Topics in kernel hypothesis testing

Kacper Chwialkowski

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Department of Computer Science, Gatsby Computational Neuroscience Unit
University College London

October 4, 2016

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

I thank my parents.

Abstract

This thesis investigates some unaddressed problems in kernel nonparametric hypothesis testing. The contributions are grouped around three main themes:

Wild Bootstrap for Degenerate Kernel Tests. A wild bootstrap method for nonparametric hypothesis tests based on kernel distribution embeddings is proposed. This bootstrap method is used to construct provably consistent tests that apply to random processes. It applies to a large group of kernel tests based on V-statistics, which are degenerate under the null hypothesis, and non-degenerate elsewhere. In experiments, the wild bootstrap gives strong performance on synthetic examples, on audio data, and in performance benchmarking for the Gibbs sampler.

A Kernel Test of Goodness of Fit. A nonparametric statistical test for goodness-of-fit is proposed: given a set of samples, the test determines how likely it is that these were generated from a target density function. The measure of goodness-of-fit is a divergence constructed via Stein's method using functions from a Reproducing Kernel Hilbert Space. Construction of the test is based on the wild bootstrap method. We apply our test to quantifying convergence of approximate Markov Chain Monte Carlo methods, statistical model criticism, and evaluating quality of fit vs model complexity in nonparametric density estimation.

Fast Analytic Functions Based Two Sample Test. A class of nonparametric two-sample tests with a cost linear in the sample size is proposed. Two tests are given, both based on an ensemble of distances between analytic functions representing each of the distributions. Experiments on artificial benchmarks and on challenging real-world testing problems demonstrate good power/time tradeoff retained even in high dimensional problems.

The main contributions to science are the following. We prove that the kernel tests based on the wild bootstrap method tightly control the type one error on the desired level and are consistent i.e. type two error drops to zero with increasing number of samples. We construct a kernel goodness of fit test that requires only knowledge of the density up to an normalizing constant. We use this test to construct first consistent test for convergence of Markov Chains and use it to quantify properties of approximate MCMC algorithms. Finally, we construct a linear time two-sample test that uses new, finite dimensional feature representation of probability measures.

Contents

The thesis are based on the following publications:

- Kacper Chwialkowski and Arthur Gretton. A kernel independence test for random processes. In *ICML*, 2014
- Kacper Chwialkowski, Dino Sejdinovic, and Arthur Gretton. A wild bootstrap for degenerate kernel tests. In *Advances in Neural Information Processing Systems 27*, pages 3608–3616. Curran Associates, Inc., 2014
- Kacper Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, pages 1972–1980, 2015
- Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *ICML*, 2016

Chapter 1

Introduction

In this chapter, we describe the research motivation, research objectives, contributions to science, and finish with an outline of the document.

Research Motivation

Wild Bootstrap for Degenerate Kernel Tests. Statistical kernel tests, that is tests based on distribution embeddings into reproducing kernel Hilbert spaces (RKHS), have been applied in many contexts, including two sample tests by Harchaoui et al. [52], Gretton et al. [47], Sugiyama et al. [107], tests of independence by Gretton et al. [44], Zhang et al. [115], Besserve et al. [14], tests of conditional independence by Fukumizu et al. [37], Gretton et al. [49], Zhang et al. [115], test for higher order (Lancaster) interactions by Sejdinovic et al. [87], and normality test by Baringhaus and Henze [8]. Another example is kernel goodnesses-of-fit test, developed in this thesis.

For these tests, consistency is usually guaranteed if the observations are independent and identically distributed (i.i.d.), an exception being Zhang et al. [116]. Much real-world data fails to satisfy the i.i.d. assumption: audio signals, EEG recordings, text documents, financial time series, and samples obtained when running Markov Chain Monte Carlo, all show significant temporal dependence patterns.

The asymptotic behaviour of statistics, used in kernel tests, may become quite different when temporal dependencies exist within the samples. In this case, the null distribution is shown to be an infinite weighted sum of *dependent* χ^2-variables, as opposed to the sum of *independent* χ^2-variables obtained in the case of i.i.d. observations [44]. The difference in the asymptotic null distri-

butions has important implications in practice: under the i.i.d. assumption, an empirical estimate of the null distribution can be obtained by repeatedly permuting the time indices of one of the signals. This breaks the temporal dependence within the permuted signal, which causes the test to return an elevated number of false positives, when used for testing time series.

To address this problem, in the preliminary work we proposed an alternative estimate of the null distribution, where the null distribution is simulated by repeatedly *shifting* one signal relative to the other (which is a form a block bootstrap). This preserves the temporal structure within each signal, while breaking the cross-signal dependence.

A serious limitation of the shift bootstrap procedure is that it is specific to the problem of independence testing: there is no obvious way to generalise it to other testing contexts. For instance, we might have two time series, with the goal of comparing their marginal distributions - this is a generalization of the two-sample setting to which the shift approach does not apply. Another example is kernel goodness-of-fit test developed in this thesis.

It is interesting to study whether some other bootstrap procedure can be used in place of the Shift bootstrap, so that all kernel tests are consistent if applied to data with temporal dependence patterns.

A Kernel Test of Goodness of Fit. A particular type of a statical test, a goodness-of-fit test, is a fundamental tool in statistical analysis, dating back to the test of Kolmogorov and Smirnov [62, 94]. Given a set of samples $\{Z_i\}_{i=1}^n$ with distribution $Z_i \sim q$, our interest is in whether q matches some reference or target distribution p, which we assume to be only known up to the normalisation constant. This setting, in which target density is not exactly known, is quite challenging and particularly relevant in Markov Chain Monte Carlo diagnostic.

Recently Gorham and Mackey [40] proposed an elegant measure of sample quality with respect to a target. This measure is a maximum discrepancy between empirical sample expectations and target expectations over a large class of test functions, constructed so as to have zero expectation over the target distribution by use of a Stein operator. This operator depends only on the derivative of the $\log q$: thus, the approach can be applied very generally, as it

does not require closed-form integrals over the target distribution (or numerical approximations of such integrals). By contrast, many earlier discrepancy measures require integrals with respect to the target. This is problematic if the intention is to perform benchmarks for assessing Markov Chain Monte Carlo methods, since these integrals will certainly not be known to the practitioner.

A challenge in applying the approach of Gorham and Mackey [40] is the complexity of the function class used, which results from applying the Stein operator to the bounded Lipschitz functions.[1].

An important application of a goodness-of-fit measure is in statistical testing, where it is desired to determine whether the empirical discrepancy measure is large enough to reject the null hypothesis (that the sample arises from the target distribution). One approach is to establish the asymptotic behaviour of the test statistic, and to set a test threshold at a large quantile of the asymptotic distribution. The asymptotic behaviour of the Lipschitz functions based Stein discrepancy remains a challenging open problem, due to the complexity of the function class used. It is not clear how one would compute p-values for this statistic, or determine when the goodness of fit test would allow us to accept the null hypothesis (at the user-specified test level).

It is interesting to study if replacing the class of bounded Lipschitz functions with some other class of functions, a natural candidate being some reproducing kernel Hilbert space, one can develop a goodness-of-fit test. Since such a test would be a very useful tool in Markov Chain Monte Carlo methods convergance diagnostics, another consideration is whether it can be applied to data with temporal dependence patterns.

Fast Analytic Functions Based Two Sample Test. Since most of nonparametric test, including kernel tests, are computationally expensive, we turn attention to computational complexity related problems. We focus on a two-sample testing problem, hoping that results for this problem can translated to other types of kernel tests. Traditional approaches to two-sample testing are based on distances between representations of the distributions, such as density functions, cumulative distribution functions, characteristic functions or mean embeddings in a reproducing kernel Hilbert space [97, 99]. These representa-

[1]The bounded Lipschitz functions give rise to the Wasserstein integral probability metric. By contrast, the Kolmogorov-Smirnov test uses functions of bounded variation 1 [73]

tions are infinite dimensional objects, which poses challenges when defining a distance between distributions. Examples of such distances include the classical Kolmogorov-Smirnov distance (sup-norm between cumulative distribution functions); the Maximum Mean Discrepancy (MMD) [45], an RKHS norm of the difference between mean embeddings, and the $\mathbb{N}$-distance (also known as energy distance) [118, 111, 7], which is also an MMD-based test for a particular family of kernels [88].

This traditional approaches usually result in tests with a quadratic time complexity. While there exists a line of work which modifies some of those quadratic tests [114, 48], it exploits solutions in which the algorithm is designed to ignore some information available e.g. use only part of a kernel matrix. It is interesting to study if it is possible to construct a linear time complexity test by considering more compact representation of probability measures (more compact compared to classical infinite dimensional representation).

Research Objectives

As outlined in the previous section, there are three problems that will be studied in this thesis.

First, we study bootstrap methods for general kernel tests for times series. We note, that many kernel tests have a test statistic with a particular structure: the Maximum Mean Discrepancy (MMD),Hilbert-Schmidt Independence Criterion (HSIC), the Lancaster interaction statistic, and kernel test of goodness of fit proposed in the chapter 4 each have empirical estimates which can be cast as normalized V-statistics,

$$\frac{1}{n^{m-1}} \sum_{1 \leq i_1, \ldots, i_m \leq n} h(Z_{i_1}, ..., Z_{i_m}),$$

where $Z_{i_1}, ..., Z_{i_m}$ are samples from a random process at the time points $\{i_1, \ldots, i_m\}$ and n is a sample size. We want to show that a method of external randomization known as the *wild bootstrap* Shao [93], Leucht and Neumann [67] may be applied to simulate from the null distribution. In particular Leucht and Neumann [67] shows that wild bootstrap mimics distribution of V-statistics of order two, i.e. when function h takes two arguments, well. Therefore the main effort will be to generalize the results of Leucht and Neumann [67] for arbitrary V-statistics, as all the kernel statistical test mentioned

above take this form.

Second, we aim to construct a goodness-of-fit test, based on the Stein operator applied to the RKHS functions, such that it only requires knowledge of an unnormalized density. To construct the test statistic, we will use a function class defined by the application of the the Stein operator to a specific space of RKHS functions, as proposed by Oates et al. [74], who addressed the problem of variance reduction in Monte Carlo integration. We want to provide a statistical tests for both uncorrelated and correlated samples, where the latter is essential if the test is to be used in assessing the quality of output of an MCMC procedure.

Third, we aim to construct a linear time complexity test based on a parsimonious representation of probability measures. Heuristics based on pseudo-distances, such as the difference between characteristic functions evaluated at a single frequency, have been studied in the context of goodness-of-fit tests [54, 55]. The objective is to improve on those inconsistent tests based on pseudo-distances between characteristic functions and propose an alternative, consistent test. To construct a test statistic we will exploit the properties of some of the infinite dimension representations of the measures and show that to distinguish between two measures it is sufficient to look a very low dimensional, randomly chosen subspace of the representation.

Contributions to Science

Wild Bootstrap for Degenerate Kernel Tests. The main result is showing that the wild bootstrap procedure yields consistent tests for time series, i.e., tests based on the wild bootstrap have a Type I error rate (of wrongly rejecting the null hypothesis) approaching the design parameter α, and a Type II error (of wrongly accepting the null) approaching zero, as the number of samples increases. We use this result to construct a two-sample test using MMD, an independence test using HSIC (resulting procedure is applied both to testing for instantaneous independence, and to testing for independence across multiple time lags) and the kernel goodness of fit test (discussed in the second chapter).

In brief, the arguments of the sum

$$\frac{1}{n^{m-1}} \sum_{1 \leq i_1, \ldots, i_m \leq n} h(Z_{i_1}, ..., Z_{i_m}),$$

are repeatedly multiplied by random, user-defined time series W_i, resulting in a bootstraped V-statistic

$$\frac{1}{n^{m-1}} \sum_{1 \leq i_1, \ldots, i_m \leq n} W_{i_1} W_{i_2} h(Z_{i_1}, \ldots, Z_{i_m}).$$

For a test of level α, the $1 - \alpha$ quantile of the empirical distribution obtained using these perturbed statistics serves as the test threshold. This approach has the important advantage over our preliminary work on shift bootstrap for HSIC that it may be applied to *all* kernel-based tests for which V-statistics are employed, and not just for independence tests. Additionally the consistency of the shift bootstrap can not be proved in general, unlike the consistency results for wild bootstrap.

In the section 3.4, we present a number of empirical comparisons: in the two sample case, we present a performance diagnostic for the output of a Gibbs sampler (the MCMC M.D.); in the independence case, we test for independence of two time series sharing a common variance (a characteristic of econometric models), and compare against the test of Besserve et al. [14] in the case where dependence may occur at multiple, potentially unknown lags. Our tests outperform both the naive approach which neglects the dependence structure within the samples, and the approach of Besserve et al. [14], when testing across multiple lags.

A Kernel Test of Goodness of Fit. The key contribution of this work is to define a statistical test of goodness-of-fit, based on a Stein discrepancy computed in a RKHS. To construct our test statistic, we apply the Stein operator to our chosen set of RKHS functions, and define our measure of goodness of fit as the largest discrepancy over this set between empirical sample expectations and target expectations (the latter being zero, due to the effect of the Stein operator). This approach is a natural extension to goodness-of-fit testing of the earlier two-sample tests [47] and independence tests [44] based on the maximum mean discrepancy, which is an integral probability metric.

