i, z_j)$. We link the empirical estimate Δ with the true differences δ evaluated at this set of evaluation points through a Gaussian likelihood of the model. This is an approximation of the true likelihood which hinges on the common “Gaussianity in the feature space” assumption in the kernel method literature and is also utilised in Flaxman et al. [2016]. We write it as

$$\Delta_{XY}|\delta, \theta \sim \mathcal{N}([\delta(z_1), \dots, \delta(z_p)]^\top, \frac{1}{n}\Sigma_\theta). \quad (6.26)$$

We proposed to estimate Σ_θ empirically in Section 6.3.2. The constant $\frac{1}{n}$ is for notational convenience to be seen later. Integrating out the prior distribution of δ , we obtain the marginal likelihood

$$\Delta_{XY}|\theta \sim \mathcal{N}(0_p, R_\theta + \frac{1}{n}\Sigma_\theta). \quad (6.27)$$

For the purpose of hyperparameter learning, we would like to compute the posterior distribution of $\theta|\{(x_i, y_i)\}_{i=1}^n$. As a first step, we will need to integrate out the model parameter δ and compute the marginal probability of the observations $\{(x_i, y_i)\}_{i=1}^n$ given the hyperparameter θ . Note that Δ implicitly depends on the parameter θ through the kernel functions, hence we follow the exposition of Flaxman et al. [2016] to apply a change of variables such that the θ dependence is separated out.

For every fixed pair of (x_i, y_i) , define the finite dimensional feature map $\phi_\theta : \mathbb{R}^D \rightarrow \mathbb{R}^p$ as

$$\phi_\theta(x) - \phi_\theta(y) = [k_\theta(x, z_1) - k_\theta(y, z_1), \dots, k_\theta(x, z_p) - k_\theta(y, z_p)]^\top \in \mathbb{R}^p. \quad (6.28)$$

Note, $k_\theta(X_i, z_j) - k_\theta(Y_i, z_j)$ are independent for all $i = 1, \dots, n$ given δ . This implies that $n\Delta_{XY}$ can be written as a sum of independent random variables: $\Delta_{XY} = \frac{1}{n} \sum_{i=1}^n (\phi_\theta(X_i) - \phi_\theta(Y_i))$. By Cramer's decomposition theorem, we obtain the distribution of each of the contributing components

$$\phi_\theta(X_i) - \phi_\theta(Y_i) | \delta, \theta \sim \mathcal{N}([\delta_\theta(z_1), \dots, \delta_\theta(z_p)]^\top, \Sigma_\theta). \quad (6.29)$$

Applying the change of variable $(x, y) \mapsto g_\theta(x, y) := (\phi_\theta(x) - \phi_\theta(y))$ and using the generalisation of the change-of-variable formula to non-square Jacobian matrices [Ben-Israel, 1999] to obtain a distribution for (x, y) conditionally on δ

$$p(x, y | \delta, \theta) = p(g_\theta(x, y) | \delta, \theta) \text{vol}[J_\theta(x, y)]. \quad (6.30)$$

The Jacobian J_θ is a $p \times 2D$ matrix with

$$J_\theta(x, y) = \begin{cases} \frac{\partial g_\theta(x, y)_{z_i}}{\partial x^j}, & \text{for } j = 1, \dots, D \\ \frac{\partial g_\theta(x, y)_{z_i}}{\partial y^j}, & \text{for } j = D + 1, \dots, 2D \end{cases} \quad (6.31)$$

and $\text{vol}[J_\theta(x, y)] := \sqrt{\det(J_\theta(x, y)^\top J_\theta(x, y))}$. This can be further simplified into

$$J_\theta(x, y) = \begin{cases} \frac{\partial k_\theta(x, z_i)}{\partial x^j}, & \text{for } j = 1, \dots, D \\ \frac{\partial k_\theta(y, z_i)}{\partial y^j}, & \text{for } j = D + 1, \dots, 2D. \end{cases} \quad (6.32)$$

The detailed computation of the Jacobian matrix will be discussed in Section 6.3.4. By the conditional independence of $g(x_i, y_i)$ for each i given δ and noting that θ represents the hyperparameter of the kernel $k_\theta(\cdot, \cdot)$, then

$$\begin{aligned} p(\{x_i, y_i\}_{i=1}^n | \delta, \theta) &= \prod_{i=1}^n p(g_\theta(x_i, y_i) | \delta, \theta) \text{vol}[J_\theta(x_i, y_i)] \\ &= \prod_{i=1}^n \mathcal{N}(g_\theta(x_i, y_i); \delta_\theta(\mathbf{z}), \Sigma_\theta) \text{vol}[J_\theta(x_i, y_i)] \\ &= \mathcal{N}(\text{vec}(G_\theta); m_\theta(\mathbf{z}), I_n \otimes \Sigma_\theta) \prod_{i=1}^n \text{vol}[J_\theta(x_i, y_i)] \end{aligned} \quad (6.33)$$

where $G_\theta = [g_\theta(x_1, y_1), \dots, g_\theta(x_n, y_n)]$ is a $p \times n$ matrix and $\text{vec}(G_\theta) \in \mathbb{R}^{np}$ is the vectorisation of the matrix G_θ . The mean vector $m_\theta(\mathbf{z}) := [\delta_\theta(\mathbf{z})^\top \dots \delta_\theta(\mathbf{z})^\top]^\top$ where each $\delta_\theta(\mathbf{z}) = [\delta_\theta(z_1), \dots, \delta_\theta(z_p)]^\top$. To obtain the marginal pseudolikelihood of $p(\{(x_i, y_i)\}_{i=1}^n | \theta)$,

we compute the integral

$$\begin{aligned}
p(\{(x_i, y_i)\}_{i=1}^n | \theta) &= \int p(\{(x_i, y_i)\}_{i=1}^n | \delta, \theta) p(\delta | \theta) d\delta \\
&= \int \mathcal{N}(\text{vec}(G_\theta); m_\theta(z), I_n \otimes \Sigma_\theta) \mathcal{N}(\delta; 0, R_\theta) \prod_{i=1}^n \text{vol}[J_\theta(x_i, y_i)] d\delta \\
&= \mathcal{N}(\text{vec}(G_\theta); 0, \mathbb{1}\mathbb{1}^\top \otimes R_\theta + I_n \otimes \Sigma_\theta) \prod_{i=1}^n \text{vol}[J_\theta(x_i, y_i)]. \quad (6.34)
\end{aligned}$$

More explicitly, the log marginal pseudolikelihood can be written as

$$\begin{aligned}
\log p(\{(x_i, y_i)\}_{i=1}^n | \theta) &\propto -\frac{1}{2} \log \det(\mathbb{1}\mathbb{1}^\top \otimes R_\theta + I_n \otimes \Sigma_\theta) \\
&\quad -\frac{1}{2} \text{vec}(G_\theta)^\top (\mathbb{1}\mathbb{1}^\top \otimes R_\theta + I_n \otimes \Sigma_\theta)^{-1} \text{vec}(G_\theta) \\
&\quad + \frac{1}{2} \sum_{i=1}^n \log \det(J_\theta(x_i, y_i)^\top J_\theta(x_i, y_i)). \quad (6.35)
\end{aligned}$$

Although the derivations presented here follow essentially the same steps as Flaxman et al. [2016], it is important to note that different from Flaxman et al. [2016], we model the difference between the empirical mean embeddings of the two distributions of interest rather than the embedding of a single distribution. This has several implications. As discussed in Flaxman et al. [2016], the marginal pseudolikelihood involves the computation of the inverse and the log determinant of an np -dimensional matrix. A naive direct implementation would require a prohibitive computation of $\mathcal{O}(n^3 p^3)$. Since we consider the difference between the empirical mean embeddings, the efficient computation utilising eigendecompositions of the relevant matrices [Flaxman et al., 2016, A.4] cannot be applied directly. Fortunately, the special form of the corresponding $np \times np$ covariance matrix allows faster computation following Kronecker product algebra and the applications of matrix determinant lemma and Woodbury identity. We detail the derivation in Section 6.3.3. Further, the Jacobian of transformation requires the concatenation of the derivative with respect to the variable X and that to the variable Y , we will present the details in Section 6.3.4.

Recall, in this chapter, we use Gaussian RBF kernel with the lengthscale parameter θ . This choice leads to a convenient form of the covariance function $r_\theta(\cdot, \cdot)$ such that the samples from a GP with r_θ lies in the correct RKHS, which we will give the details below in Section 6.3.1.

6.3.1 Covariance Matrix R_θ Computation

Flaxman et al. [2016] showed (in A.3) that the covariance function r_θ of square exponential kernels

$$k_\theta(x, y) = \exp \left(-\frac{1}{2}(x - y)^\top \tilde{\Sigma}_\theta^{-1}(x - y) \right) \quad (6.36)$$

with $x, y \in \mathbb{R}^D$ and diagonal covariance $\tilde{\Sigma}_\theta = \theta I = (\theta^{(1)}, \dots, \theta^{(D)})^\top I$ can be written as

$$r_\theta(x, y) = \pi^{D/2} \left(\prod_{d=1}^D \theta^{(d)} \right)^{1/2} \exp \left(-\frac{1}{2}(x - y)^\top (2\theta I_D)^{-1}(x - y) \right). \quad (6.37)$$

When $\theta^{(1)} = \dots = \theta^{(D)} = \theta$, the above can be further simplified as

$$r_\theta(x, y) = \pi^{D/2} \theta^{D/2} \exp \left(-\frac{1}{4\theta}(x - y)^\top (x - y) \right). \quad (6.38)$$

We use this form of the covariance function in our experiments.

6.3.2 Covariance Matrix Σ_θ Estimation

In this section, we consider two approaches of estimating the empirical covariance matrix Σ_θ . It is important to note that the inducing points $\mathbf{z} = \{z_i\}_{i=1}^p$ are treated as fixed. The first approach computes the covariance of the random variable Δ_{XY} directly and assumes the independence between the random variables X and Y . The second approach, on the other hand, computes the empirical covariance matrix of $g_\theta(X, Y)$ and does not require the independence assumption.

Method 1

In this approach, we consider the covariance of the random variable Δ_{XY} . The ij^{th} component of the estimated covariance matrix $\hat{\Sigma}_{ij} = Cov([\Delta_{XY}]_i, [\Delta_{XY}]_j)$ can be written as

$$\mathbb{E}_{XY}((\hat{\mu}_X(z_i) - \hat{\mu}_Y(z_i))(\hat{\mu}_X(z_j) - \hat{\mu}_Y(z_j))) \quad (6.39)$$

$$= \mathbb{E}_{XY}(\hat{\mu}_X(z_i)\hat{\mu}_X(z_j) - \hat{\mu}_X(z_i)\hat{\mu}_Y(z_j) + \hat{\mu}_Y(z_i)\hat{\mu}_Y(z_j) - \hat{\mu}_Y(z_i)\hat{\mu}_X(z_j)). \quad (6.40)$$

The first term of (6.40) can be expanded as

$$\mathbb{E}_X(\hat{\mu}_X(z_i)\hat{\mu}_X(z_j)) \quad (6.41)$$

$$= \mathbb{E}_X \left(\frac{1}{n^2} \sum_{a=1}^n \sum_{b=1}^n k(X_a, z_i) k(X_b, z_j) \right) \quad (6.42)$$

$$= \frac{1}{n^2} \left[\sum_{a=1}^n \mathbb{E}_X(k(X_a, z_i)k(X_a, z_j)) + \sum_{b=1}^n \sum_{b \neq a} \mathbb{E}_X(k(X_a, z_i)k(X_b, z_j)) \right] \quad (6.43)$$

$$= \frac{1}{n^2} [n\mathbb{E}_X(k(X, z_i)k(X, z_j)) + n(n-1)\mathbb{E}_X(k(X, z_i)k(X, z_j))] \quad (6.44)$$

Given observations $\{(x_i, y_i)\}_{i=1}^n$, the above is now readily estimated as

$$\approx \frac{1}{n^2} \sum_{a=1}^n k(x_a, z_i)k(x_a, z_j) + \frac{n-1}{n} \left(\frac{1}{n} \sum_{a=1}^n k(x_a, z_i) \right) \left(\frac{1}{n} \sum_{b=1}^n k(x_b, z_j) \right) \quad (6.45)$$

$$= \frac{1}{n^2} K_{z_i \mathbf{x}} K_{\mathbf{x} z_j} + \frac{n-1}{n} \left(\frac{1}{n} K_{z_i \mathbf{x}} \mathbb{1} \right) \left(\frac{1}{n} K_{z_j \mathbf{x}} \mathbb{1} \right) \quad (6.46)$$

$$= \frac{1}{n^2} K_{z_i \mathbf{x}} K_{\mathbf{x} z_j} + \frac{n-1}{n} \hat{\mu}_x(z_i) \hat{\mu}_x(z_j), \quad (6.47)$$

where $\mathbb{1}$ is a vector of 1s in $\mathbb{R}^n$ and $\mathbf{x} = [x_1 \cdots x_n]^\top \in \mathbb{R}^{n \times D}$. Similarly, the third term in (6.40) can be approximated as

$$\mathbb{E}_Y(\hat{\mu}_Y(z_i)\hat{\mu}_Y(z_j)) \approx \frac{1}{n^2} K_{z_i \mathbf{y}} K_{\mathbf{y} z_j} + \frac{n-1}{n} (\hat{\mu}_y(z_i))(\hat{\mu}_y(z_j)) \quad (6.48)$$

where $\mathbf{y} = [y_1 \cdots y_n]^\top \in \mathbb{R}^{n \times D}$. Assuming that X and Y are independent, the second term in (6.40) can be empirically estimated as

$$\mathbb{E}_{XY}(\hat{\mu}_X(z_i)\hat{\mu}_Y(z_j)) \quad (6.49)$$

$$= \left(\frac{1}{n} \sum_{a=1}^n \mathbb{E}_X(k(X_a, z_i)) \right) \left(\frac{1}{n} \sum_{b=1}^n \mathbb{E}_Y(k(Y_b, z_j)) \right) \quad (6.50)$$

$$\approx (\hat{\mu}_x(z_i))(\hat{\mu}_y(z_j)) \quad (6.51)$$

and similarly for the fourth term of (6.40). When we consider all $\mathbf{z} = \{z_1, \dots, z_p\}$, (6.40) can be written as

$$\mathbb{E}_{XY}((\hat{\mu}_X(\mathbf{z}) - \hat{\mu}_Y(\mathbf{z}))(\hat{\mu}_X(\mathbf{z}) - \hat{\mu}_Y(\mathbf{z}))^\top) \quad (6.52)$$

$$\begin{aligned} &\approx \frac{1}{n^2} K_{\mathbf{z}\mathbf{x}} K_{\mathbf{x}\mathbf{z}} + \frac{n-1}{n} \hat{\mu}_x(\mathbf{z}) \hat{\mu}_x(\mathbf{z})^\top - \hat{\mu}_x(\mathbf{z}) \hat{\mu}_y(\mathbf{z})^\top \\ &\quad + \frac{1}{n^2} K_{\mathbf{z}\mathbf{y}} K_{\mathbf{y}\mathbf{z}} + \frac{n-1}{n} \hat{\mu}_y(\mathbf{z}) \hat{\mu}_y(\mathbf{z})^\top - \hat{\mu}_y(\mathbf{z}) \hat{\mu}_x(\mathbf{z})^\top. \end{aligned} \quad (6.53)$$

The second term of $\hat{\Sigma} = Cov(\Delta_{XY})$ can be written as

$$[\mathbb{E}_{XY}(\hat{\mu}_X(\mathbf{z}) - \hat{\mu}_Y(\mathbf{z}))][\mathbb{E}(\hat{\mu}_X(\mathbf{z}) - \hat{\mu}_Y(\mathbf{z}))]^\top \quad (6.54)$$

$$\approx \hat{\mu}_x(\mathbf{z})\hat{\mu}_x(\mathbf{z})^\top + \hat{\mu}_y(\mathbf{z})\hat{\mu}_y(\mathbf{z})^\top - \hat{\mu}_x(\mathbf{z})\hat{\mu}_y(\mathbf{z})^\top - \hat{\mu}_y(\mathbf{z})\hat{\mu}_x(\mathbf{z})^\top \quad (6.55)$$

Overall the empirical estimate of $\hat{\Sigma}$ can be computed as

$$Cov(\Delta_{XY}) \quad (6.56)$$

$$= \mathbb{E}_{XY}((\hat{\mu}_X(\mathbf{z}) - \hat{\mu}_Y(\mathbf{z}))(\hat{\mu}_X(\mathbf{z}) - \hat{\mu}_Y(\mathbf{z}))^\top) - \mathbb{E}(\hat{\mu}_X(\mathbf{z}) - \hat{\mu}_Y(\mathbf{z}))\mathbb{E}(\hat{\mu}_X(\mathbf{z}) - \hat{\mu}_Y(\mathbf{z}))^\top \quad (6.57)$$

$$\approx \frac{1}{n^2}K_{\mathbf{zx}}K_{\mathbf{xz}} - \frac{1}{n}\hat{\mu}_x(\mathbf{z})\hat{\mu}_x(\mathbf{z})^\top + \frac{1}{n^2}K_{\mathbf{zy}}K_{\mathbf{yz}} - \frac{1}{n}\hat{\mu}_y(\mathbf{z})\hat{\mu}_y(\mathbf{z})^\top \quad (6.58)$$

$$= \frac{1}{n^2}K_{\mathbf{zx}}HK_{\mathbf{xz}} + \frac{1}{n^2}K_{\mathbf{zy}}HK_{\mathbf{yz}} \quad (6.59)$$

Since $\hat{\mu}_x(\mathbf{z})\hat{\mu}_x(\mathbf{z})^\top = \frac{1}{n^2}K_{\mathbf{zx}}\mathbb{1}\mathbb{1}^\top K_{\mathbf{xz}}$ and $H := I - \frac{1}{n}\mathbb{1}\mathbb{1}^\top$, the final estimation equation (6.59) follows.