As with these earlier tests, our statistic is a simple V-statistic, and can be computed in close form and in quadratic time; moreover, it is an unbiased estimate of the corresponding population discrepancy. As with all Stein-based discrepancies, only the gradient of the log-density of the target density is needed; we

do not require integrals with respect to the target density – including the normalisation constant. Given that our test statistic is a V-statistic, we may make use of the results from the Chapter 3 to provide statistical tests for both uncorrelated and correlated samples, where the latter is essential if the test is to be used in assessing the quality of output of an MCMC procedure. An identical test was obtained simultaneously in independent work by Liu et al. [69], for uncorrelated samples.

In our experiments, a particular focus is on applying our goodness-of-fit test to certify the output of approximate Markov Chain Monte Carlo (MCMC) samplers [63, 112, 6]. These methods use modifications to Markov transition kernels that improve mixing speed at the cost of worsening the asymptotic bias. The bias-variance trade-off can usually be tuned with parameters of the sampling algorithms. It is therefore important to test whether for a particular parameter setting and run-time, the samples are of the desired quality. This question cannot be answered with classical MCMC convergence statistics, such as the widely used potential scale reduction factor (R-factor) [39] or the effective sample size, since these assume that the Markov chain reaches its equilibrium distribution. By contrast, our test exactly quantifies the asymptotic bias of approximate MCMC.

In the section 4.3 we provide a number of experimental applications for our test. We begin with a simple check to establish correct test calibration on non-i.i.d. data, followed by a demonstration of statistical model criticism for Gaussian Process (GP) regression. We then apply the proposed test to quantify bias-variance trade-offs in MCMC procedures, and demonstrate how to use the test to verify whether MCMC samples are drawn from a stationary distribution.

Fast Analytic Functions Based Two Sample Test. We introduce two novel distance-like discrepancies between distributions, which both use a parsimonious representation of probability measures. The first discrepancy builds on the notion of differences in characteristic functions with the introduction of smooth characteristic functions, which can be thought of as the analytic analogues of the characteristics functions. A distance between smooth characteristic functions evaluated at a single random frequency is almost surely a distance (Definition 4 formalizes this concept) between these two distributions. In other words, there is no need to calculate the whole infinite dimensional representa-

tion - it is almost surely sufficient to evaluate it at a single random frequency (although checking more frequencies will generally result in more powerful tests) The second distance is based on analytic mean embeddings of two distributions in a characteristic RKHS; again, it is sufficient to evaluate the distance between mean embeddings at a single randomly chosen point to obtain a random variable that behaves almost surely like a distance. To our knowledge, this representation is the first mapping of the space of probability measures into a finite dimensional Euclidean space (in the simplest case, the real line) that is almost surely an injection. This injection is very appealing from a computational viewpoint, since the statistics based on it have linear time complexity (in the number of samples) and constant memory requirements.

In the section 5.2 we construct statistical tests based on empirical estimates of differences in the analytic representations of the two distributions. Our tests have a number of theoretical and computational advantages over previous approaches. The test based on differences between analytic mean embeddings is almost surely consistent for all distributions, and the test based on differences between smoothed characteristic functions is almost surely consistent for all distributions with integrable characteristic functions.

In the section A.4 we provide several experimental benchmarks for our tests. First, we compare test power as a function of computation time for two real-life testing settings: amplitude modulated audio samples, and the Higgs dataset, which are both challenging multivariate testing problems. Our tests give a better power/computation tradeoff than the characteristic function-based tests of Epps and Singleton [31], the previous sub-quadratic-time MMD tests Gretton et al. [48], Zaremba et al. [114], and the quadratic-time MMD test. In terms of power when unlimited computation time is available, we might expect worse performance for the new tests, in line with findings for linear- and sub-quadratic-time MMD-based tests [56, 45, 48, 114]. Remarkably, such a loss of power is not the rule: for instance, when distinguishing signatures of the Higgs boson from background noise [5] ('Higgs dataset'), we observe that a test based on differences in smoothed empirical characteristic functions outperforms the quadratic-time MMD. This is in contrast to linear- and sub-quadratic-time MMD-based tests, which by construction are less powerful than the quadratic-time MMD.

Structure of the Thesis

The thesis are divided into six chapters *Introduction*, *Background and Literature Review*, *Wild Bootstrap for Degenerate Kernel Tests*, *A Kernel Test of Goodness of Fit*, *Fast Analytic Functions Based Two Sample Test.* and *Conclusions and Future Work*.

In the *Background and Literature Review* chapter we review some useful statical concepts, namely: V statistics, mixing processes, dependent wild bootstrap, bootstrapped V statistics and kernel hypothesis tests. Then we discuss in more detail the previous work that is relevant to this thesis.

We begin our technical discussion in the third chapter (*Wild Bootstrap for Degenerate Kernel Tests*), in the section 3.1, where we establish a general consistency result for the wild bootstrap procedure on V-statistics, which we apply to MMD and to HSIC in Section 3.3. In the Section 3.4, we present a number of empirical comparisons, including experiments on financial time series and Markov chains convergance. The experiments concerning kernel goodness of fit test for time series are presented in the next chapter.

In the fourth chapter (*A Kernel Test of Goodness of Fit*) we begin our presentation in the section 4.1 with a high-level construction of the RKHS-based Stein discrepancy and associated statistical test. In section 4.2, we provide additional details and prove the main results. Section 4.3 contains experimental illustrations on synthetic examples, statistical model criticism, bias-variance trade-offs in approximate MCMC, and convergence in nonparametric density estimation.

In the fifth chapter (*Fast Analytic Functions Based Two Sample Test*) we first introduce two novel distances between distributions (5.1), which both use a parsimonious representation of the probability measures. We construct statistical tests in section 5.2, based on empirical estimates of differences in the analytic representations of the two distributions. We provide several experimental benchmarks (Section A.4) for our tests on various problems: audio samples, the Higgs dataset and artificial data high-dimensional distributions.

In the chapter *Conclusions and Future Work* we make some final remarks and discuss work that has been build upon this thesis as well as outline some further research directions.

Chapter 2

Background and Literature Review

In this chapter, we describe the application domain of the tests developed, outline modelling assumptions and discuss the relevant exiting results that are related to this thesis.

Application Domain

To put this thesis into a practical context we briefly comment on the application domain of statistical tests. Two-sample tests have been used e.g. in bioinformatics in comparing microarray data, in database matching and attempts have been made to use it for an unsupervised speaker verification. Independence tests have been used as a building component to algorithms that fit causal networks, to reveal features of data which might have high discriminative value, and in the independent component analysis. Goodness of fit tests are commonly used to verify whether residuals from regression, linear or non-linear, coincidence with the modelling assumptions. More generally they can be used for a statistical model criticism. In this thesis we show how to use them for the Markov Chain Monte Carlo algorithms convergence diagnostics.

Modeling techniques

We introduce necessary statistical models and tools.

Models of temporal dependence

First, we introduce the probabilistic tools which formalize notion of temporal dependence patterns. Let $(X_t, \mathcal{F}_t)_{t\in\mathbb{N}}$ be a strictly stationary sequence of random variables defined on a probability space Ω with a probability measure P

and natural filtration $\mathcal{F}_t$. Each X_t takes values in a real coordinate space $\mathbf{X}$ (the probability space is $(\mathbf{X}, \mathcal{B}(\mathbf{Z}), P_{\mathbf{X}})$, where $\mathcal{B}$ is Borel sigma algebra).

There are many ways to formalize the concept of the temporal dependence, all of which require some measure of similarity of random variables and notion of decay of the similarity as the distance in time grows. The simplest example is autocorrelation i.e.

$$Cov(X_t, X_0),$$

In this work we need more expressive notions of temporal dependence.

τ-mixing. The notion of τ-mixing is used to characterise weak dependence. It is a less restrictive alternative to classical mixing coefficients (which we discuss latter), and is covered in depth in [26]. For a random variable X on a probability space $(\Omega, \mathcal{F}, P_X)$ and $\mathcal{M} \subset \mathcal{F}$ we define

$$\tau(\mathcal{M}, X) = E\left(\sup_{g \in \Lambda} |Eg(X|\mathcal{M}) - Eg(X)|\right),$$

where Λ is the set of all one-Lipschitz continuous real-valued functions on the domain of X. A process is called τ-dependent if

$$\tau(r) = \sup_{l \in \mathbb{N}} \frac{1}{l} \sup_{r \leq i_1 \leq ... \leq i_l} \tau(\mathcal{F}_0, (X_{i_1}, ..., X_{i_l})) \overset{r \to \infty}{\longrightarrow} 0,$$

where r is a delay between $\mathcal{F}_0$ and random variables with indexes greater than r. $\tau(\mathcal{M}, X)$ can be interpreted as the minimal L_1 distance between X and X^* such that $X \overset{d}{=} X^*$ and X^* is independent of $\mathcal{M} \subset \mathcal{F}$. Furthermore, if $\mathcal{F}$ is rich enough, this X^* can be constructed [26, Lemma 5.3] .

We briefly discuss relation between classical strong mixing coefficient β and τ mixing.

Strong mixing coefficients. Strong mixing is historically the most studied type of temporal dependence – a lot of models, example being Markov Chains, are proved to be strongly mixing, therefore it's useful to relate weak mixing to strong mixing. For a random variable X on a probability space $(\Omega, \mathcal{F}, P_X)$ and $\mathcal{M} \subset \mathcal{F}$ we define

$$\beta(\mathcal{M}, \sigma(X)) = \| \sup_{A \in \mathbb{B}(R)} |P_{X|\mathcal{M}}(A) - P_X(A)| \|_1.$$

A process is called β-mixing or absolutely regular if

$$\beta(r) = \sup_{l \in \mathbb{N}} \frac{1}{l} \sup_{r \leq i_1 \leq ... \leq i_l} \beta(\mathcal{F}_0, (X_{i_1}, ..., X_{i_l})) \overset{r \to \infty}{\longrightarrow} 0,$$

[27, Equation 7.6] relates τ-mixing and β-mixing , as follows: if Q_x is the generalized inverse of the tail function

$$Q_x(u) = \inf_{t \in R} \{P(|X| > t) \leq u\},$$

then

$$\tau(\mathcal{M}, X) \leq 2 \int_0^{\beta(\mathcal{M}, \sigma(X))} Q_x(u) du.$$

While this definition can be hard to interpret, it can be simplified in the case $E|X|^p = M$ for some $p > 1$, since via Markov's inequality $P(|X| > t) \leq \frac{M}{t^p}$, and thus $\frac{M}{t^p} \leq u$ implies $P(|X| > t) \leq u$. Therefore $Q'(u) = \frac{M}{\sqrt[p]{u}} \geq Q_x(u)$. As a result, we have the following inequality

$$\frac{\sqrt[p]{\beta(\mathcal{M}, \sigma(X))}}{M} \geq C\tau(\mathcal{M}, X) \tag{2.1}$$

Models that satisfy τ-mixing. Dedecker and Prieur [27] provides examples of systems that are tau-mixing. In particular, given that certain assumptions are satisfied causal functions of stationary sequences, iterated random functions, Markov chains, expanding maps are all τ-mixing.

Of particular interest to this work are Markov chains. The assumptions provided by Dedecker and Prieur [27], under which Markov chains are tau-mixing are somehow difficult to check but we can use classical theorems about the β-mixing). In particular [17, Corollary 3.6] states that a Harris recurrent (chain returns to a fixed set of the state space an infinite number of times) and aperiodic Markov chain satisfies absolute regularity. [17, Theorem 3.7] states that geometric ergodicity [1] implies geometric decay of the β coefficient. Interestingly [17, Theorem 3.3] describes situations in which a non-stationary chain β-mixes exponentially.

Using inequality 2.1 between τ-mixing coefficient and strong mixing coeffi-

[1] $\forall_x \|P^n(x, \cdot) - \pi\|_{TV} \leq Cq^n, 0 < q < 1$

cients one can use those classical theorems show that e.g for $p = 2$ we have

$$\sqrt{\beta(\mathcal{M}, \sigma(X))} \geq \tau(\mathcal{M}, X).$$

V-statistics.

V statistics appear in the context of statistical testing, whenever a property of a probability measure p can be expressed as an expected value with respect to k independent random variables $X_1, \cdots, X_k \sim p$. For instance, variance of a measure p is given as the expected value of the function $h(x, y) = (x - y)^2$

$$Eh(X_1, X_2) \text{ where } X_1, X_2 \sim p.$$

An estimate of variance is given by a V-statistic of order 2, namely

$$\frac{1}{n^2} \sum_{i,j=1}^{n} h(X_i, X_j).$$

In general, a V-statistic [90, Section 5.1.5] of a k-argument, symmetric function h is written

$$V_n(h) = \frac{1}{n^m} \sum\nolimits_{i \in N^m} h(X_{i_1}, ..., X_{i_m}), \tag{2.2}$$

where N^m is a Cartesian power of a set $N = \{1, ..., n\}$. For simplicity, from time to time, we will drop the argument and simply write V or, V_n if we need to emphasize an aspects of the V-statistic that depends on n or, $V_n(h)$ if we want to emphasize an aspect of the V-statistic that depends on h. We will refer to the function h as to the *core* of the V-statistic $V_n(h)$. While such functions are usually called kernels in the literature, in this work we reserve the term kernel for positive-definite functions taking two arguments. The asymptotic behaviour of V-statistic depends on the degeneracy of the core. We say that a k-argument, symmetric function h is j-degenerate ($j < k$) if for each $x_1, \cdots, x_j \in \mathbf{X}$,

$$Eh(x_1, \cdots, x_j, X_{j+1}, \cdots, X_k) = 0,$$

where $X_{j+1}, \cdots, X_k$ are i.i.d. random variables. If $j = k - 1$ we say that the function is canonical.

As shown in [90, Section 5.1.5], using so called Hoeffding decomposition, any

core h can be written as a sum of canonical cores $h_1, ..., h_m$ and a constant h_0

$$h(x_1,...,x_m) = h_m(x_1,...,x_m) + \sum_{1\le i_1<...<i_{m-1}\le m} h_{m-1}(x_{i_1},...,x_{i_{m-1}}) + ... + \sum_{1\le i_1<i_2\le m} h_2(x_{i_1},x_{i_2}) + \sum_{1\le i\le m} h_1(x_i) + h_0$$

We call $h_0, ..., h_m$ components of a core h. The components are defined in terms of auxiliary functions g_c

$$g_c(x_1,...x_c) = Eh(x_1,...,x_c,X_{c+1},...,X_m), \tag{2.3}$$

where $X_1, \cdots, X_c$ are i.i.d. random variables, for each $c = 0, ..., m-1$ and we put $g_m = h$. We define components as follows

$$\begin{aligned}
h_0 &= g_0,\\
h_1(x_1) &= g_1(x_1) - h_0,\\
h_2(x_1,x_2) &= g_2(x_1,x_2) - h_1(x_1) - h_1(x_2) - h_0,\\
h_3(x_1,x_2,x_3) &= g_3(x_1,x_2,x_3) - \sum_{1\le i<j\le 3} h_2(x_i,x_j) - \sum_{1\le i\le 3} h_1(x_i) - h_0,\\
&\cdots,\\
h_m(x_1,...,x_m) &= g_m(x_1,...,x_m) - \sum_{1\le i_1<...<i_{m-1}\le m} h_{m-1}(x_{i_1},...,x_{i_{m-1}})\\
&\quad - ... - \sum_{1\le i_1<i_2\le m} h_2(x_{i_1},x_{i_2}) - \sum_{1\le i\le m} h_1(x_i) - h_0.
\end{aligned}$$

[90, Section 5.1.5] shows that components h_c are symmetric (and therefore cores) and canonical. Finally a V-statistic of a core function h can be written as a sum of V-statistics with canonical cores

$$V_n(h) = V_n(h_m) + \binom{m}{1} V_n(h_{m-1}) + ... + \binom{m}{m-2} V_n(h_2) + \binom{m}{m-1} V_n(h_1) + h_0.$$

Note that for a one-degenerate core h, the constant h_0 and the first component

h_1 are identically equal to zero

$$h_0 = g_0 = Eh(X_1, X_2, ..., X_m) = 0,$$
$$h_1(x_1) = g_1(x_1) - h_0 = Eh(x_1, X_2, X_3, ..., X_m) = 0.$$

In this case it turns out that the second component of the core h_2 is the one that governs the asymptotic distribution of the V-statistic. We say that a V-statistic with a one-degenerate core is a degenerate V-statistic and that nV_n is a normalized V-statistic. This type of degeneracy is common to many kernel statistics, when the null hypothesis holds [47, 44, 87].

Finally, we remark that asymptotic distribution of the degenerate, normalized V-statistic is known to be an infinite weighted sum of χ^2-variables [90, Section 5.4], where the weights depend on the core and random variables in the V-statistic. To our knowledge there is no close form expression for calculating quantiles of this limiting distributions and so a suitable bootstrap is used to obtain the empirical quantiles.

An Introduction to the Wild Bootstrap

Bootstrap methods aim to evaluate the accuracy of the sample estimates - they are particularly useful when dealing with complicated distributions, or when the assumptions of a parametric procedure are in doubt. Bootstrap methods randomize the dataset used for the sample estimate calculation, so that a new dataset with a similar statistical properties is obtained, e.g. one popular method is resampling. In the wild bootstrap method the observations in the dataset are multiplied by appropriate random numbers. To present the idea behind the wild bootstrap we will discuss a toy example similar to that in Shao [93], and then relate it to the wild bootstrap method used in this thesis.

Consider the following autoregressive model

$$X_t = aX_{t-1} + C + \epsilon_t$$

where ϵ_t are i.i.d. Student's t random variables and $|a| < 1$, C are unknown constants. We wish to test if $C = 0$. By [72, Theorem 1] sequence X_t is geometrically strongly mixing, which means that $\beta(m) = O(\exp^{-dm})$ for some

positive d. By [18, Theorem 0] if $C = 0$, the normalized sample mean of the process X_t has asymptotically normal distribution

$$\frac{\sum_{i=1}^{N} X_i}{\sqrt{n}} \xrightarrow{d} N(0, \sigma_\infty^2),$$

where $\sigma_\infty^2 = \sum_{j=-\infty}^{j=\infty} cov(X_0, X_j)$. We note here that using $cov(X_0, X_0)$ instead of σ_∞^2 in the testing procedure based on the asymptotic distribution $N(0, \sigma_\infty^2)$ would result in too conservative test. The variance σ_∞^2 is not easy to estimate. Using the wild bootstrap method we will construct process Y_t that mimics behaviour of the X_t process and use it to approximate the distribution of the normalized sample mean, $\frac{\sum_{i=1}^{N} X_i}{\sqrt{n}}$. Consider a triangular array of autoregressive processes, starting from $W_{1,n} = N(0, 1)$,

$$W_{t,n} = e^{-1/l_n} W_{t-1,n} + \sqrt{1 - e^{-2/l_n}} \varepsilon_{t,n}$$

where $\varepsilon_{t,n}$ are i.i.d. standard normal random variables. We call $W_{t,n}$ a wild bootstrap process.

Define $Y_{t,n} = W_{t,n} X_t$.

We need to show that the distribution of the normalized sample mean of the process $Y_{t,n}$, mimics the distribution $N(0, \sigma_\infty^2)$. For that is is necessary that the expected value and correlations of Y_t and X_t agree:

$$EY_{t,n} = EW_{t,n} X_t = 0,$$
$$cov(Y_{0,n}, Y_{t,n}) = cov(X_0, X_t) cov(W_{0,n}, W_{t,n}) = cov(X_0, X_t) e^{-t/l_n}$$

The auto-covariance structure of the process Y_t is similar to the auto-covariance structure of the process X_t. Indeed if we let l_n grow with n, we recover the same covariance structure

$$cov(Y_{0,n}, Y_{t,n}) = cov(X_0, X_t) e^{-t/l_n} \to cov(X_0, X_t),$$

and so we expect that

$$\frac{\sum_{i=1}^{N} Y_i}{\sqrt{n}} \xrightarrow{d} N(0, \sigma_\infty). \tag{2.4}$$

This has been proved in the Leucht and Neumann [67, Theorem 6.1]. This central limit theorem was used in Leucht and Neumann [67] to study some

normalized V-statistic. Suppose that h is a positive definite, degenerate kernel which allows finite Mercer decomposition [13, Theorem 40] with respect to distribution of X i.e.

$$h(x,y) = \sum_{k}^{M} \lambda_k \phi_k(x)\phi_k(y),$$
$$\lambda_k \phi_k(x) = Eh(x,Y)\phi_k(Y)$$

where $M < \infty$. In this case a V-statistic of order two, can be written as

$$\sum_{k=0}^{M} \lambda_k \left(\frac{\sum_{i=1}^{n} \phi_k(X_i)}{\sqrt{n}} \right)^2 = \frac{1}{n} \sum_{1 \le i,j \le n} h(X_i, X_j)$$

where λ_k are eigenvalues and ϕ_k are eigenfunction of the kernel h, respectively. Since

$$E\phi_k(X) = Eh(X,Y)\phi_k(Y) = E\big(\phi_k(Y)[Eh(X,Y)|Y]\big) = E[\phi_k(Y)\cdot 0|Y] = 0.$$

one may replace

$$\frac{\sum_{i=1}^{n} \phi_k(X_i)}{\sqrt{n}}$$

with a bootstrapped version

$$\frac{\sum_{i=1}^{n} W_t^n \phi_k(X_i)}{\sqrt{n}},$$

and conclude, as in the toy example, that the limiting distribution of the single component of the sum $\sum_k \lambda_k ...$ remains the same. One of the main contributions of Leucht and Neumann is in showing that the distribution of the whole sum $\sum_k \lambda_k \left(\frac{\sum_{i=1}^{n} W_t^n \phi_k(X_i)}{\sqrt{n}} \right)^2$, even with $M = \infty$, converges to the same distribution as the normalized V-statistic, $nV_n(h)$.

Bootstrapped V-statistic.

As noted above the quantiles of the asymptotic distribution of V-statistics appear to be not tractable and so the common strategy is to estimate them using some sort of bootstrap. In this work we will study two versions of the boot-

strapped V-statistics

$$B_1(h)_n = \frac{1}{n^m} \sum\nolimits_{i \in N^m} W_{i_1,n} W_{i_2,n} h(X_{i_1}, ..., X_{i_m}), \tag{2.5}$$

$$B_2(h)_n = \frac{1}{n^m} \sum\nolimits_{i \in N^m} \tilde{W}_{i_1,n} \tilde{W}_{i_2,n} h(X_{i_1}, ..., X_{i_m}), \tag{2.6}$$

where $\{W_{t,n}\}_{1 \leq t \leq n}$ is an auxiliary wild bootstrap process and $\tilde{W}_{t,n} = W_{t,n} - \frac{1}{n}\sum_{j=1}^{n} W_{j,n}$. This auxiliary process, proposed by [93, 67], satisfies the following assumption:

Definition 1 (Bootstrap process assumptions)**.**

$$\{W_{t,n}\}_{1 \leq t \leq n}$$

is a row-wise strictly stationary triangular array independent of all X_t such that $EW_{t,n} = 0$ and $\sup_n E|W_{t,n}^4| < \infty$. The autocovariance of the process is given by $EW_{s,n}W_{t,n} = \rho(|s-t|/l_n)$ for some non-negative function ρ, such that $\lim_{u \to 0} \rho(u) = 1$ and $\sum_{r=1}^{n-1} \rho(|r|/l_n) = O(l_n)$. The sequence $\{l_n\}$ is taken such that $l_n = o(\sqrt{n})$ but $\lim_{n \to \infty} l_n = \infty$. The variables $W_{t,n}$ are τ-weakly dependent with coefficients $\tau(r) \leq C\zeta^{\frac{r}{l_n}}$ for $r = 1, ..., n$, $\zeta \in (0,1)$ and $C \in \mathbb{R}$.

As noted in in [67, Remark 2], a simple realization of a process (introduced in the previous section) that satisfies this assumption is $W_{t,n} = e^{-1/l_n} W_{t-1,n} + \sqrt{1 - e^{-2/l_n}}\epsilon_t$ where $W_{0,n}$ and $\epsilon_1, \ldots, \epsilon_n$ are independent standard normal random variables. For simplicity, we will drop the index n and write W_t instead of $W_{t,n}$. A process that fulfils the *bootstrap assumption* will be called *bootstrap process*.

The versions of the bootstrapped V-statistics in (2.5) and (2.6) were previously studied in Leucht and Neumann [67] for the case of canonical cores of degree $m = 2$. We extend their results to higher degree cores (common within the kernel testing framework), which are not necessarily one-degenerate. When stating a fact that applies to both B_1 and B_2, we will simply write B, and the argument h or index n will be dropped when there is no ambiguity.

Kernel Embedding Tests

For every symmetric, positive definite function, i.e., *kernel* $k : \mathbf{E} \times \mathbf{E} \to \mathbb{R}$, there is an associated reproducing kernel Hilbert space $\mathcal{H}$ [12, p. 19]. The kernel

embedding of a probability measure P on $\mathbf{E}$ is an element $\mu(P) \in \mathcal{H}$, given by $\mu_p = \int_E k(\cdot, x)\, dP(x)$ [12, 95]. If a measurable kernel k is bounded, the mean embedding μ_P exists for all probability measures on $\mathbf{E}$, and for many interesting bounded kernels k, including the Gaussian, Laplacian and inverse multi-quadratics, the kernel embedding $P \mapsto \mu_P$ is injective. Such kernels are said to be *characteristic* [97].