Method 2

In this approach, we consider the random variable

$$g_\theta(X, Y) = (k(X, z_1) - k(Y, z_1), \dots, k(X, z_p) - k(Y, z_p))^\top.$$

The ij^{th} component of the empirical covariance matrix of $g_\theta(X, Y)$ can be written as

$$Cov(k(X, z_i) - k(Y, z_i), k(X, z_j) - k(Y, z_j)) \quad (6.60)$$

$$= \mathbb{E}_{XY}(((k(X, z_i) - k(Y, z_i))(k(X, z_j) - k(Y, z_j))) - \mathbb{E}_{XY}((k(X, z_i) - k(Y, z_i)))\mathbb{E}_{XY}((k(X, z_j) - k(Y, z_j)))) \quad (6.61)$$

The first term of (6.61) can be written as

$$\mathbb{E}_{XY}((k(X, z_i) - k(Y, z_i))(k(X, z_j) - k(Y, z_j))) \quad (6.62)$$

$$= \mathbb{E}_{XY}(k(X, z_i)k(X, z_j) - k(X, z_i)k(Y, z_j) - k(Y, z_i)k(X, z_j) + k(Y, z_i)k(Y, z_j)) \quad (6.63)$$

$$\approx \frac{1}{n}K_{z_i\mathbf{x}}K_{\mathbf{x}z_j} - \frac{1}{n}K_{z_i\mathbf{x}}K_{\mathbf{y}z_j} - \frac{1}{n}K_{z_j\mathbf{x}}K_{\mathbf{y}z_i} + \frac{1}{n}K_{z_i\mathbf{y}}K_{\mathbf{y}z_j} \quad (6.64)$$

While the second term of (6.61) can be written as

$$\mathbb{E}_{XY}(k(X, z_i) - k(Y, z_i))\mathbb{E}_{XY}(k(X, z_j) - k(Y, z_j)) \quad (6.65)$$

$$= (\mu_X(z_i) - \mu_Y(z_i))(\mu_X(z_j) - \mu_Y(z_j)) \quad (6.66)$$

$$\approx \frac{1}{n^2}K_{z_i\mathbf{x}}\mathbb{1}\mathbb{1}^\top K_{\mathbf{x}z_j} - \frac{1}{n^2}K_{z_i\mathbf{x}}\mathbb{1}\mathbb{1}^\top K_{\mathbf{y}z_j} - \frac{1}{n^2}K_{z_i\mathbf{y}}\mathbb{1}\mathbb{1}^\top K_{\mathbf{x}z_i} + \frac{1}{n^2}K_{z_i\mathbf{y}}\mathbb{1}\mathbb{1}^\top K_{\mathbf{y}z_j} \quad (6.67)$$

Hence, combining these two terms and consider all $\mathbf{z} = \{z_1, \dots, z_p\}$:

$$\text{Cov}(g_\theta(X, Y)) \tag{6.68}$$

$$\approx \frac{1}{n} K_{\mathbf{zx}} H K_{\mathbf{xz}} + \frac{1}{n} K_{\mathbf{zy}} H K_{\mathbf{yz}} - \frac{1}{n} K_{\mathbf{zx}} H K_{\mathbf{yz}} - \frac{1}{n} K_{\mathbf{zy}} H K_{\mathbf{xz}} \tag{6.69}$$

$$= \frac{1}{n} G_\theta H G_\theta^\top. \tag{6.70}$$

To see how (6.70) follows, recall $G_\theta := [g_\theta(x_1, y_1), \dots, g_\theta(x_n, y_n)] \in \mathbb{R}^{p \times n}$ which can be equivalently written in terms of kernel matrices as $G_\theta = K_{\mathbf{zx}} - K_{\mathbf{zy}}$.

Note that the distribution of $g_\theta(x_i, y_i)$ arises from the distribution of Δ due to Cramer's decomposition theorem, this implies $\text{Cov}(\Delta_{XY}) = \frac{1}{n} \text{Cov}(g_\theta(X, Y))$. If in addition, we assume the independence between X and Y , the computation of $\text{Cov}(g_\theta(X, Y))$ will be simplified and the cross terms in (6.69) will be zero. However, even in the case where independence assumption is satisfied, numerical instability leads to slight difference of the covariance computation between the two estimation methods. Nonetheless, we did not observe any difference of the resulting posterior inference in that case. The first estimation approach is used in all the experiments of Section 6.6.

6.3.3 Efficient Computation of Marginal Pseudolikelihood

In this section, we illustrate the efficient computation of the multivariate Gaussian distribution utilising the Kronecker product structure of our problem. Note, Flaxman et al. [2016] used the identity matrix as Σ_θ which results in their particular derivation of the fast computation of the marginal pseudolikelihood in their section A.4. However, we propose a different approach considering the full covariance matrix Σ_θ as estimated in the previous section.

Proposition 6.3.1. *Denote $W_\theta := \mathbf{1}\mathbf{1}^\top \otimes R_\theta + I_n \otimes \Sigma_\theta$. Then,*

$$(a) \det(W_\theta) = \det(\Sigma_\theta + nR_\theta) \det(\Sigma_\theta)^{n-1}.$$

$$(b) \text{vec}(G_\theta)^\top W^{-1} \text{vec}(G_\theta) = \text{Tr} \left((\Sigma_\theta + nR_\theta)^{-1} G_\theta G_\theta^\top + \left(\frac{1}{n} \Sigma_\theta R_\theta^{-1} \Sigma_\theta + \Sigma_\theta \right)^{-1} G_\theta H G_\theta^\top \right)$$

where $H = I - \frac{1}{n} \mathbf{1}\mathbf{1}^\top$.

Proof. (a) Using matrix determinant lemma, and note that for any $m \times m$ matrix,

$$\mathbf{1}\mathbf{1}^\top \otimes A = (\mathbf{1} \otimes I_m) A (\mathbf{1} \otimes I_m)^\top.$$

Hence, the first term of W can be written as

$$\mathbf{1}\mathbf{1}^\top \otimes R_\theta = (\mathbf{1} \otimes I_m) R_\theta (\mathbf{1} \otimes I_m)^\top.$$

For the second term of W ,

$$\det(I_n \otimes \Sigma_\theta) = \det(I_n)^m \det(\Sigma_\theta)^n = \det(\Sigma_\theta)^n$$

by the property of the determinant of Kronecker product. We can then write

$$\begin{aligned} & \det(\mathbf{1}\mathbf{1}^\top \otimes R_\theta + I_n \otimes \Sigma_\theta) \\ &= \det((\mathbf{1} \otimes I_m) R_\theta (\mathbf{1} \otimes I_m) + I_n \otimes \Sigma_\theta) \\ &= \det(I + (\mathbf{1} \otimes I_m)^\top (I_n \otimes \Sigma_\theta)^{-1} (\mathbf{1} \otimes I_m) R_\theta) \det(\Sigma_\theta)^n \end{aligned} \quad (6.71)$$

The last equation follows from the property that $\det(A+UV^\top) = \det(I+V^\top A^{-1}U) \det(A)$ given conformable matrices. Recall the properties of Kronecker products: $(\mathbf{1} \otimes I_m)^\top = \mathbf{1}^\top \otimes I_m^\top$; $(I_n \otimes \Sigma_\theta)^{-1} = I_n^{-1} \otimes \Sigma_\theta^{-1}$ and $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ for conformable matrices. By repetitive applications of these properties, the above can be simplified into

$$\begin{aligned} & \det(I + (\mathbf{1}^\top \otimes \Sigma_\theta^{-1})(\mathbf{1} \otimes I_m) R_\theta) \det(\Sigma_\theta)^n \\ &= \det(I + (\mathbf{1}^\top \mathbf{1}) \otimes (\Sigma_\theta^{-1} I_m) R_\theta) \det(\Sigma_\theta)^n \\ &= \det(I + n \Sigma_\theta^{-1} R_\theta) \det(\Sigma_\theta)^n \\ &= \det(\Sigma_\theta^{-1} (\Sigma_\theta + n R_\theta)) \det(\Sigma_\theta)^n \\ &= \det(\Sigma_\theta)^{-1} \det(\Sigma_\theta + n R_\theta) \det(\Sigma_\theta)^n \\ &= \det(\Sigma_\theta + n R_\theta) \det(\Sigma_\theta)^{n-1} \end{aligned}$$

Hence the desired equation is obtained.

(b) Using Woodbury identity and the properties of the Kronecker product,

$$W_\theta^{-1} = I_n \otimes \Sigma_\theta^{-1} - (\mathbf{1} \otimes \Sigma_\theta^{-1}) ((R_\theta)^{-1} + n \Sigma_\theta^{-1})^{-1} (\mathbf{1}^\top \otimes \Sigma_\theta^{-1}). \quad (6.72)$$

Note that $(I_n \otimes \Sigma_\theta^{-1}) \text{vec}(G_\theta) = \text{vec}(\Sigma_\theta^{-1} G_\theta)$ and

$$\{\text{vec}(G_\theta)^\top (\mathbf{1} \otimes \Sigma_\theta^{-1})\}^\top = (\mathbf{1}^\top \otimes \Sigma_\theta^{-1}) \text{vec}(G_\theta) = \text{vec}(\Sigma_\theta^{-1} G_\theta \mathbf{1}) = \Sigma_\theta^{-1} G_\theta \mathbf{1},$$

and rearranging products under the trace, we have

$$\begin{aligned}
& \text{vec}(G_\theta)^\top W_\theta^{-1} \text{vec}(G_\theta) \\
&= \text{vec}(G_\theta)^\top \text{vec}(\Sigma_\theta^{-1} G_\theta) - (\Sigma_\theta^{-1} G_\theta \mathbf{1})^\top ((R_\theta)^{-1} + n \Sigma_\theta^{-1})^{-1} \Sigma_\theta^{-1} G_\theta \mathbf{1} \\
&= \text{Tr}(G_\theta^\top \Sigma_\theta^{-1} G_\theta) - \text{Tr}(\mathbf{1}^\top G_\theta^\top \Sigma_\theta^{-1} ((R_\theta)^{-1} + n \Sigma_\theta^{-1})^{-1} \Sigma_\theta^{-1} G_\theta \mathbf{1}) \\
&= \text{Tr} \left(\Sigma_\theta^{-1} G_\theta G_\theta^\top - \Sigma_\theta^{-1} ((R_\theta)^{-1} + n \Sigma_\theta^{-1})^{-1} \Sigma_\theta^{-1} G_\theta \mathbf{1} \mathbf{1}^\top G_\theta^\top \right). \tag{6.73}
\end{aligned}$$

Denoting the matrix under the trace by A , we can further simplify

$$\begin{aligned}
A &= \Sigma_\theta^{-1} G_\theta G_\theta^\top - \Sigma_\theta^{-1} ((n R_\theta)^{-1} + \Sigma_\theta^{-1})^{-1} \Sigma_\theta^{-1} G_\theta \left(\frac{1}{n} \mathbf{1} \mathbf{1}^\top \right) G_\theta^\top \\
&= ((n R_\theta)^{-1} + \Sigma_\theta^{-1}) ((n R_\theta)^{-1} + \Sigma_\theta^{-1})^{-1} \Sigma_\theta^{-1} G_\theta G_\theta^\top \\
&\quad - \Sigma_\theta^{-1} ((n R_\theta)^{-1} + \Sigma_\theta^{-1})^{-1} \Sigma_\theta^{-1} G_\theta \left(\frac{1}{n} \mathbf{1} \mathbf{1}^\top \right) G_\theta^\top \\
&= (n R_\theta)^{-1} ((n R_\theta)^{-1} + \Sigma_\theta^{-1})^{-1} \Sigma_\theta^{-1} G_\theta G_\theta^\top \\
&\quad + \Sigma_\theta^{-1} ((n R_\theta)^{-1} + \Sigma_\theta^{-1})^{-1} \Sigma_\theta^{-1} G_\theta \left(I_n - \frac{1}{n} \mathbf{1} \mathbf{1}^\top \right) G_\theta^\top \\
&= (\Sigma_\theta + n R_\theta)^{-1} G_\theta G_\theta^\top + \left(\frac{1}{n} \Sigma_\theta R_\theta^{-1} \Sigma_\theta + \Sigma_\theta \right)^{-1} G_\theta H G_\theta^\top.
\end{aligned}$$

□

Remark 6.3.1. *The above simplifies further when $\hat{\Sigma}_\theta$ is the empirical covariance of G_θ 's, i.e. $\hat{\Sigma}_\theta = \frac{1}{n} G_\theta H G_\theta^\top$. Indeed,*

$$\begin{aligned}
& \text{Tr} \left(\left(\hat{\Sigma}_\theta + n R_\theta \right)^{-1} G_\theta G_\theta^\top + \left(\frac{1}{n} \hat{\Sigma}_\theta R_\theta^{-1} \hat{\Sigma}_\theta + \hat{\Sigma}_\theta \right)^{-1} G_\theta H G_\theta^\top \right) \\
&= \text{Tr} \left(\left(\hat{\Sigma}_\theta + n R_\theta \right)^{-1} G_\theta G_\theta^\top + \left(\hat{\Sigma}_\theta + n R_\theta \right)^{-1} n R_\theta \hat{\Sigma}_\theta^{-1} G_\theta H G_\theta^\top \right) \\
&= \text{Tr} \left(\left(\hat{\Sigma}_\theta + n R_\theta \right)^{-1} G_\theta G_\theta^\top + \left(\hat{\Sigma}_\theta + n R_\theta \right)^{-1} n^2 R_\theta \right) \\
&= \text{Tr} \left((G_\theta H G_\theta^\top + n^2 R_\theta)^{-1} n G_\theta G_\theta^\top + (G_\theta H G_\theta^\top + n^2 R_\theta)^{-1} n^3 R_\theta \right) \\
&= \text{Tr} \left(n (G_\theta H G_\theta^\top + n^2 R_\theta)^{-1} (G_\theta G_\theta^\top + n^2 R_\theta) \right).
\end{aligned}$$

Using the efficient computation presented in this section, the log marginal pseu-

likelihood (6.35) can be written as

$$\begin{aligned}
& \log p(\{(x_i, y_i)\}_{i=1}^n | \theta) \\
&= \text{const} + -\frac{1}{2} \log \det(\Sigma_\theta + nR_\theta) - \frac{1}{2}(n-1) \log \det(\Sigma_\theta) \\
&\quad - \frac{1}{2} \text{Tr} \left((\Sigma_\theta + nR_\theta)^{-1} G_\theta G_\theta^\top + \left(\frac{1}{n} \Sigma_\theta R_\theta^{-1} \Sigma_\theta + \Sigma_\theta \right)^{-1} G_\theta H G_\theta^\top \right) \\
&\quad + \frac{1}{2} \sum_{i=1}^n \log \det(J_\theta(x_i, y_i)^\top J_\theta(x_i, y_i)). \tag{6.74}
\end{aligned}$$

6.3.4 Computation of the Jacobian Matrix

In this section, we detail the computation of the Jacobian term. As the variable under consideration is the same for the null and the alternative models, the Jacobian term remains the same. The variables under consideration is

$$g_\theta : (x, y) \mapsto (k_\theta(x, z_1) - k_\theta(y, z_1), \dots, k_\theta(x, z_p) - k_\theta(y, z_p))$$

where $g_\theta : \mathbb{R}^{2D} \rightarrow \mathbb{R}^p$. For each $i = 1, \dots, n$, the Jacobian matrix $J_\theta(x_i, y_i) \in \mathbb{R}^{p \times 2D}$:

$$J_\theta(x_i, y_i) = \begin{bmatrix} J_x(x_i) & | & J_y(y_i) \end{bmatrix}$$

where $J_x(x_i) \in \mathbb{R}^{p \times D}$ and $J_y(y_i) \in \mathbb{R}^{p \times D}$ are block matrices separating the terms related to X and those that are related to Y . The θ dependence is omitted here for notational simplicity. $J_x(x_i)$ can be written as

$$J_x(x_i) = \begin{bmatrix} \frac{\partial}{\partial x_{.1}} g_\theta(x_i, y_i)_1 & \cdots & \frac{\partial}{\partial x_{.D}} g_\theta(x_i, y_i)_1 \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_{.1}} g_\theta(x_i, y_i)_p & \cdots & \frac{\partial}{\partial x_{.D}} g_\theta(x_i, y_i)_p \end{bmatrix} \tag{6.75}$$

and $J_y(y_i)$ can be written as

$$J_y(y_i) = \begin{bmatrix} \frac{\partial}{\partial y_{.1}} g_\theta(x_i, y_i)_1 & \cdots & \frac{\partial}{\partial y_{.D}} g_\theta(x_i, y_i)_1 \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial y_{.1}} g_\theta(x_i, y_i)_p & \cdots & \frac{\partial}{\partial y_{.D}} g_\theta(x_i, y_i)_p \end{bmatrix} \tag{6.76}$$

where $g_\theta(x_i, y_i)_l$ denotes the l^{th} component of $g_\theta(x_i, y_i)$, i.e. $g_\theta(x_i, y_i)_l = k_\theta(x_i, z_l) - k_\theta(y_i, z_l)$ and $\frac{\partial}{\partial x_{.d}}$ is the partial derivative with respect to the d^{th} dimension of the random variable x .