The Maximum Mean Discrepancy (MMD) [45] is defined as

$$\text{MMD}(P, Q) = \sup_{f \in B_k} \left[\int_{\mathbf{X}} f dP - \int_{\mathbf{X}} f dQ \right], \tag{2.7}$$

where P and Q are probability measures on $\mathbf{X}$, and B_k is the unit ball in the RKHS $\mathcal{H}$ associated with a positive definite kernel k. It can be shown that the MMD is equal to the RKHS distance between so called mean embeddings,

$$\text{MMD}(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}}, \tag{2.8}$$

where $\|\cdot\|_{\mathcal{H}}$ denotes the norm in the RKHS $\mathcal{H}$. When k is translation invariant, i.e., $k(x, y) = \kappa(x - y)$, the squared MMD can be written [97, Corollary 4]

$$\text{MMD}^2(P, Q) = \int_{\mathbf{R}^d} |\varphi_P(t) - \varphi_Q(t)|^2 F^{-1}\kappa(t) dt, \tag{2.9}$$

where F denotes the Fourier transform, F^{-1} is the inverse Fourier transform, and φ_P, φ_Q are the characteristic functions of P, Q, respectively. From [97, Theorem 9], if the kernel k is characteristic then MMD satisfies

$$\text{MMD}(P, Q) = 0 \text{ iff } P = Q. \tag{2.10}$$

Any bounded, continuous, translation-invariant kernel whose inverse Fourier transform is almost everywhere non-zero is characteristic [97]. By representation (2.11), it is clear that the MMD with a characteristic kernel is a metric. Finally, MMD can expressed in terms of expected value, let $z_i = x_i, y_i$,

$$\begin{aligned} \text{MMD}(P, Q) &= Eh(Z_1, Z_2) \quad \text{where } Z_1, Z_2 \sim (P, Q) \text{ and,} \\ h(z_1, z_2) &= k(x_1, x_2) - k(x_1, y_2) - k(x_2, y_1) + k(y_1, y_2). \end{aligned}$$

The Hilbert Schmidt Independence Criterion (HSIC) [43, 44], measure of sta-

tistical dependence, is simply MMD between join distribution and product of marginals. Given a measure P_{XY} over a product space, HSIC is defined

$$\text{HSIC}(P_{XY}) = \|\mu_{P_{XY}} - \mu_{P_X} \otimes \mu_{P_Y}\|_{\mathcal{H}}. \tag{2.11}$$

Again, if mean embedding are injective, HSIC is zero if and only if $P_{X,Y}$ is a product measure. Not surprisingly, Hilbert-Schmidt Independence Criterion (HSIC), can be expressed in terms of expectations of RKHS kernels [43, 44]

$$\text{HSIC}(P_{XY}) = Ek(X_1, X_2)(l(Y_1, Y_2) - 2l(Y_1, Y_3) + l(Y_3, Y_4)),$$

where $X_i, Y_i \sim P_{XY}$. Denote a group of permutations over 4 elements by S_4, with π one of its elements, i.e., a permutation of four elements. Let $z_i = x_i, y_i$ and define a symmetric function h

$$\begin{aligned} h(z_1, z_2, z_3, z_4) = \frac{1}{4!} \sum_{\pi \in S_4} k(x_{\pi(1)}, x_{\pi(2)})[l(y_{\pi(1)}, y_{\pi(2)}) \\ + l(y_{\pi(3)}, y_{\pi(4)}) - 2l(y_{\pi(2)}, y_{\pi(3)})]. \end{aligned} \tag{2.12}$$

The V-static $V_n(h)$ is an estimator of HSIC. This estimator can be also written as $\frac{1}{n^2}\text{Tr}(KHLH)$ ([43]) for kernel matrices $K_{i,j} = k(X_i, X_j)$ and $L_{i,j} = l(Y_i, Y_j)$ and the centering matrix $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$.

Related work

Wild Bootstrap for Degenerate Kernel Tests

Independence testing. Prior work on testing independence in time series may be categorized in two branches: testing serial dependence within a single time series, and testing dependence between one time series and another. The case of serial dependence turns out to be relatively straightforward, as under the null hypothesis, the samples become independent: thus, the analysis reduces to the i.i.d. case. Pinkse [77], Diks and Panchenko [29] provide a quadratic forms function-based serial dependence test which employs the same statistic as HSIC. Due to the simple form of the null hypothesis, the analysis of [91, Ch. 5] applies. Further work in the context of the serial dependency testing includes simple approaches based on rank statistics e.g. Spearman's correlation or Kendall's tau, correlation integrals e.g. [19]; criteria based on integrated

squared distance between densities e.g [84]; KL-divergence based criteria e.g. [83, 58]; and generalizations of KL-divergence to so called q-class entropies e.g. [41, 78].

In most of the tests of independence of two time series, specific conditions have been enforced, e.g that processes follow a moving average specification or the dependence is linear. Prior work in the context of dependency tests of two time series includes cross covariance based tests e.g. [53, 57, 92]; and a Generalized Association Measure based criterion [32]. Some work has been undertaken in the nonparametric case, however. A nonparametric measure of independence for time series, based on the Hilbert-Schmidt Independence Criterion, was proposed by Zhang et al. [116]. While this work established the convergence in probability of the statistic to its population value, no asymptotic distributions were obtained, and the statistic was not used in hypothesis testing. To our knowledge, the only nonparametric independence test for pairs of time series is due to Besserve et al. [14], which addresses the harder problem of testing independence across all time lags simultaneously. [2] The procedure is to compute the Hilbert-Schmidt norm of a cross-spectral density operator (the Fourier transform of the covariance operator at each time lag). The resulting statistic is a function of frequency, and must be zero at all frequencies for independence, so a correction for multiple hypothesis testing is required. It is not clear how the asymptotic analysis used in the present work would apply to this statistic, and this remains an interesting topic of future study.

Two sample testing. A two-sample test for τ-dependent time series was proposed in Leucht [66][Section 4.2]. It is a variation on the Cramer–von Mises type of test, which has a core that is a positive definite kernel and therefore the test can be thought of as a mean discrepancy test (in the Chapter 3 we provide general MMD test for time series).

In a slightly different setting, two sample problem of equality of intensities of Poisson processes was studied in [36]. While the statistical analysis involved in that work is somehow similar to ours, the formulation of the problem is different and results are not equivalent.

[2] Let X_t follow a MA(2) model and put $Y_t = X_{t-20}$. This is a case addressed by Besserve et al. [14], who will reject their null hypothesis, whereas our null is accepted

A Kernel Test of Goodness of Fit.

Several alternative approaches exist in the statistics literature to goodness-of-fit testing. A first strategy is to partition the space, and to conduct the test on a histogram estimate of the distribution [9, 11, 51, 50]. Such space partitioning approaches can have attractive theoretical properties (e.g. distribution-free test thresholds) and work well in low dimensions, however they are much less powerful than alternatives once the dimensionality increases [42]. Similar approach are testing procedures based on L_2 distance between target density and samples, an [35].

Another popular approach has been to use the smoothed L_2 distance between the empirical characteristic function of the sample, and the characteristic function of the target density. This dates back to the test of Gaussianity of Baringhaus and Henze [8], who used a squared exponential smoothing function (see Eq. 2.1 in their paper). Fan and Ullah proposed similar test for weakly depended observations. For this choice of smoothing function, their statistic is identical to the maximum mean discrepancy (MMD) with the squared exponential kernel, which can be shown using the Bochner representation of the kernel (compare with Sriperumbudur et al. 97, Corollary 4). It is essential in this case that the target distribution be Gaussian, since the convolution with the kernel (or in the Fourier domain, the smoothing function) must be available in close form. An L_2 distance between Parzen window estimates can also be used [16], giving the same expression again, although the optimal choice of bandwidth for consistent Parzen window estimates may not be a good choice for testing [1]. A different smoothing scheme in the frequency domain results in an energy distance statistic [this likewise being an MMD with a particular choice of kernel; see 88], which can be used in a test of normality [109]. The key point is that the required integrals are again computable in closed form for the Gaussian, although the reasoning may be extended to certain other families of interest, e.g. [82]. The requirement of computing closed-form integrals with respect to the test distribution severely restricts this testing strategy. Finally, a problem related to goodness-of-fit testing is that of model criticism [70]. In this setting, samples generated from a fitted model are compared via the maximum mean discrepancy with samples used to train the model, such that a small MMD indicates a good fit. There are two limitation to the method: first, it requires samples from the model (which might not be easy if this requires a

complex MCMC sampler); second, the choice of number of samples from the model is not obvious, since too few samples cause a loss in test power, and too many are computationally wasteful. Neither issue arises in the test proposed in the chapter 4, since we do not require model samples.

Fast Analytic Functions Based Two Sample Test.

The earliest work in the linear time two-sample tests includes the point wise difference between characteristic functions [54, 55]. It was shown that the power of such tests can be maximized against fully specified alternative hypotheses, where test power is the probability of correctly rejecting the null hypothesis that the distributions are the same. In other words, if the class of distributions being distinguished is known in advance, then the tests can focus only at those particular frequencies where the characteristic functions differ most. This approach was generalized to evaluating the empirical characteristic functions at multiple distinct frequencies by [31], thus improving on tests that need to know the single "best" frequency in advance (the cost remains linear in the sample size, albeit with a larger constant). This approach still fails to solve the consistency problem, however: two distinct characteristic functions can agree on an interval, and if the tested frequencies fall in that interval, the distributions will be indistinguishable.

Another alternative to the quadratic-time MMD test is a B-test [114] (block-based test): the idea is to break the data into blocks, compute a quadratic-time statistic on each block, and average these quantities to obtain the test statistic. The B-test is a variation of the MMD test in which the complexity can controlled by choosing the size of blocks used to calculate the test statistics. At one extreme is the linear-time MMD suggested by [45, 48] where we have $n/2$ blocks of size $B = 2$, and at the other extreme is the usual full MMD with 1 block of size n, which requires calculating the test statistic on the whole kernel matrix in quadratic time.

Chapter 3

Wild Bootstrap for Degenerate Kernel Tests

> This chapter is based on Kacper Chwialkowski, Dino Sejdinovic, and Arthur Gretton. A wild bootstrap for degenerate kernel tests. In *Advances in Neural Information Processing Systems 27*, pages 3608–3616. Curran Associates, Inc., 2014.

In this chapter a wild bootstrap method for nonparametric hypothesis tests based on kernel distribution embeddings is proposed. This bootstrap method is used to construct provably consistent tests that apply to random processes, for which other bootstraps methods are not directly applicable. It applies to a large group of kernel tests based on V-statistics, which are degenerate under the null hypothesis, and non-degenerate elsewhere. To illustrate this approach, we construct a two-sample test, an instantaneous independence test and a multiple lag independence test for time series. We also use the results of this chapter in construction of a kernel goodness of fit test. In experiments, the wild bootstrap gives strong performance on synthetic examples and in performance benchmarking for the Gibbs sampler.

The main results of the chapter are based around two concepts: τ-mixing, which describes the dependence within the time series, and V-statistics, which constitute our test statistics. These topics were reviewed in the Section 2.2.

Asymptotic distribution of wild bootstrapped V statistics

In this section, we present main results that describe asymptotic behaviour of V-statistics and bootstraped V-statistics for random processes. While this section aims to communicate the main ideas, we provide details and proofs in the Section 3.2. In the Section 3.3, these results will be used to construct

kernel-based statistical tests applicable to dependent observations. Tests are constructed so that the V-statistic is degenerate under the null hypothesis and non-degenerate under the alternative. Theorem 1 guarantees that the bootstrapped V-statistic will converge to the same limiting null distribution as the V-statistic. Following Leucht and Neumann [67], we will establish the convergence of the bootstrapped distribution to the desired asymptotic distribution. Throughout this chapter we will make one mild assumption

$$\sup_{i \in N^m} Eh(Z_i)^2 < \infty,$$

where $Z_i = (Z_{i_1}, \cdots Z_{i_m})$. This assumption is almost always automatically satisfied, since most of the kernels used in practice are bounded.

Theorem 1. *Assume that the stationary process Z_t is τ-dependent with $\sum_{r=1}^{\infty} r^2 \sqrt{\tau(r)} < \infty$. If the core h is a Lipschitz continuous, one-degenerate and its h_2-component is a positive definite kernel, such that $Eh_2(Z_0, Z_0) < \infty$, then nB_n (2.5), (2.6), and nV_n (2.2) converge weakly to the same distribution V. Moreover $nB_n(h_2)$ and $nV_n(h_2)$ converge weakly to $\binom{m}{2}^{-1} V$.*

On the other hand, if the V-statistic is not degenerate, which is usually true under the alternative, it converges to some non-zero constant.

Theorem 2. *Assume that the stationary process Z_t is τ-dependent with $\tau(r) = o(r^{-4})$. If the core h is a Lipschitz continuous, and h_0 component is positive then V_n converges in mean squared to h_0.*

In this setting, Theorem 3 guarantees that the bootstrapped V-statistic will converge to zero in probability. This property is necessary in testing, as it implies that the test thresholds computed using the bootstrapped V-statistics will also converge to zero, and so will the corresponding Type II error.

Theorem 3. *Assume that the stationary process $\{Z_t\}$ is τ-dependent with a coefficient $\tau(r) = o(r^{-4})$. If the core h is a function of $m > 1$ arguments then $B_1(h)$ and $o(n)B_2(h)$ converge to zero in mean squared.*

Although both B_2 and B_1 converge to zero, the rate does not seem to be that same. As a consequence, tests that utilize B_2 usually give lower Type II error than the ones that use B_1. On the other hand, B_1 seems to better approximate

V-statistic distribution under the null hypothesis. This agrees with our experiments in Section 3.4 as well as with those in [67, Section 5]). These results are sufficient for adopting kernel tests developed for i.i.d. data to tests that work on random processes. In particular Theorem 1 justifies usage of bootstraped V-statistics for estimating quantiles of the null distribution, while Theorems 23 guarantee consistency.