For observations of D dimension, the Gaussian RBF kernel under consideration is of the form:

$$k_\theta(x_i, z_l) = \exp\left(-\frac{1}{2\theta^2}\|x_i - z_l\|_2^2\right) \quad (6.77)$$

$$= \exp\left(-\frac{1}{2\theta^2}[(x_{i1} - z_{l1})^2 + \cdots (x_{iD} - z_{lD})^2]\right) \quad (6.78)$$

Let the Gram matrix between $\{x_i\}_{i=1}^n$ and $\{z_j\}_{j=1}^p$ be K_{xz} such that $[K_{xz}]_{il} = k_\theta(x_i, z_l)$. Similarly, $[K_{yz}]_{il} = k_\theta(y_i, z_l)$.

For each dimension $d = 1, \dots, D$, we denote the difference between x_{id} and z_{jd} for $i = 1, \dots, n$ and $j = 1, \dots, p$ as $D_d^x \in \mathbb{R}^{n \times p}$:

$$D_d^x = \begin{bmatrix} x_{1d} - z_{1d} & \cdots & x_{1d} - z_{pd} \\ & \vdots & \\ x_{nd} - z_{1d} & \cdots & x_{nd} - z_{pd} \end{bmatrix}. \quad (6.79)$$

Similarly, we have D_d^y for $d = 1, \dots, D$.

Given the form of the kernel function, each term in (6.75) can be written as

$$[J_x(x_i)]_{lm} = \frac{\partial g_\theta(x_i, y_i)_l}{\partial x_{\cdot m}} = \frac{\partial k_\theta(x_i, z_l)}{\partial x_{\cdot m}} \quad (6.80)$$

$$= \exp\left(-\frac{1}{2\theta^2}[(x_{i1} - z_{l1})^2 + \cdots (x_{iD} - z_{lD})^2]\right) \left(-\frac{2}{2\theta^2}(x_{im} - z_{lm})\right) \quad (6.81)$$

$$= k_\theta(x_i, z_l) \left(-\frac{1}{\theta^2}(x_{im} - z_{lm})\right) \quad (6.82)$$

Similarly for the Jacobian term associated with Y :

$$[J_y(y_i)]_{lm} = \frac{\partial g_\theta(x_i, y_i)_l}{\partial y_{\cdot m}} = -\frac{\partial k_\theta(y_i, z_l)}{\partial y_{\cdot m}}.$$

In computing the probability of the observations $\{x_i, y_i\}_{i=1}^n$ given the hyperparameter θ , we are interested in computing the log determinant of the Jacobian term $\log \det(J_\theta(x_i, y_i)^\top J_\theta(x_i, y_i))$.

Note, for a fixed i , the square matrix can be written as

$$J_\theta(x_i, y_i)^\top J_\theta(x_i, y_i) = \left[\begin{array}{c|c} J_x(x_i)^\top J_x(x_i) & J_x(x_i)^\top J_y(y_i) \\ \hline J_y(y_i)^\top J_x(x_i) & J_y(y_i)^\top J_y(y_i) \end{array} \right] \quad (6.83)$$

where each $J(\cdot)_i^\top J(\cdot)_i$ is a block matrix of dimension $D \times D$. For notational convenience, we will omit the θ in the kernel function $k_\theta(\cdot, \cdot)$ when the context is clear.

$$J_x(x_i)^\top J_x(x_i) = \begin{bmatrix} \sum_{j=1}^p \frac{\partial k(x_i, z_j)}{\partial x_{\cdot 1}} \frac{\partial k(x_i, z_j)}{\partial x_{\cdot 1}} & \dots & \sum_{j=1}^p \frac{\partial k(x_i, z_j)}{\partial x_{\cdot 1}} \frac{\partial k(x_i, z_j)}{\partial x_{\cdot D}} \\ \vdots & & \vdots \\ \sum_{j=1}^p \frac{\partial k(x_i, z_j)}{\partial x_{\cdot D}} \frac{\partial k(x_i, z_j)}{\partial x_{\cdot 1}} & \dots & \sum_{j=1}^p \frac{\partial k(x_i, z_j)}{\partial x_{\cdot D}} \frac{\partial k(x_i, z_j)}{\partial x_{\cdot D}} \end{bmatrix} \quad (6.84)$$

The term $J_y(y_i)^\top J_y(y_i)$ is similar, while the cross-term can be computed as

$$J_x(x_i)^\top J_y(y_i) = \begin{bmatrix} -\sum_{j=1}^p \frac{\partial k(x_i, z_j)}{\partial x_{\cdot 1}} \frac{\partial k(y_i, z_j)}{\partial y_{\cdot 1}} & \dots & -\sum_{j=1}^p \frac{\partial k(x_i, z_j)}{\partial x_{\cdot 1}} \frac{\partial k(y_i, z_j)}{\partial y_{\cdot D}} \\ \vdots & & \vdots \\ -\sum_{j=1}^p \frac{\partial k(x_i, z_j)}{\partial x_{\cdot D}} \frac{\partial k(y_i, z_j)}{\partial y_{\cdot 1}} & \dots & -\sum_{j=1}^p \frac{\partial k(x_i, z_j)}{\partial x_{\cdot D}} \frac{\partial k(y_i, z_j)}{\partial y_{\cdot D}} \end{bmatrix} \quad (6.85)$$

with

$$\begin{aligned} & [J_x(x_i)^\top J_x(x_i)]_{st} \\ &= \sum_{j=1}^p \frac{\partial k(x_i, z_j)}{\partial x_{\cdot s}} \frac{\partial k(x_i, z_j)}{\partial x_{\cdot t}} \end{aligned} \quad (6.86)$$

$$= \sum_{j=1}^p k_\theta(x_i, z_j) \left(-\frac{1}{\theta^2} (x_{is} - z_{js}) \right) k_\theta(x_i, z_j) \left(-\frac{1}{\theta^2} (x_{it} - z_{jt}) \right) \quad (6.87)$$

$$= \sum_{j=1}^p \frac{1}{\theta^4} k_\theta(x_i, z_j)^2 (x_{is} - z_{js})(x_{it} - z_{jt}) \quad (6.88)$$

$$= \sum_{j=1}^p \frac{1}{\theta^4} [K_{xz}]_{ij}^2 [D_s^x]_{ij} [D_t^x]_{ij}. \quad (6.89)$$

Essentially, the above equation indicates that we compute the row sum of the elements-wise product of these three matrices of interest. Similarly,

$$[J_x(x_i)^\top J_y(y_i)]_{st} = - \sum_{j=1}^p \frac{\partial k(x_i, z_j)}{\partial x_{\cdot s}} \frac{\partial k(y_i, z_j)}{\partial y_{\cdot t}} \quad (6.90)$$

$$= - \sum_{j=1}^p \frac{1}{\theta^4} [K_{xz}]_{ij} [K_{yz}]_{ij} [D_s^x]_{ij} [D_t^y]_{ij} \quad (6.91)$$

Equations (6.89) and (6.91) enable the computation of the square matrix $J_\theta(x_i, y_i)^\top J_\theta(x_i, y_i)$ for each $i = 1, \dots, n$.

6.4 Null Model: $p(\theta|M=0, \mathcal{D})$

In this section, we derive the marginal pseudolikelihood under the null model. The term “pseudolikelihood” is used since it relies on the evaluation of the empirical

embedding at a finite set of inducing points and hence it is an approximation to the likelihood of the infinite dimensional empirical embedding [Flaxman et al., 2016]. The model simplifies significantly as $\delta = 0$, we propose to model the data directly with a Gaussian noise model

$$\Delta|\theta \sim \mathcal{N}(0, \frac{1}{n}\Sigma_\theta) \quad (6.92)$$

where as before, we rewrite the covariance matrix as $\frac{1}{n}\Sigma_\theta$. By a similar argument as in Section 6.3, we obtain

$$g_\theta(x_i, y_i)|\theta \sim \mathcal{N}(0, \Sigma_\theta) \quad (6.93)$$

for $i = 1, \dots, n$. This implies that the marginal pseudolikelihood can be written as

$$p(\{x_i, y_i\}_{i=1}^n|\theta) = \prod_{i=1}^n p(g_\theta(x_i, y_i)|\theta) \text{vol}(J_\theta(x_i, y_i)) \quad (6.94)$$

$$= \mathcal{N}(\text{vec}(G_\theta); 0, I_n \otimes \Sigma_\theta) \prod_{i=1}^n \text{vol}(J_\theta(x_i, y_i)). \quad (6.95)$$

Expanding out the multivariate Gaussian distribution and taking the logarithm

$$\begin{aligned} & \log p(\{x_i, y_i\}_{i=1}^n|\theta) \\ & \propto -0.5 \log \det(I_n \otimes \Sigma_\theta) - 0.5 \text{vec}(G_\theta)^\top (I_n \otimes \Sigma_\theta)^{-1} \text{vec}(G_\theta) + \sum_{i=1}^n \log \text{vol}(J_\theta(x_i, y_i)). \end{aligned} \quad (6.96)$$

A naive computation of the above again incur a prohibitive computational burden that scales like $\mathcal{O}(m^3 n^3)$. We mimic the case of the alternative model (Proposition 6.3.1 from Section 6.3.3) and derive the efficient computation of the marginal pseudolikelihood.

Lemma 6.4.1. $\text{vec}(G_\theta)^\top (I_n \otimes \Sigma_\theta)^{-1} \text{vec}(G_\theta) = \text{Tr}(G^\top \Sigma^{-1} G)$.

Proof. By properties of matrix inversion:

$$(I_n \otimes \Sigma_\theta)^{-1} = I_n^{-1} \otimes \Sigma_\theta^{-1} = I_n \otimes \Sigma_\theta^{-1}.$$

Note that $(I_n \otimes \Sigma_\theta^{-1}) \text{vec}(G) = \text{vec}(\Sigma_\theta^{-1} G)$, we have

$$\text{vec}(G)^\top (I_n \otimes \Sigma_\theta^{-1}) \text{vec}(G) = \text{vec}(G)^\top \text{vec}(\Sigma_\theta^{-1} G) = \text{Tr}(G^\top \Sigma_\theta^{-1} G).$$

□

For the log determinant term, $\det(I_n \otimes \Sigma_\theta) = \det(I_n)^m \otimes \det(\Sigma_\theta)^n$ by properties of matrix determinant. Since the $\det(I_n)^m = 1$, then the marginal pseudolikelihood can be written as

$$\begin{aligned} & \log p(\{x_i, y_i\}_{i=1}^n | \theta) \\ & \propto -0.5n \log(\det(\Sigma_\theta)) - 0.5 \text{Tr}(\Sigma_\theta^{-1} G G^\top) + \sum_{i=1}^n \log \text{vol}(J_\theta(x_i, y_i)) \end{aligned} \quad (6.97)$$

6.5 Posterior Inference

6.5.1 Bayes Factor with Fixed θ

Given the framework set out before and assume equal prior probabilities for the null and alternative model, when the lengthscale parameter θ is fixed, the Bayes factor can be computed directly by utilizing Gaussian distributions of Δ as

$$BF = \frac{P(\Delta | M = 0, \theta)}{P(\Delta | M = 1, \theta)} \quad (6.98)$$

$$\Delta | M = 0, \theta \sim N(0, \frac{1}{n} \Sigma_\theta) \quad (6.99)$$

$$\Delta | M = 1, \theta \sim N(0, R_\theta + \frac{1}{n} \Sigma_\theta). \quad (6.100)$$

We consider some simple 1 dimensional experiments where under the null, both samples come from $N(0, 1)$ and under the alternative, the data sample $\{y_i\}_{i=1}^n$ come from distributions listed in Table.6.1 and the resulting Bayes factor is given. We fix

Alternative Model	Bayes Factor	$P(M = 0 \Delta, \theta)$
$\mathcal{N}(0, 1)$	2.60×10^{25}	1.00
$\mathcal{N}(0, 2^2)$	2.85×10^{-04}	2.85×10^{-04}
$\mathcal{N}(0, 3^2)$	2.98×10^{-48}	2.98×10^{-48}
$\mathcal{N}(1, 1)$	6.10×10^{-17}	6.10×10^{-17}
$\mathcal{N}(2, 1)$	9.52×10^{-175}	9.52×10^{-175}
$\mathcal{N}(3, 1)$	0	0
$Laplace(0, 1)$	4.94×10^{17}	1.00

Table 6.1: Table of Bayes factors. Recall a Bayes factor with a value greater than 1 is supporting the null hypothesis while a value smaller than 1 is supporting the alternative hypothesis. The larger the value of the Bayes factor the stronger the support it is for the null [Jeffreys, 1961].

the θ parameter to the median heuristic. The results were obtained using 500 samples with 40 evaluation points and averaged across 100 simulations. Equivalently, by the

relationship between the posterior probability of $M|\Delta = 0$ and the BF given in (6.6), the model clearly recover the ground truth for distinguishing Gaussian distributions that differ in mean or variance. However, when comparing the null distribution with Laplace distribution, the model is struggling to detect the difference at 500 samples. Although not presented here, the Bayes factor approach as above is struggling to distinguish between $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, 4)$ for a smaller sample size of 400 or less. Also, we noticed that when increasing the number of samples to approximately 2500, the Bayes factor is again able to distinguish between the Laplace distribution and the standard Gaussian distribution. On the other hand, for the number of evaluation points in $\{30, 40, 50, 60, 80\}$, the results do not exhibit noticeable difference.

6.5.2 Bayes Factor with Inferred θ

In the previous section, the lengthscale parameter θ was fixed, and the computation of the posterior distribution of $M|\mathcal{D}$ was straightforward. In this section, we propose to use a Metropolis Hasting within Gibbs type of approach for the joint posterior inference of M and θ . Iteratively, we would like to sample from $p(\theta|M, \mathcal{D})$ and $p(M|\theta, \mathcal{D})$.

To sample from $p(\theta|M, \mathcal{D})$, recall the marginal pseudolikelihood of the alternative model (6.74) as a function of θ from Section 6.3:

$$\begin{aligned} \log p(\{x_i, y_i\}_{i=1}^n|\theta) = & \text{const} - \frac{1}{2} \log \det(\Sigma_\theta + nR_\theta) - \frac{1}{2}(n-1) \log \det(\Sigma_\theta) \\ & - \frac{1}{2} \text{Tr} \left((\Sigma_\theta + nR_\theta)^{-1} G_\theta G_\theta^\top + \left(\frac{1}{n} \Sigma_\theta R_\theta^{-1} \Sigma_\theta + \Sigma_\theta \right)^{-1} G_\theta H G_\theta^\top \right) \\ & + \frac{1}{2} \sum_{i=1}^n \log \det(J_\theta(x_i, y_i)^\top J_\theta(x_i, y_i)) \end{aligned}$$

and the null model (6.97) from Section 6.4 :

$$\log p(\{x_i, y_i\}_{i=1}^n|\theta) \propto -0.5n \log(\det(\Sigma_\theta)) - 0.5 \text{Tr}(\Sigma_\theta^{-1} G G^\top) + \sum_{i=1}^n \log \text{vol}(J_\theta(x_i, y_i)).$$

Hence, we can evaluate $\log p(\theta|\{(x_i, y_i)\}_{i=1}^n, M)$ by adding the log prior distribution $\text{Gamma}(2, 2)$. To sample from $p(M|\theta, \mathcal{D})$, recall the relationship between Bayes factor and the posterior distribution of the null (6.6) and alternative model (6.7) respectively

$$\begin{aligned} p(M = 1|\theta, \mathcal{D}) &= \frac{1}{1 + BF_\theta}, \\ p(M = 0|\theta, \mathcal{D}) &= \frac{BF_\theta}{1 + BF_\theta}. \end{aligned}$$

We index BF by θ to emphasize the dependence of the Bayes factor on the kernel hyperparameter θ . Equivalently, this says $M|\theta, \mathcal{D} \sim \text{Bernoulli}\left(\frac{1}{1+BF_\theta}\right)$.