The general testing procedure is

- Calculate the test statistic $nV_n(h)$.
- Obtain wild bootstrap samples $\{B_n(h)\}_{i=1}^D$ and estimate the $1-\alpha$ empirical quantile of these samples.
- If $nV_n(h)$ exceeds the quantile, reject.

Proofs

In this section we prove the main theorems. As for the notation, n denotes number of observations, $N = \{1, \cdots, n\}$, if h is function then $h \times h$ denotes a product of h with itself, $\lim_{n\to\infty} X_n \overset{L_2}{=} X$ denotes convergence in mean square

Proof of the Theorem 1

Hoeffding decomposition reduces any V-statistic to a sum of canonical V-statistics with canonical cores h_c, which are easier to study in context of non-iid data. As an illustration, consider a canonical core h of m arguments and fix some indexes $i_1 \leq \cdots \leq i_{m-1} \ll i_m$, for a sake of example we may assume that indexes represent time. If observations $Z_{i_1}, \cdots, Z_{i_{m-1}}$ are independent of the observation Z_{i_m}, then the expected value of $h(Z_{i_1}, \cdots, Z_{i_m})$, by degeneracy, is equal to zero. If it is reasonable to assume that Z_{i_m} is almost independent of $Z_{i_1}, \cdots, Z_{i_{m-1}}$, maybe because it is so distant in time, then it is also reasonable to expect that for a canonical core h (which is not too complicated)

$$Eh(Z_{i_1}, \cdot, Z_{i_m}) \approx 0.$$

which follows from the following approximate calculation

$$\int h(z_{i_1}, \cdot, z_{i_m}) dP_{Z_{i_1}, \cdots, Z_{i_m}} \approx \int h(z_{i_1}, \cdots, z_{i_m}) dP_{Z_{i_1}, \cdots, Z_{i_{m-1}}} dP_{Z_{i_m}} = 0$$

We formalize this intuition.

Definition 2. *Associate with any set of indexes $i_1, \cdots, i_m$ its nearest neighbor within the set. Suppose i_r is is an index with the most distant nearest neighbor. We will call i_r the most isolated index, and we will refer to its distance to the nearest neighbor as an isolation distance.*

Consider a following example, for the set $\{1, 5, 7\}$, 1 is the most isolated index and the isolation distance is 4.

Definition 3. *Given a sequence of random variables Z_t and a function h, if for all sets of indexes $i_1, \cdots, i_m$, with the isolation distance equal to r*

$$|Eh(Z_{i_1}, \cdots, Z_{i_m})| \leq \Delta(h, r)$$

for a function Δ, then we say that the pair (h, Z_t) is of type Δ.

The next theorem shows a growth rate of a canonical V-statistic when a pair h, Z_t is of type Δ.

Theorem 4. *Let (Z_t, h), where h is a function of $m > 1$ arguments, be of type Δ, with $\Delta(h, r) = o(r^{-k})$ for some k, then*

$$\sum_{i \in N^m} |Eh(Z_i)| = O\left(n^{\lfloor \frac{m}{2} \rfloor}\right) + o\left(n^{2\lfloor \frac{m}{2} \rfloor + 2 - k}\right).$$

Proof. The proof uses a technique similar to [4, Lemma 3]. We will focus on ordered m-tuples $1 \leq i_1 \leq \ldots \leq i_m \leq n$, and by considering all possible permutations of their indices, we obtain an upper bound

$$\sum_{i \in N^m} |Eh\left(Z_{i_1}, \ldots, Z_{i_m}\right)| < \sum_{1 \leq i_1 \leq \ldots \leq i_m \leq n} \sum_{\pi \in S_m} \left| Eh\left(Z_{i_{\pi(1)}}, \ldots, Z_{i_{\pi(m)}}\right)\right|,$$

where (strict) inequality stems from the fact that the m-tuples with some coinciding entries appear multiple times on the right.

Since (h, Z_t) is a of type Δ

$$\forall i \in N^m \sum_{\pi \in S_m} \left| Eh\left(Z_{i_{\pi(1)}}, \ldots, Z_{i_{\pi(m)}}\right)\right| = O(\Delta(h, w(i))),$$

where $w(i)$ is an isolating distance of the index set $i = i_1, \cdots i_m$. We need to estimate order of the sum

$$\sum_{1 \le i_1 \le \ldots \le i_m \le n} O(\Delta(h, w(i))).$$

Let us upper bound the number of ordered m-tuples i with $w(i) = w$. Denote $s = \lfloor \frac{m}{2} \rfloor + 1$. i_1 can take n different values, but since $i_2 \le i_1 + w$, i_2 can take at most $w+1$ different values. For $2 \le l \le s-1$, since $\min\{i_{2l} - i_{2l-1}, i_{2l-1} - i_{2l-2}\} \le w$, we can either let i_{2l-1} take up to n different values and let i_{2l} take up to $w+1$ different values (if $i_{2l} - i_{2l-1} \le i_{2l-1} - i_{2l-2}$) or let i_{2l-1} take up to $w+1$ different values and let i_{2l} take up to n different values (if $i_{2l} - i_{2l-1} > i_{2l-1} - i_{2l-2}$), upper bounding the total number of choices for $[i_{2l-1}, i_{2l}]$ by $2n(w+1)$. Finally, the last term i_m can always have at most $w+1$ different values. This brings the total number of m-tuples with $w(i) = w$ to at most $2^{s-2} n^{s-1} (w+1)^s$. Thus, the number of m-tuples with $w(i) = 0$ is $O(n^{s-1})$ and since $Eh(Z_{i_1}, \ldots, Z_{i_m}) < \infty$, we have

$$\begin{aligned}
&\sum_{1 \le i_1 \le \ldots \le i_m \le n} O(\Delta(h, w(i))) \\
&\le O(n^{s-1}) + \sum_{w=1}^{n-1} \sum_{\substack{1 \le i_1 \le \ldots \le i_m \le n: \\ w(i) = w}} O(\Delta(h, w(i))) \\
&\le O(n^{s-1}) + n^{s-1} \sum_{w=1}^{n-1} (w+1)^s O(\Delta(h, w)) \\
&\le O(n^{s-1}) + n^{s-1} \sum_{w=1}^{n-1} o(w^{s-k}) \\
&\le O(n^{s-1}) + n^{s-1} \max(o(n^{s-k+1}), O(1)) \\
&\le O(n^{s-1}) + o(n^{2s-k}) + O(n^{s-1})) \\
&= O(n^{s-1}) + o(n^{2s-k}),
\end{aligned}$$

which proves the claim. We have used $\Delta(h, w) = o(w^{-k})$. □

The previous theorem states sufficient conditions for a V-statistic or a bootstrapped V-statistic to converge to zero.

Lemma 1. *Let h be a function of $m > 1$ arguments and let $(\{Z_t\}_{t \in N}, h \times h)$ be a of type Δ, with $\Delta(h \times h, r) = o(r^{-4})$. If $\{G_i\}_{i \in N}$ is a random process, independent*

of Z_t, such that $\sup_i EG_i^4 < \infty$, with notation $T_n = \frac{1}{n^{m-1}} \sum_{i \in N^m} G_{i_1} G_{i_2} h(Z_i)$,

$$\begin{cases} \lim_{n\to\infty} o(1) T_n \overset{L_2}{=} 0 & m = 2, \\ \lim_{n\to\infty} T_n \overset{L_2}{=} 0 & m > 2 \end{cases}$$

since,

$$\begin{cases} ET_n^2 = O(1) & m = 2, \\ ET_n^2 = o(1) & m > 2. \end{cases}$$

Proof. First we verify that for $i, j \in N^m$

$$a_{i,j} = EG_{i_1} G_{i_2} G_{j_1} G_{j_2}$$

is uniformly bounded. We get the bound by applying Cauchy-Schwarz iteratively and using assumption $\sup_i EG_i^4 < \infty$.

We check that the second non-central moment converges to zero,

$$\begin{aligned} & E\left(T_n\right)^2 \\ &= \frac{1}{n^{2m-2}} \sum_{i,j \in N^m} EG_{i_1} G_{i_2} G_{j_1} G_{j_2} Eh(Z_i) h(Z_j) \\ &\leq \frac{1}{n^{2m-2}} \sum_{i,j \in N^m} |a_{i,j} Eh(Z_i) h(Z_j)| \\ &\leq \left(\sup_n \sup_{i,j \in N^m} |a_{i,j}| \right) \frac{1}{n^{2m-2}} \sum_{i,j \in N^m} |Eh(Z_i) h(Z_j)|. \end{aligned}$$

Supremum over n is needed since $EG_{i_1} G_{i_2} G_{j_1} G_{j_2}$ might change with n. Lemma 4, by the assumption that $(h(\cdots) \times h(\cdots), Z_t)$ is of type Δ, the growth of the inner sum $\sum_{i,j \in N^m} |Eh(Z_i) h(Z_j)|$ is at most of order

$$O(n^m) + o(n^{2m+2-k}).$$

Since $\Delta(h \times h, r) = o(r^{-4})$, the growth rate is

$$E\left(T_n\right)^2 = \frac{O\left(n^m\right) + o(n^{2m-2})}{n^{2m-2}} = \begin{cases} O(1) & m = 2 \\ o(1) & m > 2 \end{cases}$$

For $m = 2$ we have assumed existence of an extra term $o(1)$, which concludes the proof. □

We next prove that the asymptotic distribution of a V-statistic depends on number of terms in the Hoeffding decomposition that are equal to zero.

Lemma 2. *Let h be a core with m arguments. If $h_0 = h_1 = 0$, and for all $c > 2$ component $(h_c \times h_c, Z_t)$ is of type Δ, with $\Delta(h_c \times h_c, r) = o(r^{-4})$ then*

$$\lim_{n\to\infty} \left(nV_n(h) - \binom{m}{2} nV_n(h_2) \right) \overset{L_2}{=} 0$$

Proof. Using Hoeffding decomposition we write the core h as a sum of the components h_c ,

$$\begin{aligned} nV_n(h) =& nV_n(h_m) + \binom{m}{1} nV_n(h_{m-1}) + ... \\ &+ \binom{m}{m-2} nV_n(h_2) + \binom{m}{m-1} nV_n(h_1) + h_0. \end{aligned}$$

$h_0 = 0$ and $h_1 = 0$. By Lemma 1, for $c \geq 3$, $nV_n(h_c)$ converges to zero in mean squared. To see that it suffices to put $Q = 1$ and verify that $(h_c \times h_c, Z_t)$ is of Δ type, which is explicitly assumed. □

Before we study the asymptotic distribution of a bootstrapped statistic B_n we need to state three simple lemmas that will be frequently used.

Lemma 3. *If W_i is a bootstrap process then*

$$\lim_{n\to\infty} \frac{l_n}{n} \sum_{i=1}^{n} W_i \overset{L_2}{=} 0.$$

Proof. By the definition of W_i, $E(\sum_{i=1}^n W_i)^2 \leq n2\sum_{r=1}^n Cov(W_0, W_r) = nO(l_n)$, where $\sum_{r=1}^n Cov(W_0, W_r) = O(l_n)$ follows from bootstrap assumption. Also, by the bootstrap assumptions (Definition 1) we have $\lim_{n\to\infty} \frac{l_n^3}{n^2} = 0$. Therefore $\frac{1}{n}\sum_{i=1}^n W_i$ converges to zero in mean squared. □

Lemma 4. *If $\{W_i\}$ is a bootstrap process then*

$$\sum_{i=1}^{n} \tilde{W}_i = \sum_{i=1}^{n} \left(W_i - \frac{1}{n} \sum_{j=1}^{n} W_j \right) = 0.$$

Lemma 5. *Let f be a function and let $j = \{j_1, \ldots, j_q\}$ be a subset of $\{1, \ldots, m\}$. Then*

$$\sum_{i \in N^m} f(Z_{i_{j_1}}, ..., Z_{i_{j_q}}) = n^{m-q} \sum_{i \in N^q} f(Z_{i_1}, ..., Z_{i_q})$$

Proof. Each element $f(Z_{i_{j_1}}, ..., Z_{i_{j_q}})$ is repeated exactly n^{m-q} times. □

We now prove an analogue of the Lemma 2 for bootstrapped statistics B.

Lemma 6. *Let h be a core of a m arguments and let Q_i denote W_i or $\tilde{W}_i$. If*

$$\frac{1}{n^2} \sum_{i \in N^2} Q_{i_1} Q_{i_2} h_0 = 0,$$
$$\frac{1}{n^m} \sum_{i \in N^m} \sum_{1 \le j \le m} Q_{i_1} Q_{i_2} h_1(Z_{i_j}) = 0.$$

and (h_c, Z_t) for $c > 2$ are of type Δ, with $\Delta(h_c \times h_c, r) = o(r^{-4})$ then

$$\lim_{n \to \infty} \left(nB(h) - \binom{m}{2} nB(h_2) \right) \overset{L_2}{=} 0$$

Proof. Where it is necessary, we check claims for both W_i and $\tilde{W}_i$ separately. We will frequently use the fact that $\frac{l_n}{n} \sum_{i=1}^n Q_i, \frac{1}{n} \sum_{i=1}^n Q_i$ converge to zero in mean square.