6.5.3 The Procedure for Posterior Inference

We present pseudocode of our posterior inference procedure in Algorithm 1. We observe from our experiments, that increasing the number of HMC steps inside Gibbs $\tilde{n}$ improves the posterior convergence of the chain. Due to the computational resources available, $\tilde{n}$ has only been set to values up to 30. The posterior marginal probability $P(\theta|\mathcal{D})$ is approximated by the posterior MCMC samples $\{\theta_1, \dots, \theta_{\tilde{n}}\}$. On the other hand, the posterior marginal probability $P(M = 1|\mathcal{D})$ is the proportion of $M = 1$ in all the samples $\{M_0, \dots, M_{\tilde{n}}\}$.

Data: A paired sample $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$; The number of inducing points m ;
The number of simulation $\tilde{m}$; The number of HMC steps $\tilde{n}$.
Output: A sample $\{\theta_i, M_i\}_{i=1}^{\tilde{m}}$ from the posterior distribution of $p(\theta, M|\mathcal{D})$.
1 Initialise $\theta_0 = \text{Median heuristic on the set } \{x_1, \dots, x_n, y_1, \dots, y_n\}$;
2 Initialise M_0 by computing the Bayes factor (BF) in (6.98) and sample $\text{Bernoulli}\left(\frac{1}{1+BF_\theta}\right)$;
3 **for** $i \leftarrow 1$ **to** $\tilde{n}$ **do**
4 Simulate a chain $\{\theta_1, \dots, \theta_{\tilde{n}}\}$ from $p(\theta|M_{i-1}, \mathcal{D})$ using NUTS in Stan;
5 Set $\theta_i = \theta_{\tilde{n}}$;
6 Simulate $M_i \sim \text{Bernoulli}\left(\frac{1}{1+BF_\theta}\right)$;
7 **end**

Algorithm 1: Posterior Inference of the lengthscale parameter θ of the Gaussian RBF kernel and the model indicator M .

6.6 Experiments

In this section, we would like to investigate the performance of the proposed posterior inference scheme for M and θ on synthetic data experiments. For each of the synthetic experiments, we generate 100 independent data sets of size n for each of the models. We examine the distribution of the probability of the alternative hypothesis (i.e. $p(M = 1|\mathcal{D})$) while varying the number of observed data points n . The number of evaluation points is fixed at $p = 40$, with half sampled from the distribution of X and the other half sampled from the distribution of Y .

For the posterior sampling, we run the algorithm for $\tilde{m} = 2000$. The initial 500 samples are discarded as burn-in and the thinning factor is 2. For every Gibbs sampling step, we take 9 steps in HMC which contains 3 warmup steps for step size adaption. Note, we have experimented with 1 HMC step for every step of Gibbs sampling, the convergence of the parameters M and θ is much slower in that case. On the other hand, increasing the number of HMC steps beyond 9 does not seem to improve the performance by a significant amount. We used 9 steps for a balance between computational complexity and performance.

6.6.1 Simple 1 Dimensional Gaussian Distributions

This section investigates the same experiment as Section 6.5.1 while inferring both M and θ . We refer to Table.6.1 for details of the null and alternative models. As expected, when the samples X and Y come from the same distribution (i.e. first plot of Figure.6.1), the probability of $M = 1$ is consistently zero across all sample sizes. When there is a difference in the variance of the Gaussian distribution, for example when the alternative model is $\mathcal{N}(0, 4)$, the probability of $M = 1$ is zero when 50 or 100 samples are given. As we increase the sample size to 300, we observe that the test is uncertain since probability of $M = 1$ varying from 0 to approximately 0.95 with median at around 0.15. However, as we increase the sample size to 500, the probability of $M = 1$ is more concentrated towards 1. This phenomena is observed for all the alternative models considered.

Essentially, when the number of samples is small, there is not enough evidence to determine if the null hypothesis should be rejected. In such a case, the Bayes factor favours the simpler null hypothesis. This is to be expected since Bayesian modelling encompasses a natural Occam factor in the prior predictive [MacKay, 2003, Chapter 28] Filippi and Holmes [2016]. We will see this reflected in all the synthetic experiments in this section.

Note, for a Gaussian distribution, the difference in mean is easier to detect comparing to a difference in variance. This is reflected in the experiments: we observe that the probability of the alternative hypothesis becomes very close to 1 at a much smaller sample size for the experiments with a difference in mean.

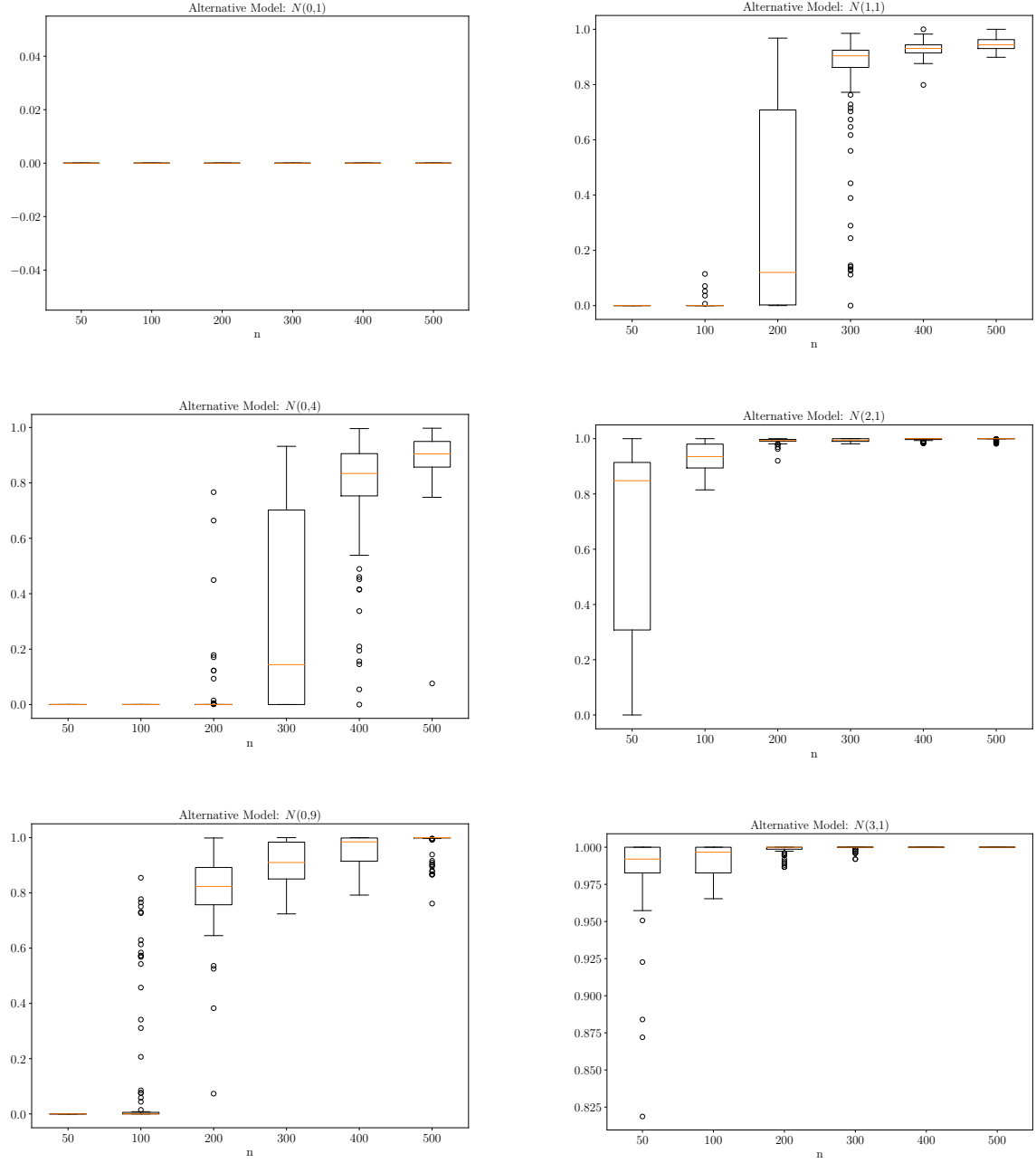


Figure 6.1: 1-dimensional Gaussian experiment. The distribution of the probability of the alternative hypothesis $p(M=1|\mathcal{D})$ for a different number of observations n .

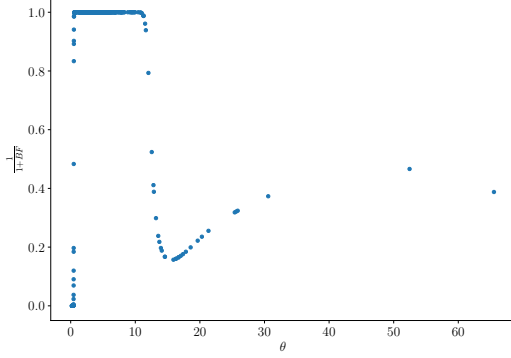


Figure 6.2: 1-dimensional Gaussian experiment with the alternative model $\mathcal{N}(0, 9)$ and 200 samples. The plot illustrates $\frac{1}{1+BF_\theta}$ as a function of θ .

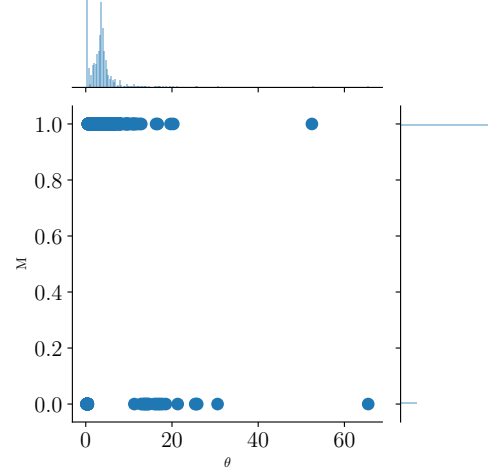


Figure 6.3: Joint distribution of M and θ given observed data $\mathcal{D}$. The histogram of posterior samples for θ is given on the top. The histogram of posterior samples for M is given on the right.

We emphasize that, unlike the frequentist kernel two-sample test where a single value of the lengthscale parameter needs to be predetermined, the proposed Bayesian framework integrates over all possible θ values and alleviates the need for kernel lengthscale selection. However, some θ values are more informative in distinguishing the difference between the two distributions while others are less informative. As a specific example, we consider the case when $X \sim \mathcal{N}(0, 1)$ and $Y \sim \mathcal{N}(0, 9)$ with 200 samples each. For this specific simulation, we obtain $P(M = 1|\mathcal{D}) \approx 0.839$. Figure 6.2 illustrates the change of the probability of $M = 1|\theta, \mathcal{D}$ as a function of θ . Clearly, the region of θ from approximately 0.05 to 11 is most informative for distinguishing these two distributions. This is also reflected in the marginal distribution of $\theta|M = 1$ and $\theta|M = 0$ from Figure 6.4 and Figure 6.3. Rather than selecting a single lengthscale parameter, the proposed method is able to highlight the range of informative lengthscales. As we will see, this is more useful in the cases when multiple lengthscale parameters are of interest for a single testing problem.

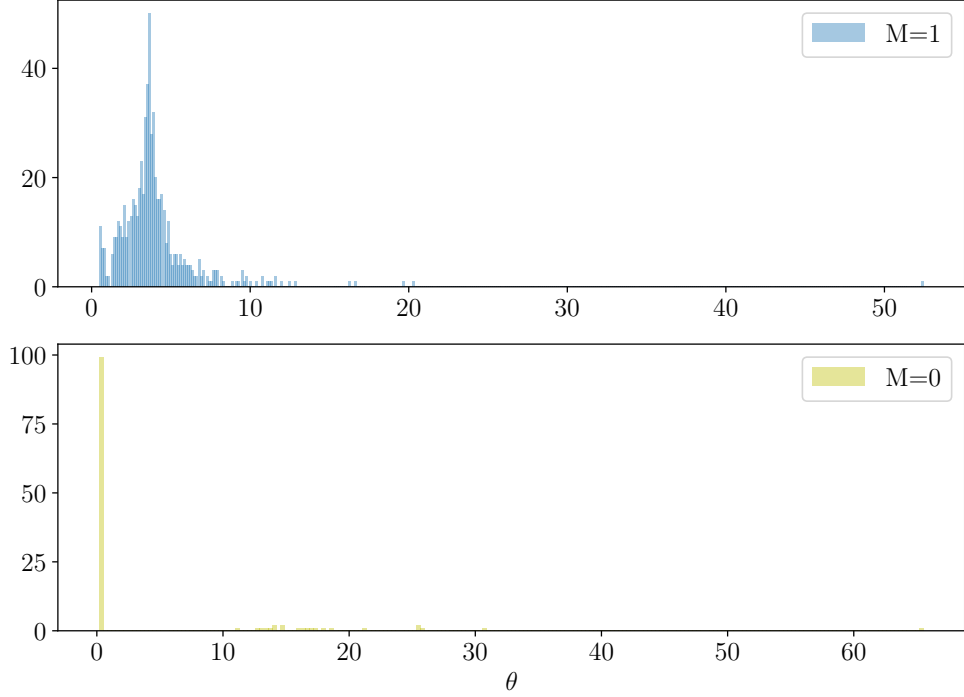


Figure 6.4: 1-dimensional Gaussian distribution. Top: The marginal distribution of $\theta|M=1, \mathcal{D}$. Bottom: The marginal distribution of $\theta|M=0, \mathcal{D}$.

6.6.2 2 Dimensional Gaussian Distributions

Before investigate examples where multiple lengthscales are of interest, we first consider the Gaussian distributions in 2 dimensions with

$$X \sim \mathcal{N}\left(\begin{pmatrix} 10 \\ 10 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \quad (6.101)$$

$$Y \sim \mathcal{N}(M_y, QS_\epsilon Q^\top) \quad (6.102)$$

where $M_y \in \mathbb{R}^2$ is the mean vector, Q is the rotation matrix

$$Q = \begin{pmatrix} \cos(\frac{\pi}{2}) & \sin(\frac{\pi}{2}) \\ -\sin(\frac{\pi}{2}) & \cos(\frac{\pi}{2}) \end{pmatrix} \quad (6.103)$$

and S_ϵ is the diagonal matrix of the form

$$S_\epsilon = \begin{pmatrix} \epsilon & 0 \\ 0 & 1 \end{pmatrix}. \quad (6.104)$$

For the ease of notation, we denote $\Sigma(\epsilon) := QS_\epsilon Q^\top$ highlighting the dependency on the parameter ϵ . More specifically, we consider the following parameter settings for the simulated data as shown in Table 6.2. Under the null hypothesis, we simulated both X and Y independently from a 2 dimensional Gaussian distribution with

	Null	Alternative					
$M_y^\top$	(10, 10)	(11.5, 11.5)	(12, 10)	(10, 10)	(10, 10)	(10, 10)	(10, 10)
ϵ	1	1	1	2	6	10	20

Table 6.2: The different parameters for the 2-dimensional Gaussian distribution experiment.

$M_y = (10, 10)^\top$ and $\epsilon = 1$. As expected, the distribution of the probability of $M = 1$ is consistently zero for all sample sizes as shown in Figure 6.12. Under the alternative hypothesis, as expected, we observe in Figure 6.5 that the probability distribution of the alternative hypothesis become concentrated around 1 as the number of samples increase. In other words, as more samples are seen, the Bayes factor is able to favor the alternative hypothesis with more certainty.

Note, for the case when $\epsilon = 2$, the difference between the two distributions is not detected even at the maximum sample size considered ($n = 800$) and the Bayes factor consistently favours the null hypothesis. Since more samples are needed for the model to be able to detect the difference between the two distributions, we argue that large scale approximations such as random Fourier feature extension of the proposed method is needed. We will provide a brief discussion of this in Section 6.7.