Using Hoeffding decomposition we write core h as a sum of components h_c (the ones with h_0, h_1 are equal to zero and therefore omitted)

$$\begin{aligned} nB_1(h) = \frac{1}{n^{m-1}} \sum_{i \in N^m} \Big[& Q_{i_1} Q_{i_2} h_m(Z_{i_1}, ..., Z_{i_m}) + \\ & \sum_{1 \le j_1 < ... < j_{m-1} \le m} Q_{i_1} Q_{i_2} h_{m-1}(Z_{i_{j_1}}, ..., Z_{i_{j_{m-1}}}) + ... + \\ & \sum_{1 \le j_1 < j_2 \le m} Q_{i_1} Q_{i_2} h_2(Z_{i_{j_1}}, Z_{i_{j_2}}) \Big]. \end{aligned}$$

Consider the sum associated with h_c

$$\frac{1}{n^{m-1}} \sum_{i \in N^m} \sum_{1 \le j_1 < ... < j_c \le m} Q_{i_1} Q_{i_2} h_c(Z_{i_{j_1}}, ..., Z_{i_{j_c}}). \tag{3.1}$$

We will show that for almost all fixed $j_1 < \cdots < j_c$ the sum 3.1 converges to zero.

Suppose $j_1 > 2$. The sum 3.1 can be written

$$\frac{1}{n^{m-1}} \sum_{i \in N^m} Q_{i_1} Q_{i_2} h_c(Z_{i_{j_1}}, ..., Z_{i_{j_c}}) \overset{L.5}{=\!=} \frac{1}{n^{c+1}} \sum_{i \in N^{c+2}} Q_{i_1} Q_{i_2} h_c(Z_{i_3}, ..., Z_{i_{c+2}})$$
$$= \left(\frac{1}{n^{c-1}} \sum_{i \in N^c} h_c(Z_{i_1}, ..., Z_{i_c}) \right) \left(\frac{1}{n} \sum_{i=1}^{n} Q_i \right)^2 = \frac{n}{l_n} V_n(h_c) \left(\frac{l_n}{n} \sum_{i=1}^{n} Q_i \right)^2.$$

By Lemma 1, for $c \geq 3$, $\frac{n}{l_n} V_n(h_c)$ converges to zero in mean squared. Indeed, it is sufficient to put $G_i = 1$ and $T_n = n V_n(h_c)$ and notice that $\frac{n}{l_n} V_n(h_c) = \frac{1}{l_n} = o(1) T_n$, since $l_n \to \infty$. Consequently, since $(\frac{1}{n} \sum_{i=1}^{n} Q_i)^2$ converges to zero in mean square 3, the product, converges to zero in mean square i.e.

$$V_n(h_c) \left(\frac{1}{n} \sum_{i=1}^{n} Q_i \right)^2 \overset{L_2}{\to} 0$$

Suppose $j_1 = 2$. The sum 3.1 can be written

$$\frac{1}{n^{m-1}} \sum_{i \in N^m} Q_{i_1} Q_{i_2} h_c(Z_{i_2}, ..., Z_{i_{j_c}}) \overset{L.5}{=\!=} \frac{1}{n^c} \sum_{i \in N^{c+1}} Q_{i_1} Q_{i_2} h_c(Z_{i_2}, \cdots, Z_{i_{j_c}}) =$$
$$\left(\frac{1}{l_n n^{c-1}} \sum_{i \in N^c} Q_{i_1} h_c(Z_{i_1}, \cdots, Z_{i_c}) \right) \left(\frac{l_n}{n} \sum_{i=1}^{n} Q_i \right). \tag{3.2}$$

The latter expression $\frac{l_n}{n} \sum_{i=1}^{n} Q_i$ converges to zero in mean square. The former expression can be further decomposed

$$\frac{1}{l_n} n^{-c+1} \sum_{i \in N^c} Q_{i_1} h_c(Z_{i_1}, \cdots, Z_{i_c}) = \frac{1}{4} \frac{1}{l_n} (T_+ - T_-) \text{ where,}$$
$$\frac{1}{l_n} T_- = \frac{1}{l_n} n^{-c+1} \sum_{i \in N^2} (Q_{i_1} - 1) h_c(Z_{i_1}, \cdots, Z_{i_c})(Q_{i_2} - 1),$$
$$\frac{1}{l_n} T_+ = \frac{1}{l_n} n^{-c+1} \sum_{i \in N^2} (Q_{i_1} + 1) h_c(Z_{i_1}, \cdots, Z_{i_c})(Q_{i_2} + 1),$$

We use Lemma 1 for $\frac{1}{l_n} T_+$ and $\frac{1}{l_n} T_-$, to show that they converge to zero. We need to check that

$$\sup_i E(Q_i + / - 1)^4 < \infty$$

If $Q_i = W_i$ this follows from the bootstrap assumption (see Definition 1)

$\sup_n \sup_{i\leq n} EW_{i,n}^4 < \infty$. If $Q_i = \tilde{W}_i$ we check that

$$E(\frac{1}{n}\sum_{i=1}^{n} W_i)^4 \leq \sup_n \sup_{i\leq n} EW_{i,n}^4,$$

and so $\leq \sup_i E(\tilde{W}_i) < \infty$. Now we conclude that both $\frac{1}{l_n}T_+$ and $\frac{1}{l_n}T_-$ converge to zero. Therefore their sum (even though they are not independent) converges to zero.

Suppose $j_1 = 1$ and $j_2 > 2$. This case is identical to the previous case, up to swapping i_1, i_2 in the equation 3.2.

Finally, suppose $j_1 = 1$ and $j_2 = 2$ and $c > 2$. The sum 3.1 can be written

$$\frac{1}{n^{m-1}}\sum_{i\in N^m} Q_{i_1}Q_{i_2}h_c(Z_{i_1}, Z_{i_2}, ..., Z_{i_{j_c}}) \overset{L.5}{=\!=} \frac{1}{n^c}\sum_{i\in N^{c+1}} Q_{i_1}Q_{i_2}h_c(Z_{i_1}, Z_{i_2}, ..., Z_{i_{j_c}})$$

We again use Lemma 1 to see that this sum converges to zero in mean squared (we checked the assumptions above). We have proved that

$$\lim_{n\to\infty}\left(nB(h) - \binom{m}{2}nB(h_2)\right) \overset{L_2}{=} 0$$

□

So far we avoided expressing results in terms of τ-mixing and degeneracy of a core, now we relate Δ formalism to those concepts. We start with a technical lemma.

Lemma 7. *If h is a Lipschitz continuous core then its components are also Lipschitz continuous.*

Proof. Note that the auxiliary function 2.3 used in the Hoeffding decomposition

$$g_c(z_1, ...z_c) = Eh(z_1, ..., z_c, Z_{c+1}^*, ..., Z_m^*).$$

is Lipschitz, since h is Lipschitz continuous.

$$
\begin{aligned}
&|g_c(z_1,...z_c) - g_c(z'_1,...z'_c)| \\
&\leq \left| \int [h(z_1,...,z_c,z_{c+1},...,z_m) - h(z'_1,...,z'_c,z_{c+1},...,z_m)] dP(z_{c+1}) \cdots dP(z_m) \right| \\
&\leq \left| \int Lip(h) \left(\sum_{i=1}^{c} |z_i - z'_i| + \sum_{i=c+1}^{m} |z_i - z_i| \right) dP(z_{c+1}) \cdots dP(z_m) \right| \\
&\leq \left| \int Lip(h) \left(\sum_{i=1}^{c} |z_i - z'_i| \right) dP(z_{c+1}) \cdots dP(z_m) \right| \\
&= |Lip(h) \sum_{i=1}^{c} |z_i - z'_i| \int dP(z_{c+1}) \cdots dP(z_m)| \\
&= |Lip(h) \sum_{i=1}^{c} |z_i - z'_i||.
\end{aligned}
$$

h_0 is obviously Lipschitz continuous. If h_k for $k < c$ are Lipschitz continuous then, since g_c is Lipschitz continuous, h_c is also Lipschitz continuous as a sum of Lipschitz continuous functions. □

Lemma 8. *Let $\{Z_t\}$ be a τ-dependent stationary process and h be a Lipschitz core of m arguments, If for all $c > 0$ $(h_c \times h_c, Z_t)$ and (h, Z_t) are of type Δ with the rate $O(\tau(d))$ then*

$$\Delta(h,d) = \Delta(h_c \times h_c, d) = O(\tau(d))$$

Proof. Let $f = h_c \times h_c$ or $f = h$. f is canonical and Lipschitz continuous (if $f = h_c \times h_c$ it follows from Lemma 7). Suppose i_r is the isolating index. Further suppose there are k indexes $a_1, \cdots, a_k$ smaller than i_r and $m - k - 1$ indexes greater than i_r, namely $a_{k+2}, \cdots, a_m$. In this notation $a_{k+1} = i_r$.

Let us partition the vector $(Z_{i_1}, \ldots, Z_{i_m})$ into three parts:

$$A = \left(Z_{a_1}, \ldots, Z_{a_k}\right),\ B = Z_{a_{k+1}},\ C = \left(Z_{a_{k+2}}, \ldots, Z_{a_m}\right).$$

where a_{k+1} is the isolating index. If $k = 0$, A is empty and if $k = m - 1$, C is empty but this does not change our arguments below. Using Lemma [26, Lemma 5.3], we will construct B^* and C^{**} that are independent of A and

independent of each other and

$$E\left\|(A,B,C)-(A,B^*,C^{**})\right\|_1 = O(\tau(w)), \tag{3.3}$$

where w is an isolating distance [1]. Let $D=(B,C)$ The [26, Lemma 5.3] guarantees that there exists D^* independent of A, such that

$$\begin{aligned}\|Ed(D,D^*)|\sigma(A)\|_1 &= E|Ed(D,D^*)|\sigma(A)| \\ &= E(Ed(D,D^*)|\sigma(A)) = Ed(D,D^*) = O(\tau(w)),\end{aligned}$$

where d is the L_1 distance on Euclidean space (non-negativity justifies dropping absolute value). By definition of τ-mixing, $\tau(w)\geq\tau(\sigma(A),D)$. Since $D^*=(B^*,C^*)$ has the same distribution as D (in particular it has the same τ dependence structure) we use the lemma again to construct C^{**}, independent of A and B^*, such that

$$Ed(C,C^{**}) = O(\tau(w)).$$

By the triangle inequality we obtain equation 3.3.

$$\begin{aligned}&Ed\big((A,B,C)-(A,D^*)+(A,D^*)-(A,B^*,C^{**})\big)\leq \\ &Ed\big((A,B,C)-(A,D^*)\big)+Ed\big((A,D^*)-(A,B^*,C^{**})\big)= \\ &Ed(D,D^*)+Ed(C,C^{**}) = O(\tau(w)).\end{aligned}$$

Since B^* is a singleton, independent of both A and C^{**}, by degeneracy of f

$$Ef(A,B^*,C^{**}) = 0. \tag{3.4}$$

Note that $f(A,B^*,C^{**})$ is just a shorthand, random variables A,B^*,C^{**} are inserted in the right order. Thus, we have that

$$\begin{aligned}|Ef(Z_{i_1},\ldots,Z_{i_m})| &\leq E\left|f(A,B,C)-f(A,B^*,C^{**})\right|+|Ef(A,B^*,C^{**})| \\ &\leq \mathrm{Lip}(f)E\left\|(A,B,C)-(A,B^*,C^{**})\right\|_1+0 \\ &= O(\tau(w)).\end{aligned}$$

[1] [26, Lemma 5.3] assumes that there exists a random variable δ independent of the vector (A,B,C). This assumption is important only if CDF of the vector is not continuous, we can assume that our space is endowed with such δ.

□

Finally we can prove Theorem 1.

Proof. In the proof we are going to use [67][Theorems 2.1, 3.1], which characterise asymptotic properties of $nV_n(h_2)$ and $nB(h_2)$. Both theorems use similar set of assumptions which we verify upfront.
Assumption A2.

- (i) h_2 is one-degenerate and symmetric - this follows from the Hoeffding decomposition;
- (ii) h_2 is a kernel - is one of the assumptions of this theorem;
- (iii) $Eh_2(Z_1, Z_1) < \infty$ – follows from $\sup_{i \in N^6} |Eh(Z_i)| < \infty$;
- (iv) h_2 is Lipschitz continuous - follows from the Lemma 7.

Assumption B1, A1. Assumption $B1$, $\sum_{r=1}^{n} r^2 \sqrt{\tau(r)} < \infty$, is the same as ours, assumption $A1$, $\sum_{r=1}^{n} \sqrt{\tau(r)} < \infty$ is implied.
Assumption B2. This assumption about the bootstrap process W_t is the same as our Definition 1.