6.6.3 1 Dimensional Mixture of Gaussian Distributions

We have seen that the proposed Bayesian two-sample testing method works well in distinguishing 1 and 2 dimensional Gaussian distributions. In this section, we consider 1 dimensional mixture of Gaussian distributions as our null and alternative models. More specifically, we consider the following list of null and alternative models:

- Null Model: $X \sim 0.5\mathcal{N}(0, 1) + 0.5\mathcal{N}(4, 1)$
- Alternative Model: $Y \sim 0.5\mathcal{N}(0, 4) + 0.5\mathcal{N}(4, 4)$
- Alternative Model: $Y \sim 0.5\mathcal{N}(0, 1) + 0.5\mathcal{N}(8, 1)$
- Alternative Model: $Y \sim 0.5\mathcal{N}(0, 4) + 0.5\mathcal{N}(8, 4)$
- Alternative Model: $Y \sim 0.5\mathcal{N}(2, 1) + 0.5\mathcal{N}(6, 1)$
- Alternative Model: $Y \sim 0.5\mathcal{N}(2, 4) + 0.5\mathcal{N}(8, 4)$
- Alternative Model: $Y \sim 0.5\mathcal{N}(-4, 1) + 0.5\mathcal{N}(8, 1)$

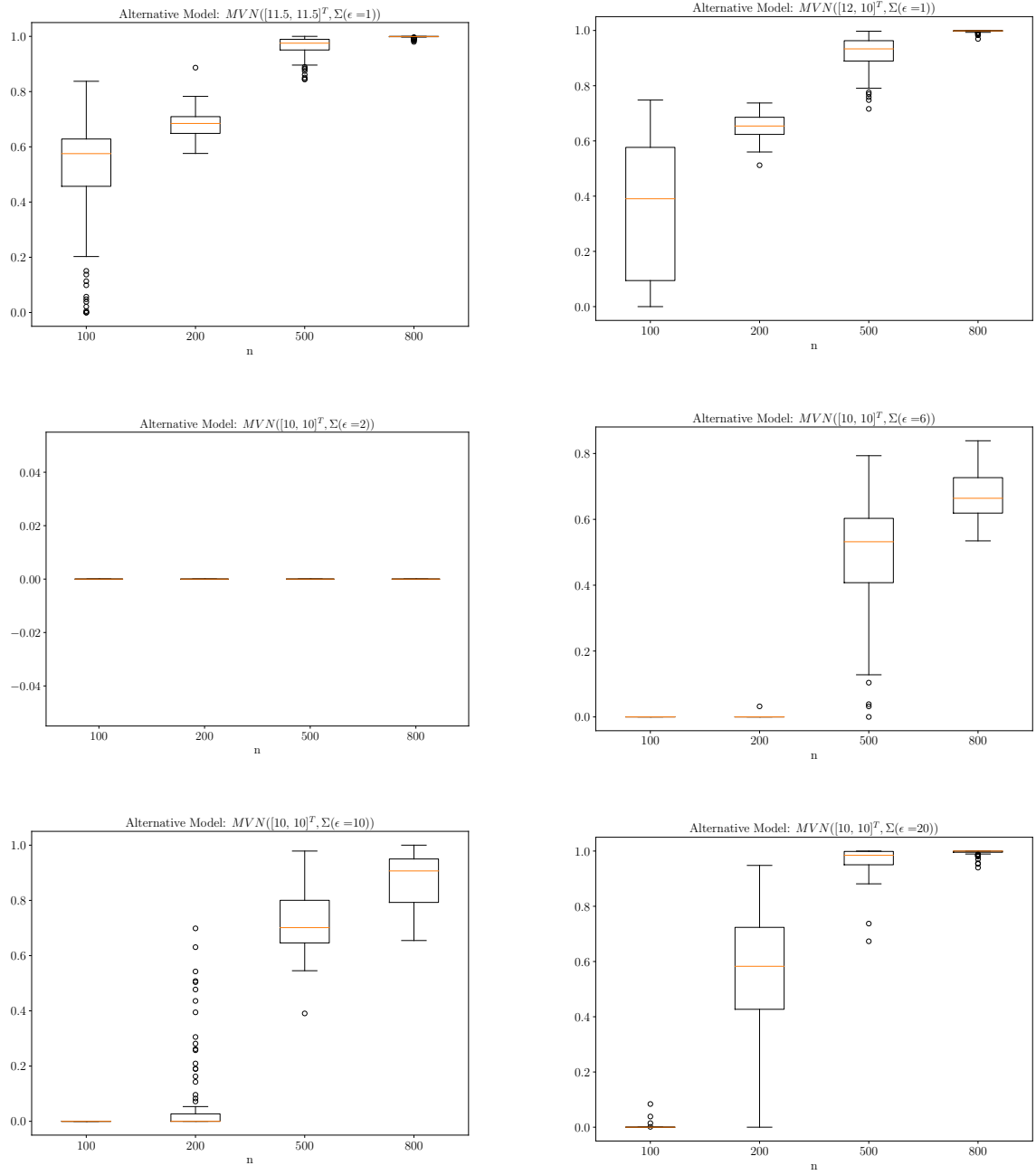


Figure 6.5: 2-dimensional Gaussian distributions. The distribution of the probability of the alternative hypothesis $p(M=1|\mathcal{D})$ for a different number of observations n .

- Alternative Model: $Y \sim 0.5\mathcal{N}(-4, 4) + 0.5\mathcal{N}(8, 4)$.

Figure.6.13 of Appendix 6.8 provides illustration of samples from these models. For each of the models, Figure.6.6 shows the distribution of $M = 1|\mathcal{D}$ as a function of the number of samples n . For the first plot of Figure.6.6, we observe that the probability consistently equals to zero for all sample sizes and for all simulated datasets. This is reassuring as the data X and Y are indeed simulated from the same distribution, i.e. we are under the null hypothesis. When the alternative model is $Y \sim 0.5\mathcal{N}(0, 4) + 0.5\mathcal{N}(4, 4)$, the problem is more challenging. At 800 samples, the test is uncertain, but preferring the null hypothesis. This comes back to the aforementioned idea of Bayesian modelling naturally prefers simple hypothesis. In the other plots of Figure 6.6, we observed that the distribution of $M = 1|\mathcal{D}$ gradually concentrates around 1 as the number of samples increases. This aligns with our expectations.

6.6.4 2 by 2 Blobs of 2 Dimensional Gaussian Distributions

The performance of the frequentist kernel two-sample test using MMD depends heavily on the choice of kernel. When a Gaussian kernel is used, this boils down to choosing an appropriate lengthscale parameter. Often, median heuristic is used. However, Gretton et al. [2012b] showed that MMD with median heuristic failed to reject the null hypothesis when comparing samples from a grid of isotropic Gaussian v.s. a grid of non-isotropic Gaussian. The framework proposed by Flaxman et al. [2016] showed, by choosing the lengthscale that optimise the Bayesian kernel learning marginal log likelihood (i.e. an empirical Bayes type of approach), MMD is able to correctly reject the null hypothesis at the desired significance level. Intuitively, the algorithm needs to look locally at each blob to detect the difference rather than at the lengthscale that covers all of the blobs which is given by the median distance between points.

We repeat this experiment using the proposed Bayesian two-sample test with P_X being a mixture of 2 dimensional isotropic Gaussian distributions and P_Y a mixture of 2 dimensional Gaussian distributions centered at slightly shifted locations with rotated covariance matrix. Note, this is not the same dataset used in Flaxman et al. [2016] and Gretton et al. [2012b], we shift the dataset to have multiple lengthscales relevant. We center the blobs of 2 dimensional Gaussian distributions of P_X at $\{(10, 10)^\top, (10, 20)^\top, (20, 10)^\top, (20, 20)^\top\}$ and shift such locations by $(-1, -1)$ for P_Y . An equal number of observations is sampled from each of the blobs. The covariance

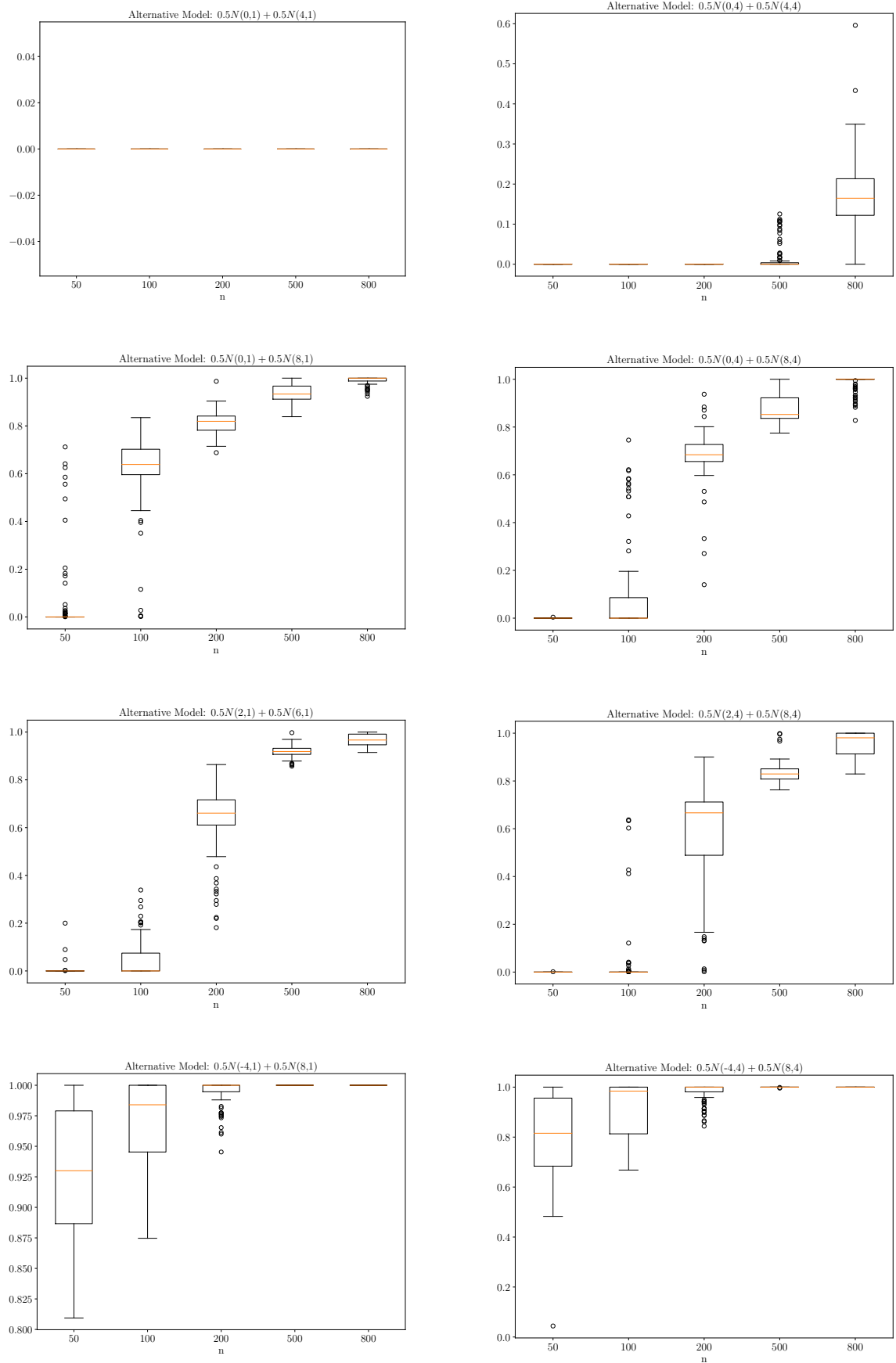


Figure 6.6: 1 dimensional mixture of Gaussian distributions. The distribution of the probability of the alternative hypothesis $p(M = 1|\mathcal{D})$ for a different number of observations n .

matrix of P_Y follows the same form as in Section 6.6.2 with $\epsilon = \{2, 6, 10, 20\}$. Figure 6.15 illustrates samples from these distributions.

As an example, in Figure 6.8, we see that a wide range of lengthscales is informative for this two-sample testing problem when $\epsilon = 2$. If we further observe the marginal distributions $P(\theta|M = 0, \mathcal{D})$ and $P(\theta|M = 1, \mathcal{D})$ in Figure 6.10, the method takes advantage of the large lengthscales to detect shift in location and the small lengthscales to detect the difference in covariance. But when the lengthscale is too small (approximately less than 0.5), the method regards the samples as identically distributed.

Figure 6.7 shows that our approach is able to detect the difference between the distributions since the probability of the alternative hypothesis becomes more concentrated around 1 as the number of samples increases. Note, when $\epsilon = 2$, the distribution of the probability of the alternative hypothesis is around 0.6. We expect this to increase to 1 as we increase the number of samples to around 300 samples per blob given the pattern observed.

6.6.5 Higher Dimensional Gaussian Distributions

We have seen that the proposed method is able to utilise informative value of the lengthscale parameter and make correct decisions about the probability of the alternative hypothesis given large enough samples. In this section, we investigate the effect of dimensionality of the problem. We use the Gaussian blobs experiment from the previous section and append simple $\mathcal{N}(0, 1)$ to both X and Y (i.e. the difference exist only in the first two dimensions). In particular, we consider the cases when the total number of dimensions are 3, 4 and 5 due to the availability of computational resources. The results are presented in Figure 6.11.

Comparing to the plot of $\epsilon = 6$ from Figure 6.7 where no noise is added, the performance at 25 samples per blobs is very similar. At 75 samples per blobs, the posterior distribution is concentrated at 0 insisting the probability of the alternative hypothesis is zero. This is more conservative comparing to the $\epsilon = 6$ case of Figure 6.7 where the median of the distribution is around 0.4. At 125 samples per blob, the variance of the distribution of $\frac{1}{1+BF_\theta}$ increases and the median of the distribution of $\frac{1}{1+BF_\theta}$ decreases as the number of dimension increase from 2 to 4. When the total

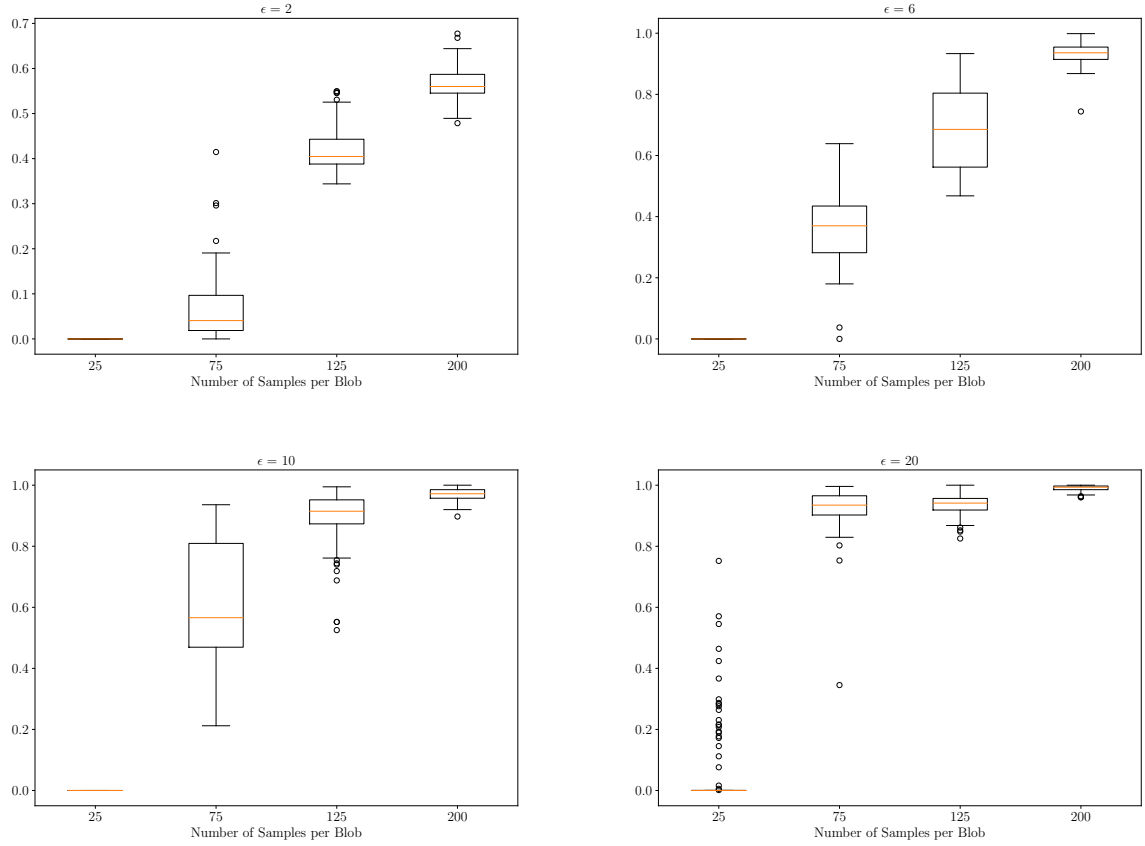


Figure 6.7: 2 by 2 blobs of bivariate Gaussian distributions. The distribution of the probability of the alternative hypothesis $p(M = 1|\mathcal{D})$ for a different number of observations.

number of dimensions is 5, although the median of the distribution is similar to when the dimension is 4, we see the first and third quantiles of the distribution shifting downwards. At 200 samples per blob, we observe a similar performance as Figure 6.7. However, the medians of the distribution are slightly lower in 4 and 5 dimensions. The noise in the additional dimensions has indeed made the problem harder for the given number of samples. But the proposed method still manages to discover the difference as the number of samples increases. This illustrates the robustness of our proposed method to the dimensionality of the problem.