Denote by V the weak limit of $nV_n(h_2)$, which exits by the [67][Theorem 2.1], and let $\mathcal{F} = \sigma(Z_1, \cdots, Z_n)$. By Leucht and Neumann [67, Theorem 3.1], since the distribution of V is continuous, we have

$$\sup_{x \in R} |P(nB_n(h_2) < x|\mathcal{F}) - P(V < x)| \to 0$$

in probability. We show that $nB_n(h_2)$ converges to V weakly, by showing pointwise convergence of CDF

$$\begin{aligned} \lim_{n\to\infty} P(nB_n(h_2) < x) &= \lim_{n\to\infty} EP(nB_n(h_2) < x|\mathcal{F}) \\ &= E \lim_{n\to\infty} P(nB_n(h_2) < x|\mathcal{F}) = EP(V < x) = P(V < x) \end{aligned}$$

To change the order of limit and expectation we have dominated convergence Theorem, justified since $P(nB_n(h) < x|\mathcal{F})$ are bounded by 1. The difference

$n(B_n(h) - V_n(h))$ is

$$n\left(B_n(h) - \binom{m}{2} B_n(h_2)\right) + \binom{m}{2}\left(nB_n(h_2) - V\right) + \left(\binom{m}{2} V - nV_n(h)\right)$$

By Lemma 6 and Lemma 2 respectively, both

$$n(B(h) - \binom{m}{2} B(h_2)), n(V_n(h) - n\binom{m}{2} V_n(h_2))$$

converge to zero in mean square. We check assumptions: since Z_t is tau mixing and h is Lipschitz continuous, by Lemma 8 all self products of components and Z_t, $(h_c \times h_c, Z_t)$ for $c > 0$, are Δ type of order $\tau(r)$, of order at least $o(r^{-4})$ (since $\sum_{r=1}^{n} r^2\sqrt{\tau(r)} < \infty$). Since h is one degenerate, first and zero component h_0, h_1 are equal to zero (and so are $B(h_0), B(h_1)$).

This shows that $nB_n(h_2)$ converges weakly to V. □

Proof of Theorem 2

Proof. Using Hoeffding decomposition we write the core h as a sum of the components h_c ,

$$\begin{aligned} nV_n(h) =& nV_n(h_m) + \binom{m}{1} nV_n(h_{m-1}) + ... \\ &+ \binom{m}{m-2} nV_n(h_2) + \binom{m}{m-1} nV_n(h_1) + h_0. \end{aligned}$$

By the Lemma 1, for $c \geq 1$, $V_n(h_c)$ converges to zero in probability. The sum associated with h_1 is

$$V_n(h_1) = \frac{1}{n}\sum_{i=1}^{N} h_1(Z_i).$$

By Lemma 8 $(h_1 \times h_1, Z_t)$ is Δ type of order $o(r^{-4})$. Using Lemma 1 we get the growth rate of $E(V_n(h_1))^2 = O(\frac{1}{n})$, thus $V_n(h_1)$ converges in mean square to zero. □

Proof of Theorem 3

Proof. We show that the second non central moment of B_1 converges to 0. The second non central moment is

$$\begin{aligned} EB_1 &= E\frac{1}{n^{2m}} \sum_{i \in N^{2m}} W_{i_1} W_{i_2} W_{i_{m+1}} W_{i_{m+2}} Eh(Z_{i_1}, ..., Z_{i_m}) h(Z_{i_{m+1}}, ..., Z_{i_{2m}}) \\ &= \frac{1}{n^{2m}} \sum_{i \in N^{2m}} EW_{i_1} W_{i_2} W_{i_{m+1}} W_{i_{m+2}} Eh(\cdots) h(\cdots) \\ &\leq CE\frac{1}{n^4} \sum_{i \in N^4} |EW_{i_1} W_{i_2} W_{i_{m+1}} W_{i_{m+2}}| \\ &= CE\left(\frac{1}{n}\sum_{i=1}^{n} W_i\right)^4. \end{aligned}$$

The inequality in the third line follows from the fact that correlations of the bootstrap process W_i are positive (Definition 1) and

$$C = \sup_n \sup_{i \in N^m} Eh(Z_{i_1}, ..., Z_{i_m}) h(Z_{i_{m+1}}, ..., Z_{i_2 m}),$$

is finite. By Lemma 3

$$\frac{1}{n}\sum_{i=1}^{n} W_i \to 0,$$

and therefore $EC\left(\frac{1}{n}\sum_{i=1}^{n} W_i\right)^4 \to 0$.

We now prove that $o(n)B_2(h)$ converges to zero. Using Hoeffding decomposition we write core h as a sum of components h_c and h_0

$$nB_2(h) = \frac{1}{n^{m-1}} \sum_{i \in N^m} \Big[h_0 \tilde{W}_{i_1} \tilde{W}_{i_2} + \sum_{1 \leq j \leq m} \tilde{W}_{i_1} \tilde{W}_{i_2} h_1(Z_{i_j}) \tag{3.5}$$

$$\sum_{1 \leq j_1 < j_2 \leq m} \tilde{W}_{i_1} \tilde{W}_{i_2} h_2(Z_{i_{j_1}}, Z_{i_{j_2}}) + ... + \tilde{W}_{i_1} \tilde{W}_{i_2} h_m(Z_{i_1}, ..., Z_{i_m}) \Big]. \tag{3.6}$$

We examine terms of the above sum starting form the one with h_0 - it is equal to zero

$$\frac{1}{n^{m-1}} \sum_{i \in N^m} h_0 \tilde{W}_{i_1} \tilde{W}_{i_2} \overset{L.5}{=\!=} \frac{1}{n} h_0 \sum_{i \in N^2} \tilde{W}_{i_1} \tilde{W}_{i_2} = \frac{1}{n} h_0 \left(\sum_{i=1} \tilde{W}_i\right)^2 \overset{L.4}{=\!=} 0.$$

Term with h_1 is zero as well, to see that fix j and consider

$$T_j = \frac{1}{n^{m-1}} \sum_{i \in N^m} \tilde{W}_{i_1} \tilde{W}_{i_2} h_1(Z_{i_j}).$$

If $j = 1$ then

$$T_1 \overset{L.5}{=\!=} \frac{1}{n} \sum_{i \in N^2} \tilde{W}_{i_1} \tilde{W}_{i_2} h_1(Z_{i_1}) = \frac{1}{n} \left(\sum_{i=1}^{n} \tilde{W}_i h_1(Z_i) \right) \left(\sum_{i=1} \tilde{W}_i \right) \overset{L.4}{=\!=} 0.$$

If $j = 2$ the same reasoning holds and if $j > 2$

$$T_j \overset{L.5}{=\!=} \frac{1}{n^2} \sum_{i \in N^3} \tilde{W}_{i_1} \tilde{W}_{i_2} h_1(Z_{i_3}) = \frac{1}{n} \left(\sum_{i=1}^{n} h_1(Z_i) \right) \left(\sum_{i=1} \tilde{W}_i \right)^2 \overset{L.4}{=\!=} 0.$$

By Lemma 6, since $B(h_0) = B(h_1) = 0$, $(nB(h) - \binom{m}{2} nB(h_2)) \to 0$ in mean square and the only term that remains is

$$T_n = \frac{1}{n} \sum_{i,j \in N} \tilde{W}_i \tilde{W}_j h_2(Z_i, Z_j)$$

Now we can use the Lemma 1 to show that $o(1)T_n$ converges to zero. □

Applications to Kernel Tests

In this section, we describe how the wild bootstrap for V-statistics can be used to construct kernel tests for independence and the two-sample problem, in presence of weakly dependent observations. The main concepts underpinning the kernel testing framework are reviewed in the section 2.2.

Wild Bootstrap For MMD

Denote the observations by $\{X_i\}_{i=1}^n \sim P_x$, and $\{Y_j\}_{j=1}^n \sim P_y$. Our goal is to test the null hypothesis $\mathbf{H}_0 : P_x = P_y$ vs. the alternative $\mathbf{H}_1 : P_x \neq P_y$. The empirical MMD can be written as a V-statistic with the core of degree two on pairs $z_i = (x_i, y_i)$ given by

$$h(z_1, z_2) = k(x_1, x_2) - k(x_1, y_2) - k(x_2, y_1) + k(y_1, y_2).$$

It is clear that whenever k is Lipschitz continuous and $\sup_{i,j} Ek(X_i, X_j), \sup_{i,j} Ek(Y_i, Y_j)$ are finite, so is h. Moreover, h is a valid positive definite kernel, since it can be

represented as an RKHS inner product

$$h(z_1, z_2) = \langle k(\cdot, x_1) - k(\cdot, y_1), k(\cdot, x_2) - k(\cdot, y_2) \rangle_{\mathcal{H}_k}.$$

Under the null hypothesis, h is also one-degenerate, i.e., $Eh((x_1, y_1), (X_2, Y_2)) = 0$. Therefore, by the Theorems 1, 3, we can use the bootstrapped statistics in (2.5) and (2.6) to approximate the null distribution and attain a desired test level.

Wild Bootstrap For HSIC

Recall (2.12) that the core of the test static for HSIC, with notation $z_i = (x_i, y_i)$, is

$$\begin{aligned} h(z_1, z_2, z_3, z_4) = \frac{1}{4!} \sum_{\pi \in S_4} k(x_{\pi(1)}, x_{\pi(2)}) [l(y_{\pi(1)}, y_{\pi(2)}) \\ + l(y_{\pi(3)}, y_{\pi(4)}) - 2l(y_{\pi(2)}, y_{\pi(3)})]. \end{aligned}$$

One-degeneracy of the core under the null hypothesis was stated in [44, Theorem 2], [44, Section A.2, following eq. (11)] shows that h_2 is a kernel; $h_0 \geq 0$ follows from the fact that HSIC is a distance. Using Theorems 1,3,2 we can construct an independence test using h. Drawback of this test, when implemented in the most straightforward way, is its quadruple computational complexity. To achieve quadratic time complexity, that matches time complexity of HSIC test for i.i.d. data, we modify our bootstrapped statistic.

Quadratic time HSIC. In this section we assume that kernels k, l are positive and bounded. We define empirical mean embedding $\tilde{\mu}_X(x) = \frac{1}{n} \sum_i^n k(x, X_i)$ and centred kernels

$$\begin{aligned} \bar{k}(x, x') =& k(x', x) - Ek(x, X) - Ek(X', x') + Ek(X, X') \\ =& \langle k(x, \cdot) - \mu_X, k(x', \cdot) - \mu_X \rangle. \\ \tilde{k}(x, x') =& k(x, x') - \frac{1}{n} \sum_i^n k(x, X_i) - \frac{1}{n} \sum_i^n k(x', X_i) + \frac{1}{n^2} \sum_{i,j}^n k(X_j, X_i) \\ =& \langle k(x, \cdot) - \tilde{\mu}_X, k(x', \cdot) - \tilde{\mu}_X \rangle. \end{aligned}$$

where X, X' are i.i.d. copies of X_1. Same definitions hold for the kernel l. Let Q_i denote W_i or $\tilde{W}_i$ (where it is necessary, we check claims for both W_i and

$\tilde{W}_i$ separately). We further define

$$S_n = \frac{1}{\sqrt{n}} \sum_{i \in N} Q_i(\phi(X_i) - \tilde{\mu}_X) \otimes (\phi(Y_i) - \tilde{\mu}_Y), \tag{3.7}$$

$$T_n = \frac{1}{\sqrt{n}} \sum_{i \in N} Q_i(\phi(X_i) - \mu_X) \otimes (\phi(Y_i) - \mu_Y). \tag{3.8}$$

First, we relate T_n to $B(h_2)$.

Statement 1. *[44, section A.2, following eq. (11)] The second component of h is $h_2(z_1, z_2) = \frac{1}{6}\bar{k}(x_1, x_2)\bar{l}(y_1, y_2)$.*

Lemma 9. *Squared norm of T_n is equal to $6B(h_2)$.*

Proof.

$$\begin{aligned}
\|T_n\|^2 =& \frac{1}{n} \sum_{i,j \in N} Q_i Q_j \Big\langle (\phi(X_i) - \mu_X) \otimes (\phi(Y_i) - \mu_Y), (\phi(X_j) - \mu_X) \otimes (\phi(Y_j) - \mu_Y) \Big\rangle \\
=& \frac{1}{n} \sum_{i,j \in N} Q_i Q_j \bar{k}(X_i, X_j) \bar{l}(Y_i, Y_j) \\
=& 6B(h_2).
\end{aligned}$$

□

Next we relate S_n to T_n – we show that the difference between them is asymptotically negligible. We start with a technical lemma.