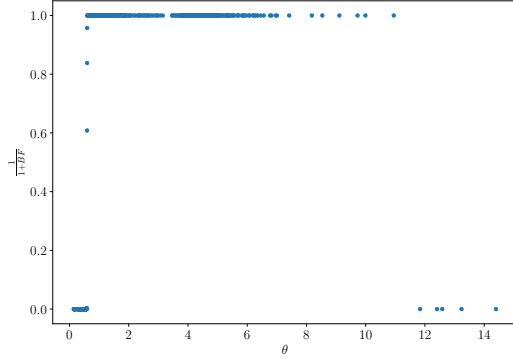


Figure 6.8: 2 by 2 blobs of bivariate Gaussian experiment with the alternative model $\epsilon = 2$ and 200 samples per blob. The plot illustrates $\frac{1}{1+BF_\theta}$ against the value of θ .

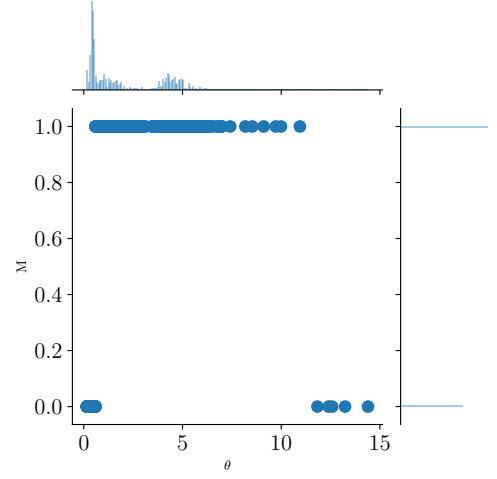


Figure 6.9: Joint distribution of M and θ given $\mathcal{D}$. The histogram of posterior samples for θ is given on the top. The histogram of posterior samples for M is given on the right.

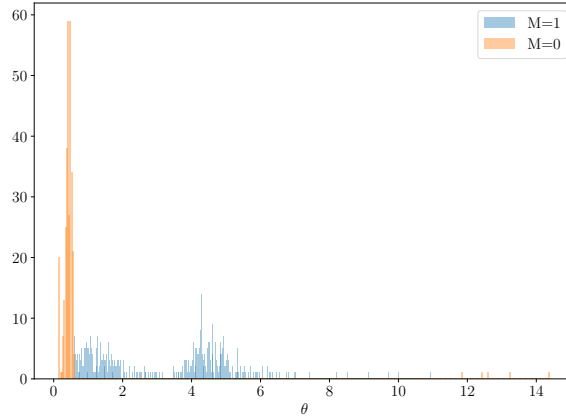


Figure 6.10: 2 by 2 blobs of bivariate Gaussian experiment with the alternative model $\epsilon = 2$ and 200 samples per blob. Blue: Histogram of samples from the marginal distribution of $\theta|M=1, \mathcal{D}$. Orange: Histogram of samples from the marginal distribution of $\theta|M=0, \mathcal{D}$.

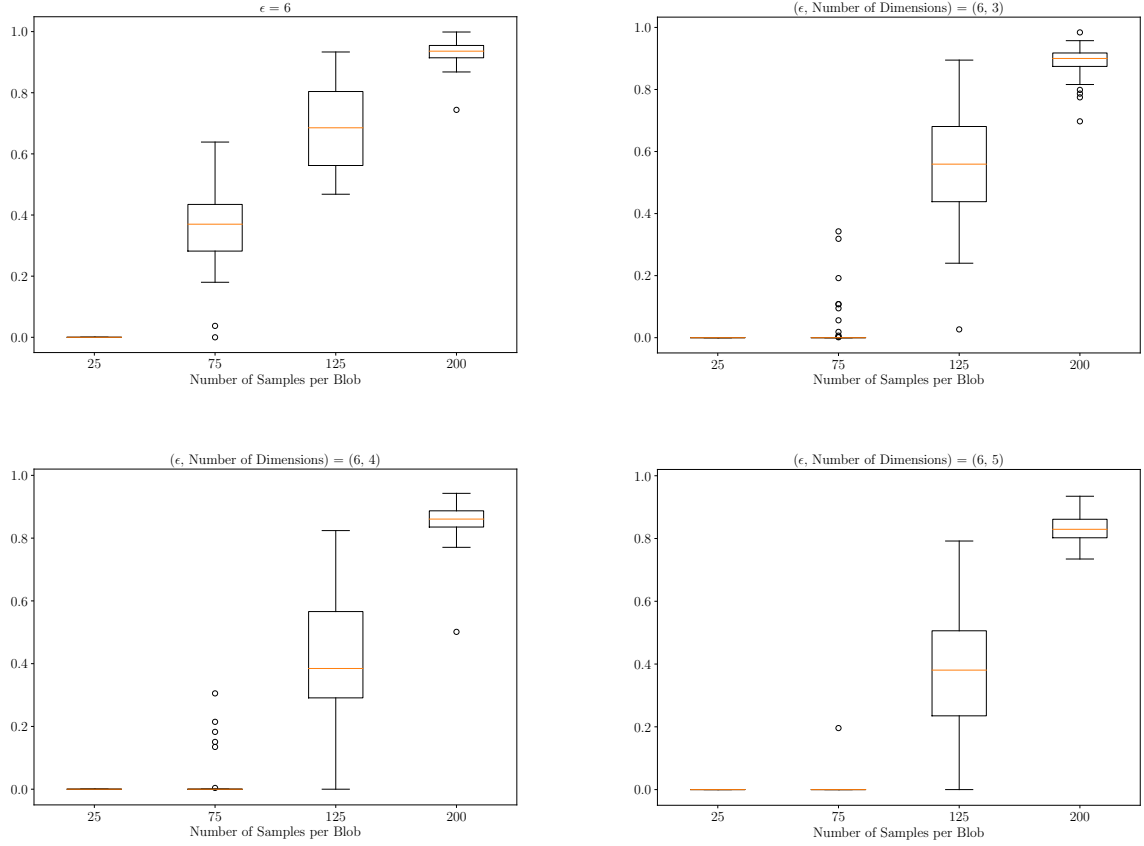


Figure 6.11: $\epsilon = 6$: Distribution of the probability of $M = 1|\mathcal{D}$ for 2 by 2 blobs of bivariate Gaussian distributions with standard Gaussian noise in the remaining dimensions for both samples. Top left figure is copied from Figure 6.7 for the ease of performance comparison.

6.7 Conclusion

In this chapter, we have proposed a Bayesian two-sample testing framework utilising the Bayes factor. Rather than directly considering the observations, we have proposed to consider the differences between the empirical kernel mean embeddings (KME) evaluated at a set of inducing points. Following the learning procedure of the empirical KME [Flaxman et al., 2016], we have derived the Bayes factors when the kernel hyperparameter is inferred and when it is given as well as when it is treated in a fully Bayesian way and marginalised over. Further, we have obtained efficient computation methods of the marginal pseudolikelihood through the Kronecker structure of the covariance matrices. The posterior inference of the model label and the kernel hyperparameter is done by HMC within Gibbs. We have showed in a range of synthetic experiments that our proposed Bayesian test is able to simultaneously utilis-

ing multiple lengthscales and correctly uncover the ground truth given sufficient data.

There are many possible future developments following this work. We have used the Bayes factor formulation for Bayesian model selection as the base framework of our two-sample test. Alternatively, we can adopt the mixture approach of Kamary et al. [2014] and it would be interesting to see the difference in performance between these two methods. As we have seen from the synthetic experiments, a larger sample size is required for the more challenging problems. A random Fourier feature [Rahimi and Recht, 2007] extension of the above framework can be easily developed to enable larger samples to be used. Since explicit feature maps are available, the difference between feature maps $\phi(x_i) - \phi(y_i)$ can be modelled directly with a Gaussian distribution with mean $\mathbb{E}(\phi(X)) - \mathbb{E}(\phi(Y))$ and an estimated covariance matrix. The rest of the inference procedure would follow similarly as presented here. The framework developed is not restrictive to the two-sample testing only, it can be modified for independence testing with the respective RFF extensions.

6.8 Appendix

Null Case of 2-Dimensional Gaussian Distributions

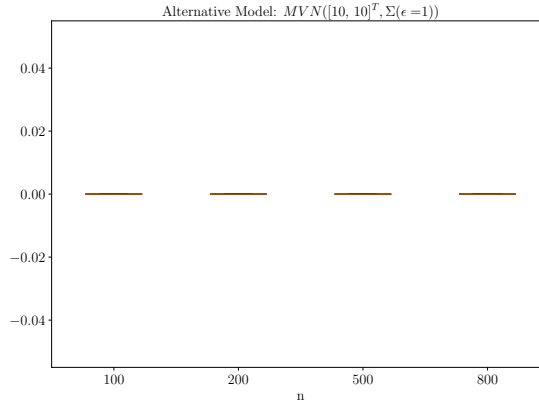


Figure 6.12: Null case of 2-dimensional Gaussian distributions. The distribution of the probability of the alternative hypothesis $p(M = 1|\mathcal{D})$ for a different number of observations n . The results align with our expectation, i.e. the probability of the alternative hypothesis is zero when both samples are simulated from the same distribution.

Data Visualisations

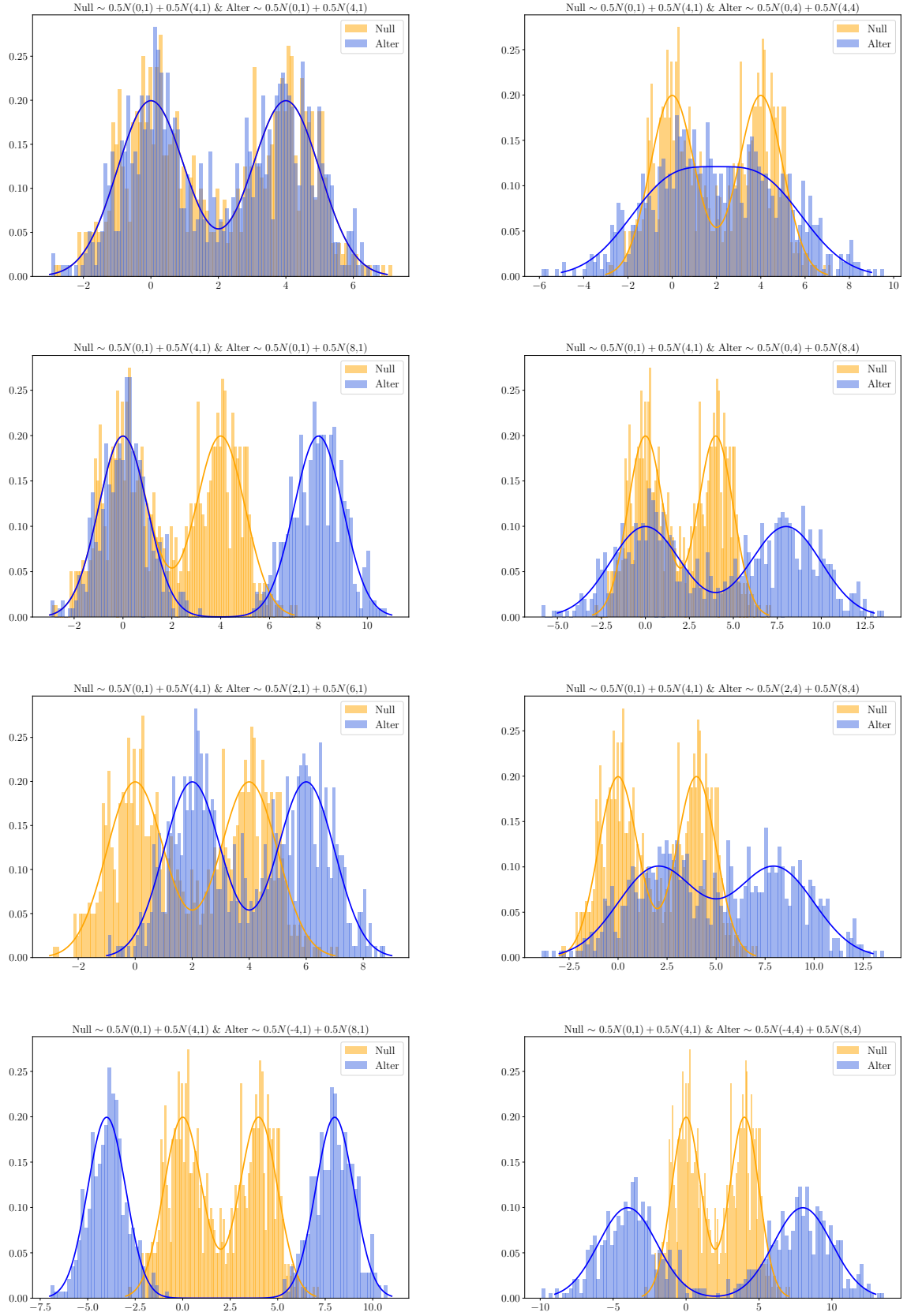


Figure 6.13: Visualisation of 1-dimensional mixture of Gaussian distributions from Section 6.6.3 with 800 samples each.

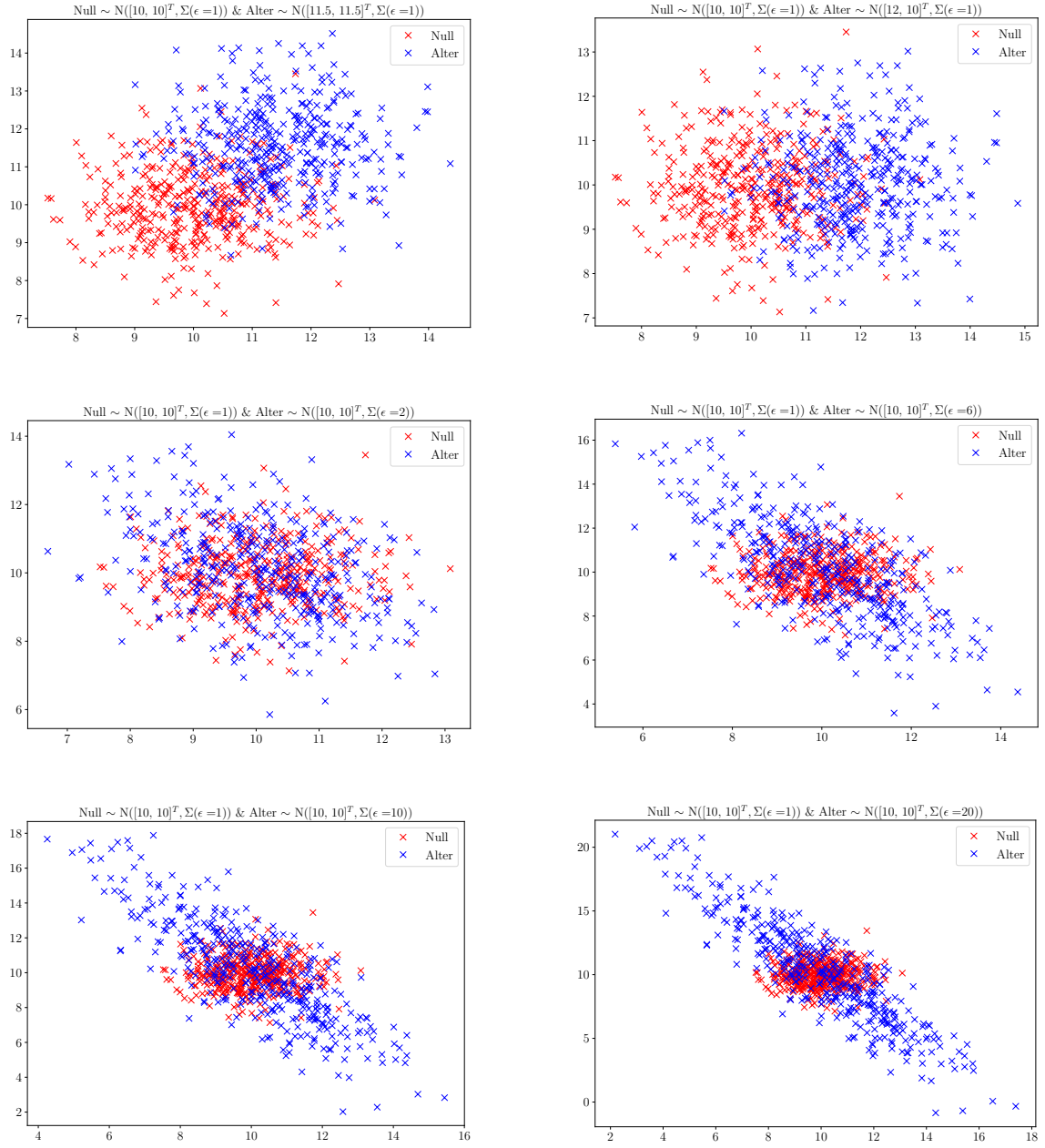


Figure 6.14: Visualisation of 2-dimensional Gaussian distributions from Section 6.6.2 with 400 samples each.

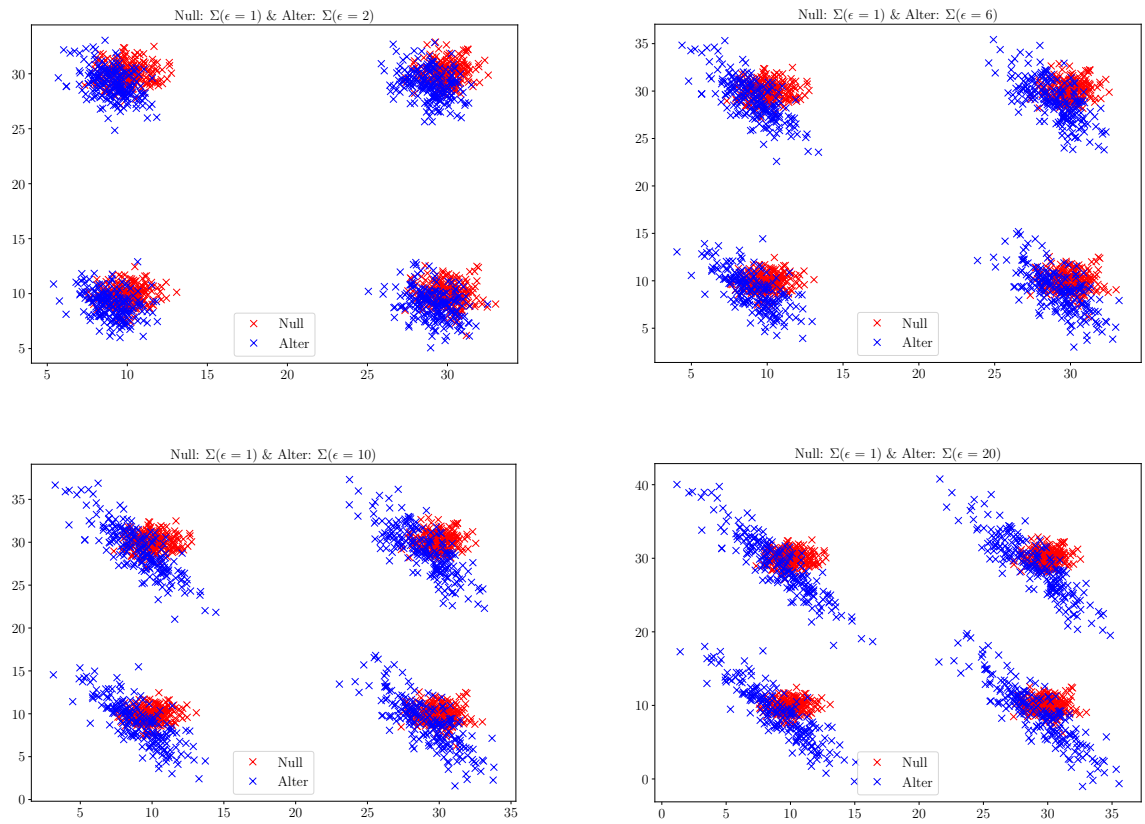


Figure 6.15: Visualisation of 2 by 2 Blobs of 2-dimensional Gaussian distributions from Section 6.6.4 with 200 samples in each blob.

Chapter 7

Conclusion and Future work

In this thesis, we have made several contributions to the field of hypothesis testing with kernel methods: we have presented a range of methods for scaling up kernel independence tests and normalised kernel independence tests; proposed a two-step conditional independence test which corrects type I control of previous two-step approaches; and also proposed a Bayesian two-sample test which allows integrating over kernel hyperparameters. We have focused on independence testing in the first part of this thesis. In Chapter 3, we have proposed three large scale estimators of the Hilbert Schmidt Independence Criterion (HSIC): the block-based estimator, the Nyström estimator, the random Fourier features (RFF) estimator. Suitable independence testing procedures has been established for each method. For the block HSIC statistics, one has taken advantage of its asymptotic normality which is a consequence of the Central Limit Theorem (CLT). For the Nyström and RFF HSIC statistics, as an alternative to the permutation approach, we have proposed to directly estimate the eigenvalues from the approximated Gram matrices and appeal to the kernel spectral approach to arrive at the appropriate form of the asymptotic null distributions. All three proposed tests have achieved significant reduction in computational complexity in memory and time over the standard HSIC-based test. In Chapter 4, we have mimicked the large scale estimation procedure for HSIC and proposed the RFF Normalised Cross COvariance (NOCCO) test. The advantage of the proposed statistic is that it enables us to derive a null distribution independent of the kernel, significantly simplifying the testing procedure. The resulting large scale normalised kernel independence test has runtime complexity that is linear in the number of samples, in contrast to the standard NOCCO-based test which is cubic in the number of samples. From a range of synthetic experiments, we have observed that the normalised approach is more robust to kernel bandwidth selection in comparison to HSIC-based test and its large scale counterparts. In Chapter 5, we have proposed a conditional

independence test Kernel RESIT (or KRESIT) that extends the popular two-step approach REgression with Subsequent (unconditional) Independence Test (RESIT). The proposed test uses RKHS-valued ridge regressions that attempt to regress out the dependence on the conditioning variable before conducting an independence test using HSIC. While maintaining simple and effective testing procedures of RESIT, the resulting test has correct Type I control even when the conditioning variables induce challenging switching associations, where RESIT is inappropriate.

The final part of this thesis was focused on two-sample testing. While the method developed is not restrictive to this particular hypothesis testing problem, we have left its application to independence testing as future work. In Chapter 6, we have proposed a Bayesian two-sample testing framework utilising the Bayes factor. Instead of directly modelling the input observations, we have proposed to consider the difference between mean embeddings evaluated at a set of inducing points. While Flaxman et al. [2016] models the mean embeddings directly with a Gaussian Process (GP), we have employed their framework on the difference between mean embeddings. We have derived efficient computation methods of the marginal pseudolikelihood through the Kronecker structure of the covariance matrices. As a part of the proposed Bayesian test, we have developed a posterior inference scheme that enables the joint learning of the posterior distributions of the model label and at the same time, the bandwidth parameter of the kernel. This has provided a different perspective on traditional bandwidth selection in frequentist kernel-based methods more broadly, where typically a single value of kernel hyperparameters is utilised. The proposed test is able to uncover sample differences in a variety of synthetic experiments given sufficient data and also lead to discovering the full distribution of lengthscales relevant for the problem.

7.1 Future Work

For the large-scale normalised independence work, one possible topic of research is to further explore the connection with the work of Dauxois and Nkiet [1998]. The work considers a deterministic orthonormal basis, however, it takes almost exactly the same form as the RFF NOCCO statistics we have proposed for fixed finite basis. This set of orthonormal basis leads to an operator whose unit eigenvector can be easily removed. The significance of working with operator centered in such way is unclear to us and it

would be interesting to understand if there are equivalent notions in our framework. In particular, RFF are not orthonormal, but approaches utilizing Mercer’s expansions may be considered. One may also ask if we can still use Hotelling’s T-statistics to derive the null distribution. More generally, is there an RKHS equivalent to Dauxois and Nkiet [1998]’s system of orthonormal basis? These are just a few questions that we would like to answer. Although asymptotic consistency to the population quantity was considered in Dauxois and Nkiet [1998], they never considered the consistency as the number of basis increases (equivalently in our framework, as the number of random Fourier feature increases). Establishing such consistency can also be a nice direction for future work.

Recently, Li et al. [2019] proposed a unified analysis of the generalisation properties of random Fourier features, which is applicable to kernel ridge regression, kernel support vector machines and kernel logistic regression. They provided the sharpest results on the number of features required in terms of the regularisation parameter and the number of effective degrees of freedoms. No such analysis has yet been established in the setting of kernel hypothesis testing and this provides an important direction for future research, since it would further shed light on the tradeoffs between computational efficiency and testing power which have motivated our development of large-scale tests. For example, one question to ask is could the analysis in the work of Li et al. [2019] be applied to the kernel testing framework. In the same work, they also used a modified data-dependent sampling distribution for RFF called the empirical ridge leverage score distribution (correspondingly, leverage weighted RFF) and similar ideas have previously been used to achieve a sharper convergence rate of the Nyström approximated estimator. This sampling scheme enabled the use of less features to achieve the same learning risk comparing to a plain random Fourier feature sampling scheme. It would be interesting to investigate if such phenomena can be empirically observed in our kernel hypothesis testing setting. Further, if one can establish similar theoretical guarantees comparing the plain RFF and leveraged weighted RFF.

Parallel to our large-scale independence tests based on HSIC, Jitkrittum et al. [2017] proposed a linear time dependence measure with an adaptive statistical test of independence. Their test statistics, the Finite Set Independence Criterion (FSIC), is the square L^2 norm between the kernel mean embedding of the joint probability distributions and that of the product of the marginals, which can be understood

as an average of covariances of analytic functions (i.e. features) defined on each of X and Y . Their test is adaptive in the sense that their features are chosen on a held-out validation set such that they maximise a lower bound on the test power. However, the asymptotic null distribution of the FSIC statistics is a sum of dependent weighted χ^2 random variables which will lead to spectral type of approach for testing. They instead proposed an independence test based on Normalised FSIC (NFSIC), this parallels that of our RFF NOCCO test proposed in 4, as they both achieve a distribution-free test. Due to the optimised features, their test performs comparably to the quadratic HSIC test and outperforms our large-scale kernel test from Chapter 3 where features are sampled randomly.

The framework that has been established in Chapter 3 and 4 can also be applied to a range of other recent results in the statistics and machine learning community. For example, it can be directly applied to the recently proposed joint independence test of Pfister et al. [2017]. The work establishes as a test statistics, d -variable HSIC (dHSIC) — the square Hilbert Schmidt norm between the joint distribution of d multivariate random variables and the product of their marginals. In the population case, the value of dHSIC is zero if and only if the d variables are jointly independent given characteristic kernels. This joint independence test includes HSIC as a special case. The work has direct application in causal inference as a goodness-of-fit test and for model selection and our large-scale framework will enable its use with large datasets. Similarly, it can also be applied to scale up the procedure of Nguyen and Eisenstein [2017] that measures geographical language variation through the use of HSIC between geographical and linguistic variables. For the text-matching problem considered in Roberts et al. [2019], balance checking is an important step for assessing whether the matching is successful. It requires a similarity measure between the matched units and the pre-treatment confounders. MMD can be applied and our large-scale approximation techniques can be applied as well (see the references in Chapter 3).

The developments in this thesis have already found applications in the work of others – in particular, the proposed block estimator for HSIC from Chapter 3 has been subsequently applied in Yamada et al. [2018] for the problem of post selective inference (PSI). Block HSIC follows an asymptotic normal distribution when the number of blocks is finite and the number of samples goes to infinity, this fits the assumption of the PSI framework which allows Yamada et al. [2018] to propose HSIC based PSI

algorithm that is able to handle nonlinear, multi-label, multi-class and structured output. We also see its use in the work of Climente-Gonzalez et al. [2019] to scale up the HSIC LASSO [Yamada et al., 2014] for biomarker discovery (i.e. feature selection with redundancy between features) while maintaining the convexity of the final optimisation problem.

For the Bayesian two-sample testing work, there are several possible directions for future research. We have seen in Section 6.6 that larger sample sizes are required for more challenging problems. A random Fourier feature approximation of the above framework can be easily developed to enable the use of large sample size without having prohibitive runtime. In this case, explicit finite dimensional feature maps are available, the difference between the mean embeddings $\delta = \mu_X - \mu_Y$ can be written more explicitly as $\delta = \mathbb{E}(\phi(X)) - \mathbb{E}(\phi(Y))$. Assume a GP prior with an appropriate covariance matrix for δ , $\phi(x_i) - \phi(y_i)$ can be modelled by a Gaussian distribution with mean $\mathbb{E}(\phi(X)) - \mathbb{E}(\phi(Y))$ and covariance estimated as in Section 6.3.2. While the rest of the inference procedure can follow similarly as presented in Chapter 6, this large scale approximation requires careful specification of the covariance matrix for the GP model of δ to ensure that draws from such GP lie in the correct RKHS.

Recently, Kamary et al. [2014] proposed a mixture modelling framework for Bayesian model selection. The proposed Bayesian two-sample testing framework using Bayes factor can be equivalently formulated as a mixture model:

$$\Delta|\theta, \pi \sim \pi N(0, \frac{1}{n}\Sigma_\theta) + (1 - \pi)N(0, R_\theta + \frac{1}{n}\Sigma_\theta).$$

The posterior distribution of the mixture proportion $0 \leq \pi \leq 1$ indicates the model preferred. A joint inference of π and the kernel bandwidth parameter θ can be easily done through MCMC. It would be interesting to see if there is a difference in performance between the mixture approach and the Bayes factor approach proposed here. Lastly, the Bayesian testing framework developed here and the directions for future work can all be applied to independence testing.

Bibliography

- World Health Organization Statistical Information Systems (WHOSIS). World Health Organization Statistical Information Systems (WHOSIS), 2009. www.who.int/whosis/en/.
- S. Akaho. A Kernel Method for Canonical Correlation Analysis. In *In Proceedings of International Meeting on Psychometric Society (IMPS2001)*, 2001.
- S.M. Ali and S.D. Silvey. A General Class of Coefficients of Divergence of One Distribution from Another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1):131–142, 1966.
- D. Alpay. *The Schur Algorithm, Reproducing Kernel Spaces and System Theory*. American Mathematics Society, 2001.
- N.H. Anderson, P. Hall, and D.M. Titterington. Two-Sample Test Statistics for Measuring Discrepancies Between Two Multivariate Probability Density Functions Using Kernel-Based Density Estimates. *Journal of Multivariate Analysis*, 50:41–54, 1994.
- M.A. Arcones and E. Giné. On the Bootstrap of U and V Statistics. *The Annals of Statistics*, 20(2):655–674, 1992.
- N. Aronszajn. Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.
- F.R. Bach and M.I. Jordan. Kernel Independent Component Analysis. *Journal of Machine Learning Research (JMLR)*, 10:1–48, 2002.
- C. R. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.
- L. Baringhaus and C. Franz. On a New Multivariate Two-Sample Test. *Journal of Multivariate Analysis*, 88(1):190–206, 2004.

- A. Ben-Israel. The Change-of-Variables Formula Using Matrix Volume. *SIAM Journal on Matrix Analysis and Applications*, 21(1):300–312, 1999.
- W. Bergsma. Testing Conditional Independence for Continuous Random Variable. *EURANDOM-report 2004-049*, 2004.
- A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.
- J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. Wiley, 2000.
- T. Bertin-Mahieux, D.P. Ellis, B. Whitman, and P. Lamere. The Million Song Dataset. *International Conference on Music Information Retrieval (ISMIR)*, 2011.
- M. Blaschko and A. Gretton. Learning Taxonomies by Dependence Maximization. In *Advances in Neural Information Processing Systems 21 (NIPS)*, pages 153–160. 2009.
- K.M. Borgwardt and Z. Ghahramani. Bayesian Two-Sample Tests. *ArXiv e-prints: 0906.4032*, 2009.
- K.M. Borgwardt, A. Gretton, M.J. Rasch, H P. Kriegel, B. Schölkopf, and A.J. Smola. Integrating Structured Biological Data by Kernel Maximum Mean Discrepancy. *Bioinformatics*, 22(14):49–57, 2006.
- L. Breiman and J.H. Friedman. Estimating Optimal Transformations for Multiple Regression and Correlation. *Journal of the American Statistical Association*, 80 (391):580–598, 1985.
- Y. Cho and L.K. Saul. Kernel Methods for Deep Learning. In *Advances in Neural Information Processing Systems 22 (NIPS)*, pages 342–350. 2009.
- K. Chwialkowski and A. Gretton. A Kernel Independence Test for Random Processes. In *Proceedings of The 31st International Conference on Machine Learning (ICML)*, 2014.
- K. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast Two-Sample Testing with Analytic Representations of Probability Measures. In *Advances in Neural Information Processing Systems (NIPS)*, volume 28, 2015.

- K. Chwialkowski, H. Strathmann, and A. Gretton. A Kernel Test of Goodness of Fit. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.
- H. Climente-Gonzalez, C. A. Azencott, S. Kaski, and M. Yamada. Block HSIC Lasso: Model-Free Biomarker Detection for Ultra-High Dimensional Data. 2019.
- C. Cortes, M. Mohri, and A. Rostamizadeh. Algorithms for Learning Kernels Based on Centered Alignment. *Journal of Machine Learning Research (JMLR)*, 13(1): 795–828, March 2012.
- T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 1991.
- I. Csiszár. Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Magyar. Tud. Akad. Mat. Kutató Int. Kozl.*, 8(85-108), 1963.
- J.J.M. Cuppen. A Divide and Conquer Method for the Symmetric Tridiagonal Eigenproblem. *Numerische Mathematik*, 36:177–195, 1981.
- J. Dattorro. *Convex Optimisation & Euclidean Distance Geometry*. Meboo Publishing USA, second edition, 2016.
- J.J. Daudin. Partial Association Measures and an Application to Qualitative Regression. *Biometrika*, 67(3):581–590, 1980.
- J. Dauxois and G.M. Nkiet. Nonlinear Canonical Analysis and Independence Tests. *The Annals of Statistics*, 26(4):1254–1278, 1998.
- M.H. DeGroot. Doing What Comes Naturally: Interpreting a Tail Area as a Posterior Probability or as a Likelihood Ratio. *Journal of the American Statistical Association*, 68(344):966–969, 1973.
- G. Doran, K. Muandet, K. Zhang, and B. Schölkopf. A Permutation-Based Kernel Conditional Independence Test. In *Uncertainty in Artificial Intelligence*, 2014.
- R.M. Dudley. *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2 edition, 2002.
- S. Filippi and C.C. Holmes. A Bayesian Nonparametric Approach to Testing for Dependence Between Random Variables. *Bayesian Analysis Advance Publication*, 2016.

- S. Filippi, C.C. Holmes, and L.E. Nieto-Barajas. Scalable Bayesian Nonparametric Measures for Exploring Pairwise Dependence via Dirichlet Process Mixtures. *Electronic Journal of Statistics*, 10(2):3338–3354, 2016.
- S. Flaxman, D.B. Neill, and A.J. Smola. Gaussian Processes for Independence Tests with Non-iid Data in Causal Inference. *ACM Transactions on Intelligent Systems and Technology*, 7(2):22:1–22:23, 2015.
- S. Flaxman, D. Sejdinovic, J.P. Cunningham, and S. Filippi. Bayesian Learning of Kernel Embeddings. In *Uncertainty in Artificial Intelligence (UAI)*, pages 182–191, 2016.
- J.G.F. Francis. The QR Transformation, I. *The Computer Journal*, 4(3):265–271, 1961.
- J.G.F. Francis. The QR Transformation, II. *The Computer Journal*, 4(4):332–345, 1962.
- K. Fukumizu, F.R. Bach, and M.I. Jordan. Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces. *Journal of Machine Learning Research (JMLR)*, 5:73–99, 2004a.
- K. Fukumizu, F.R. Bach, and M.I. Jordan. Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Space. *Journal of Machine Learning Research (JMLR)*, 5:73 – 99, 2004b.
- K. Fukumizu, F. R. Bach, and A. Gretton. Statistical Consistency of Kernel Canonical Correlation Analysis. *Journal of Machine Learning Research (JMLR)*, 8:361–383, May 2007.
- K. Fukumizu, A. Gretton, X.H. Sun, and B. Schölkopf. Kernel Measures of Conditional Dependence. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.
- K. Fukumizu, L. Song, and A. Gretton. Kernel Bayes Rule. In *Advances in Neural Information Processing Systems 24 (NIPS)*, pages 1737–1745. 2011.
- A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring Statistical Dependence with Hilbert-Schmidt Norms. In *Proceedings of the 16th International Conference on Algorithmic Learning Theory*, pages 63–77, 2005a.

- A. Gretton, A. Smola, O. Bousquet, R. Herbrich, A. Belitski, M. Augath, Y. Murayama, J. Pauls, B. Schölkopf, and N. Logothetis. Kernel Constrained Covariance for Dependence Measurement. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 1–8, January 2005b.
- A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A Kernel Method for the Two-Sample problem. In *Advances in Neural Information Processing Systems (NIPS)*, pages 513–520, 2007.
- A. Gretton, K. Fukumizu, B. Schölkopf, C.H. Teo, L. Song, and A. Smola. A Kernel Statistical Test of Independence. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.
- A. Gretton, K. Fukumizu, Z. Harchaoui, and B. Sriperumbudur. A Fast, Consistent Kernel Two-Sample Test. In *Advances in Neural Information Processing Systems 22 (NIPS)*, 2009.
- A. Gretton, K. Borgwardt, M.J. Rasch, B. Schölkopf, and A. Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research (JMLR)*, 13:723–773, 2012a.
- A. Gretton, B. Sriperumbudur, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, and K. Fukumizu. Optimal Kernel Choice for Large-Scale Two-Sample Tests. *Advances in Neural Information Processing Systems (NIPS)*, 2012b.
- D. Harrison Jr. and D.L. Rubinfeld. Hedonic Housing Prices and the Demand for Clean Air. *Journal of Environmental Economics and Management*, 5:81–102, 1978.
- D.M. Hausman and J. Woodward. Independence, Invariance, and the Causal Markov Condition. *British Journal of Philosophy and Science*, 50(4):521–583, 1999.
- C.C. Holmes, F. Caron, J.E. Griffin, and D.A. Stephens. Two-Sample Bayesian Nonparametric Hypothesis Testing. *Bayesian Analysis*, 10(2):297–320, 06 2015.
- R.A. Horn and C.R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.
- H. Hotelling. The Generalization of Student’s Ratio. *The Annals of Mathematical Statistics*, 2(3):360–378, 1931.
- H. Hotelling. Relations Between Two Sets of Variates. *Biometrika*, pages 321–377, 1936.

- P.O. Hoyer, D. Janzing, J.M. Mooij, J. Peters, and B. Schölkopf. Nonlinear Causal Discovery with Additive Noise Models. In *Advances in Neural Information Processing Systems 21 (NIPS)*, pages 689–696. 2009.
- S.Y. Huang, M. H. Lee, and C.K. Hsiao. Nonlinear Measures of Association with Kernel Canonical Correlation Analysis and Applications. *Journal of Statistical Planning and Inference*, 139(7):2162 – 2174, 2009.
- W.H. Jefferys and J.O. Berger. Ockham’s Razor and Bayesian Analysis. *American Journal of Science*, 80(1):64–72, 1992.
- H. Jeffreys. Some Tests of Significance, Treated by the Theory of Probability. *Proceedings of the Cambridge Philosophy Society*, (31):203–222, 1935.
- H. Jeffreys. *Theory of Probability*. Oxford University Press, 1961.
- W. Jitkrittum, Z. Szabó, and A. Gretton. An Adaptive Test of Independence with Analytic Kernel Embeddings. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 70, 2017.
- M. Kalisch and P. Buhlmann. Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. *Journal of Machine Learning Research (JMLR)*, 8:613–636, 2007.
- K. Kamary, K. Mengersen, C.P. Robert, and J. Rousseau. Testing Hypotheses via a Mixture Estimation Model. *ArXiv e-prints: 1412.2044*, 2014.
- R.E. Kass and A.E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795, June 1995.
- A. Kraskov, H. Stoegbauer, and P. Grassberger. Estimating Mutual Information. 2003.
- E. Kreyszig. *Introductory Functional Analysis with Applications*. Wiley, 1989.
- V.N. Kublanovskaya. On Some Algorithms for the Solution of the Complete Eigenvalue Problem. *USSR Computational Mathematics and Mathematical Physics*, 1(3):637–657, 1961.
- P. Lai and C. Fyfe. Kernel and Nonlinear Canonical Correlation Analysis. *International Journal of Neural Systems*, 3, 2000. ISSN 5.

- Z. Li, J. F. Ton, D. Oglic, and D. Sejdinovic. Towards A Unified Analysis of Random Fourier Features. In *International Conference on Machine Learning (ICML)*, 2019.
- F. Liese and I. Vajda. On Divergences and Informations in Statistics and Information Theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, Oct 2006.
- D. Lopez-Paz. *From Dependence to Causation*. PhD thesis, Max Planck Institute for Intelligent Systems and the University of Cambridge, 2016.
- D. Lopez-Paz, P. Hennig, and B. Schölkopf. The Randomized Dependence Coefficient. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 1–9. 2013.
- D. Lopez-Paz, S. Sra, A. Smola, Z. Ghahramani, and B. Schölkopf. Randomized Nonlinear Component Analysis. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 1359–1367, 2014.
- M.N. Lukić and J.H. Beder. Stochastic Processes with Sample Paths in Reproducing Kernel Hilbert Spaces. *Transaction of the American Mathematical Society*, 353(10):3945–3969, 2001.
- R. Lyons. Distance Covariance in Metric Spaces. *The Annals of Probability*, 41(5):3284–3305, 2013.
- D.J.C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.
- K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis*. Academic Press, 1979.
- T. Melzer, M. Reiter, and H. Bischof. Nonlinear Feature Extraction Using Generalized Canonical Correlation Analysis. In *In Proceedings of International Conference on Artificial Neural Networks (ICANN)*, page 353–360, 2001.
- T. Morimoto. Markov Processes and the H-Theorem. *Journal of the Physical Society of Japan*, 18:328–331, 1963.
- A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probabilities*, 29:429–443, 1997.
- D. Nguyen and J. Eisenstein. A Kernel Independence Test for Geographical Language Variation. *Computational Linguistics*, 43(3):567–592, 2017.

- R.K. Pace and O.W. Gilley. Using the Spatial Configuration of the Data to Improve Estimation. *The Journal of Real Estate Finance and Economics*, 14(3):333–340, 1997.
- J. Peters, J.M. Mooij, D. Janzing, and B. Schölkopf. Causal Discovery with Continuous Additive Noise Models. *Journal of Machine Learning Research (JMLR)*, 15(1):2009–2053, 2014.
- N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters. Kernel-Based Tests for Joint Independence. *Journal of the Royal Statistical Society, Series B*, 80(1):5–31, 2017.
- B. Póczos, Z. Ghahramani, and J.G. Schneider. Copula-based Kernel Dependency Measures. In *Proceedings of the 29th International Conference on Machine Learning, (ICML)*, 2012.
- A. Rahimi and B. Recht. Random Features for Large-Scale Kernel Machines. *Advances in Neural Information Processing Systems 20 (NIPS)*, 2007.
- M. Reed and B. Simon. *Methods of Modern Mathematical Physics. I: Functional Analysis*. Academic Press, second edition, 1980.
- D.N. Reshef, Y.A. Reshef, H.K. Finucane, S.R. Grossman, G. McVean, P.J. Turnbaugh, E.S. Lander, M. Mitzenmacher, and P.C. Sabeti. Detecting Novel Associations in Large Data Sets. *Science*, 334(6062):1518–1524, 2011.
- M.E. Roberts, B.M. Stewart, and R.A. Nielsen. Adjusting for Confounding with Text Matching, 2019.
- H. Rosling. Indicators in Gapminder World. Gapminder, 2008.
- P.K. Rubenstein, K.P. Chwialkowski, and A. Gretton. A Kernel Test for Three-Variable Interactions with Random Processes. In *Uncertainty in Artificial Intelligence (UAI)*, 2016.
- K. Schäcke. On the Kronecker Product. Technical report, University of Waterloo, 2013.
- B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

- D. Sejdinovic, A. Gretton, and W. Bergsma. A Kernel Test for Three-Variable Interactions. In *Advances in Neural Information Processing Systems (NIPS)*, volume 26, pages 1124–1132. 2013a.
- D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of Distance-Based and RKHS-Based Statistics in Hypothesis Testing. *The Annals of Statistics*, 41(5):2263–2291, 2013b.
- D. Sejdinovic, H. Strathmann, S. De, W. Zaremba, M. Blaschko, and A. Gretton. Big Hypothesis Tests with Kernel Embeddings: an Overview. Technical Report, 2014. Gatsby Unit, UCL.
- R.J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, Inc., 2002.
- C. Shalizi. *Advanced Data Analysis from an Elementary Point of View*. 2016. URL <http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/>.
- J. Shao. *Mathematical Statistics*. Springer, 2007.
- J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*, volume 19 of 22. Cambridge University Press, 2004.
- A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert Space Embedding for Distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory*, pages 13–31, 2007.
- E. Snelson and Z. Ghahramani. Sparse Gaussian Processes using Pseudo-Inputs. In *Advances in Neural Information Processing Systems 18 (NIPS)*, pages 1257–1264. MIT press, 2006.
- L. Song, A. Smola, A. Gretton, and K.M. Borgwardt. A Dependence Maximization View of Clustering. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pages 815–822, 2007.
- L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert Space Embeddings of Conditional Distributions with Applications to Dynamical Systems. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.
- L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature Selection via Dependence Maximization. *Journal of Machine Learning Research*, 13:1393–1434, 2012.

- P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.
- B.K. Sriperumbudur. Reproducing Kernel Space Embeddings and Metrics on Probability Measures. *PhD Thesis, University of California–San Diego*, 2010.
- B.K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G.R.G. Lanckriet. Hilbert Space Embeddings and Metrics on Probability Measures. *Journal of Machine Learning Research (JMLR)*, 11:1517–1561, 2010.
- B.K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G.R.G. Lanckriet. On the Empirical Estimation of Integral Probability Metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.
- O. Stegle, K.J. Denby, E.J. Cooke, D.L. Wild, Z. Ghahramani, and K.M. Borgwardt. A Robust Bayesian Two-Sample Test for Detecting Intervals of Differential Gene Expression in Microarray Time Series. *Journal of Computational Biology*, 17(3), 2010.
- I. Steinwart. On the Influence of the Kernel on the Consistency of Support Vector Machines. *Journal of Machine Learning Research (JMLR)*, 2:67–93, 2001.
- I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.
- E.V. Strobl, S. Visweswara, and K. Zhang. Approximate Kernel-based Conditional Independence Tests for Fast Non-Parametric Causal Discovery. *Journal of Causal Inference*, 7(1), 2018.
- D.J. Sutherland and J. Schneider. On the Error of Random Fourier Features. *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2015.
- G.J. Székely and M.L. Rizzo. Testing for Equal Distributions in High Dimensions. *InterStat*, 2004.
- G.J. Székely and M.L. Rizzo. A New Test for Multivariate Normality. *Journal of Multivariate Analysis*, 93(1):58–80, 2005.
- G.J. Székely and M.L. Rizzo. Brownian Distance Covariance. *The Annals of Applied Statistics*, 3(4):1236–1265, 2009.
- G.J. Székely, M.L. Rizzo, and N.K. Bakirov. Measuring and Testing Dependence by Correlation of Distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

- Aad van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.
- G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.
- H. Wendland. *Scattered Data Approximation*. Cambridge University Press, Cambridge, UK, 2005.
- C.K.I. Williams and M. Seeger. Using the Nyström Method to Speed Up Kernel Machines. In *Advances in Neural Information Processing Systems 13 (NIPS)*, pages 682–688. 2001.
- M. Yamada, W. Jitkrittum, L. Sigal, E.P. Xing, and M. Sugiyama. High-Dimensional Feature Selection by Feature-Wise Kernelized Lasso. *Neural Computation*, 26(1): 185–207, 2014.
- M. Yamada, Y. Umezū, K. Fukumizu, and I. Takeuchi. Post Selection Inference with Kernels. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, volume 84, pages 152–160, 2018.
- G.A. Young and R.L. Smith. *Essentials of Statistical Inference*. Cambridge University Press, 2005.
- A. Zaremba and T. Aste. Measures of Causality in Complex Datasets with Application to Financial Data. *Entropy*, 16(4):2309, 2014.
- W. Zaremba, A. Gretton, and W. Blaschko. B-Test: A Non-Parametric, Low Variance Kernel Two-Sample Test. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.
- K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based Conditional Independence Test and Application in Causal Discovery. In *Uncertainty in Artificial Intelligence (UAI)*, pages 804–813, 2011.
- Q. Zhang, S. Filippi, S. Flaxman, and D. Sejdinovic. Feature-to-Feature Regression for a Two-Step Conditional Independence Test. In *Uncertainty in Artificial Intelligence (UAI)*, 2017.
- Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic. Large-Scale Kernel Methods for Independence Testing. *Statistics and Computing*, 28(1):113–130, January 2018.
- J. Zhao and D. Meng. FastMMD: Ensemble of Circular Discrepancy for Efficient Two-sample Test. *Neural Computation*, 27(6):1345–1372, June 2015.