Lemma 10. *If $(\bar{k} \times \bar{k}, Z_i)$ is of type Δ of order $O(r^{-4})$ (see Definition 3), then*

$$\lim_{n \to \infty} E \left\| \sqrt{n}(\tilde{\mu}_X - \mu_X) \right\|^4 = O(1).$$

Proof.

$$\begin{aligned}
E \left\| \sqrt{n}(\tilde{\mu}_X - \mu_X) \right\|^4 &= E \left\| \frac{1}{\sqrt{n}} \sum_{i \in N} \phi(X_i) - \mu_X \right\|^4 \\
&= E \left(\frac{1}{n} \sum_{i \in N} \langle \phi(X_j) - \mu_X, \phi(X_i) - \mu_X \rangle \right)^2 \\
&= \frac{1}{n^2} E \sum_{i \in N^4} \bar{k} \times \bar{k}(Z_i).
\end{aligned}$$

Since $(\bar{k} \times \bar{k}, X_i)$ is of type Δ, by Lemma 1, the expected value is of order $O(1)$. □

Lemma 11. *If $(\bar{k} \times \bar{k}, Z_i)$, $(\bar{l} \times \bar{l}, Z_i)$ are of type Δ of order $O(r^{-4})$, then, under the null, $\|S_n\|^2 - \|T_n\|^2$ converges to zero in mean square. Under the alternative $\frac{1}{n}(\|S_n\|^2 - \|T_n\|^2)$ converges to zero in mean square.*

Proof. We first show that $E\|S_n - T_n\|^2 \to 0$ both under the null and the alternative. Then, using the fact that $\|T_n\|^2 < \infty$ under the null and $\frac{1}{n}\|T_n\|^2 < \infty$ under alternative we will conclude the proof. The difference $S_n - T_n$ is

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i \in N} Q_i \Big[(\phi(X_i) - \tilde{\mu}_X) \otimes (\phi(Y_i) - \tilde{\mu}_Y) - (\phi(X_i) - \mu_X) \otimes (\phi(Y_i) - \mu_Y) \Big] \\
&\quad = \frac{1}{\sqrt{n}} \sum_{i \in N} Q_i \Big[\phi(X_i) \otimes \mu_Y - \phi(X_i) \otimes \tilde{\mu}_Y \Big] \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i \in N} Q_i \Big[\phi(Y_i) \otimes \mu_X - \phi(Y_i) \otimes \tilde{\mu}_X \Big] \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i \in N} Q_i (\tilde{\mu}_X \otimes \tilde{\mu}_Y - \mu_Y \otimes \mu_X).
\end{aligned}$$

We examine differences separately – it is sufficient to show that each difference converges to zero in mean square.

The expected norm of the first difference is

$$\begin{aligned}
&E \left\| \frac{1}{\sqrt{n}} \sum_{i \in N} Q_i \Big[\phi(X_i) \otimes \mu_Y - \phi(X_i) \otimes \tilde{\mu}_Y \Big] \right\|^2 \\
&\quad = E \left\| \sqrt{n}(\mu_Y - \tilde{\mu}_Y) \otimes \frac{1}{\sqrt{n}} \sum_{i \in N} Q_i \phi(X_i) \right\|^2 \\
&\quad \leq \sqrt{ E \left\| \sqrt{n}(\tilde{\mu}_Y - \mu_Y) \right\|^4 E \left\| \frac{1}{\sqrt{n}} \sum_{i \in N} Q_i \phi(X_i) \right\|^4 }.
\end{aligned}$$

We used $\|v \otimes u\| = \|v\|\|u\|$ and Cauchy-Schwarz inequality. By Lemma 10 the first term is $O(1)$. The second term is equal to

$$E \| \frac{1}{n} \sum_{i \in N} Q_i \phi(X_i) \|^4 = E \left(\frac{1}{n^2} \sum_{i,j} k(X_i, X_j) Q_i Q_j \right)^2 .$$

The expected value converges to zero in mean square by Lemma 1 (the assump-

tion $\sup_{i,j} k(X_i, X_j) < \infty$ is satisfied). Using similar reasoning, the second term

$$E\left\| \frac{1}{\sqrt{n}} \sum_{i \in N} Q_i \left[\phi(Y_i) \otimes \tilde{\mu}_X - \phi(Y_i) \otimes \mu_X \right] \right\|^2$$

also converges to zero. The final term is

$$\begin{aligned} & E\left\| \frac{1}{\sqrt{n}} \sum_{i \in N} Q_i (\tilde{\mu}_X \otimes \tilde{\mu}_Y - \mu_Y \otimes \mu_X) \right\|^2 \\ & = E\left| \frac{1}{n} \sum_{i \in N} Q_i \right| E\left\| \sqrt{n} (\tilde{\mu}_X \otimes \tilde{\mu}_Y - \mu_Y \otimes \mu_X) \right\|^2 \end{aligned}$$

$\frac{1}{n} \sum_{i \in N} Q_i$ converges in mean square to zero (Lemmas 3, 4). We rewrite the second term

$$E\left\| \sqrt{n} (\tilde{\mu}_X \otimes \tilde{\mu}_Y - \tilde{\mu}_Y \otimes \mu_X + \tilde{\mu}_Y \otimes \mu_X - \mu_Y \otimes \mu_X) \right\|^2$$

It is sufficient to bound

$$\begin{aligned} E\left\| \sqrt{n} \tilde{\mu}_Y \otimes (\tilde{\mu}_X - \mu_X) \right\|^2 & \leq E \sqrt{\left\| \tilde{\mu}_Y \right\|^4 E\left\| \sqrt{n} (\tilde{\mu}_X - \mu_X) \right\|^4} \\ E\left\| \sqrt{n} \mu_X \otimes (\tilde{\mu}_Y - \mu_Y) \right\|^2 & = \left\| \mu_X \right\|^2 E\left\| \sqrt{n} (\tilde{\mu}_Y - \mu_Y) \right\|^2 \end{aligned}$$

$E\|\tilde{\mu}_Y\|^4 = E \frac{1}{n^4} \sum_{i \in N^4} (l \times l)(Y_i) = O(1)$, since l is bounded. By Lemma 10 $E\left\| \sqrt{n} (\tilde{\mu}_X - \mu_X) \right\|^4$ and $E\left\| \sqrt{n} (\tilde{\mu}_Y - \mu_Y) \right\|^2$ are finite. Thus, the whole expression converges to zero. We proved that $T_n - S_n$ converges in mean square to zero. We have

$$\begin{aligned} E\left| \|T_n\|^2 - \|S_n\|^2 \right| & \leq E\left| \|T_n\| - \|S_n\| \right| \left| \|T_n\| + \|S_n\| \right| \\ & \leq \sqrt{E\left| \|T_n\| - \|S_n\| \right|^2 E\left| \|T_n\| + \|S_n\| \right|^2} \end{aligned}$$

To show that the above expression converges to zero it is sufficient to show that $E\|T_n\|^2 < \infty$ and $E\|S_n\|^2 < \infty$. Under the null hypothesis, by Lemma 1, expected value of $E\|T_n\|^2 = nB_n(h_2)$ is finite. Since $E\|T_n - S_n\|^2 \to 0$ we

also have $E\|T_n - S_n\| \to 0$. Therefore we have (by simply squaring triangle inequality)

$$\begin{aligned} E\|S_n\|^2 &\leq E\|S_n - T_n + T_n\|^2 \\ &\leq E\|S_n - T_n\|^2 + E\|T_n - S_n\|E\|T_n\| + E\|T_n\|^2 < \infty \end{aligned}$$

Under the alternative we have

$$\begin{aligned} n^{-1}E|\|T_n\|^2 - \|S_n\|^2| &\leq n^{-1}E\big|\|T_n\| - \|S_n\|\big|\|T_n\| + \|S_n\|\big| \\ &\leq \sqrt{E\big|\|T_n\| - \|S_n\|\big|^2 n^{-1}E\big|\|T_n\| + \|S_n\|\big|^2} \end{aligned}$$

it is sufficient to show that $n^{-1}E\|T_n\|^2 < \infty$ and $n^{-1}E\|S_n\|^2 < \infty$. By Theorem 3, $n^{-1}E\|T_n\|^2 < \infty$ is finite and, using the reasoning similar to the one above, we have that $n^{-1}E\|S_n\|^2 < \infty$. □

This shows that we can use squared norm of S_n as a bootstrapped test statistic. For HSIC we redefine B_n

$$B_n^* := \|S_n\|^2 = \frac{1}{n}\sum_{i,j \in N} Q_i, Q_j \tilde{k}(X_i, X_j)\tilde{l}(X_i, X_j). \tag{3.9}$$

B_1^* corresponds to $Q_i = W_i$, B_2^* corresponds to $Q_i = \tilde{W}_i$. This bootstrapped statistic interestingly coincides with $V_n(h)$. Gretton et al. [44] showed that

$$V_n(h) = \frac{1}{n}\sum_{i,j \in N} \tilde{k}(X_i, X_j)\tilde{l}(X_i, X_j). \tag{3.10}$$

Finally, notice that both statistics 3.9 and 3.10 can be calculated in quadratic time.

Proposition 1. *Let $Z_t = (X_t, Y_t)$ be a stationary process that is τ-dependent such that $\sum_{r=1}^{\infty} r^2\sqrt{\tau(r)} < \infty$. Under the null hypothesis B_n^* (3.9) and $nV_n(h)$ (3.10)converge weakly to the same distribution. Under the alternative hypothesis B_n^* converges to zero in probability, while $V_n(h)$ converges to a positive constant.*

Proof. We calculate

$$nV_n(h) - B_n^* = nV_n(h) - 6nB_n(h_2) + 6nB_n(h_2) - B_n^*.$$

By Lemma 9, $6nB_n(h_2) = \|T_n\|^2$. By definition (3.9), $B_n^* = \|S_n\|^2$. By Lemma 11, $6nB_n(h_2) - B_n^*$ converges to zero in mean square. We check assumptions; since process Z_t is τ-mixing (of order $o(r^{-4})$) and both $\bar{k}$, $\bar{l}$ are canonical, Lemma 8 guarantees that $(\bar{k}, Z_i)$, $(\bar{l}, Z_i)$ are of type Δ of order $O(r^{-4})$.

Under the null hypothesis, by Theorem 1, $nV_n(h) - 6nB_n(h_2)$ converges to zero. We check assumptions; by Lemma 1, h_2 is a symmetric, one-degenerate, bounded kernel, assumptions concerning τ-mixing are satisfied.

Under the alternative, by Theorem 3 and Lemma 11 respectively, $6B_n(h_2)$ and $\frac{1}{n}B_n^* - 6B_n(h_2)$ converge to zero in mean square. By Theorem 3, $V_n(h)$ converges to a positive constant. □

We consider two types of tests: instantaneous independence and independence at multiple time lags.

Test of instantaneous independence Here, the null hypothesis $\mathbf{H_0}$ is that X_t and Y_t are independent at all times t, and the alternative hypothesis $\mathbf{H_1}$ is that they are dependent. We use Proposition1 directly to bootstrap an appropriate quantile and compare it with a test statistic.

Lag-HSIC Proposition 1 allows us to construct a test of time series independence that is similar to one designed by [14]. Here, we will be testing against a broader null hypothesis: X_t and $Y_{t'}$ are independent for $|t - t'| < M$ for an arbitrary large but fixed M.

Since the time series $Z_t = (X_t, Y_t)$ is stationary, it suffices to check whether there exists a dependency between X_t and Y_{t+m} for $-M \leq m \leq M$. Since each lag corresponds to an individual hypothesis, we will require a Bonferroni correction to attain a desired test level α. We therefore define $q = 1 - \frac{\alpha}{2M+1}$. The shifted time series will be denoted $Z_t^m = (X_t, Y_{t+m})$. Let $S_{m,n} = nV_n(h, Z^m)$ denote the value of the normalized HSIC statistic calculated on the shifted process Z_t^m. Let $F_{b,n}$ denote the empirical cumulative distribution function obtained by the bootstrap procedure using B_n^* (3.9). The test will then reject the null hypothesis if the event $\mathcal{A}_n = \left\{\max_{-M \leq m \leq M} S_{m,n} > F_{b,n}^{-1}(q)\right\}$ occurs. By a simple application of the union bound, it is clear that the asymptotic probability of the Type I error will be $\lim_{n\to\infty} P_{\mathbf{H_0}}(\mathcal{A}_n) \leq \alpha$. On the other hand, if the alternative holds, there exists some m with $|m| \leq M$ for which

$V_n(h, Z^m) = n^{-1} S_{m,n}$ converges to a non-zero constant. In this case

$$P_{\mathbf{H_1}}(\mathcal{A}_n) \geq P_{\mathbf{H_1}}(S_{m,n} > F^{-1}_{b,n}(q)) = P_{\mathbf{H_1}}(n^{-1} S_{m,n} > n^{-1} F^{-1}_{b,n}(q)) \to 1 \tag{3.11}$$

as long as $n^{-1} F^{-1}_{b,n}(q) \to 0$, which follows from the convergence of B^*_n (3.9) to zero in probability shown in Proposition 1. Therefore, the Type II error of the multiple lag test is guaranteed to converge to zero as the sample size increases. Our experiments in the next Section demonstrate that while this procedure is defined over a finite range of lags, it results in tests more powerful than the procedure for an infinite number of lags proposed in [14]. We note that a procedure that works for an infinite number of lags, although possible to construct, does not add much practical value under the present assumptions. Indeed, since the τ-mixing assumption applies to the joint sequence $Z_t = (X_t, Y_t)$, dependence between X_t and Y_{t+m} is bound to disappear at a rate of $o(m^{-6})$, i.e., the variables both within and across the two series are assumed to become gradually independent at large lags.

Experiments

The MCMC M.D. We employ MMD in order to diagnose how far an MCMC chain is from its stationary distribution [89, Section 5], by comparing the MCMC sample to a benchmark sample. Note that in next chapter we develop more realistic test that does not require benchmark sample. A hypothesis test of whether the sampler has converged based on the standard permutation-based bootstrap leads to too many rejections of the null hypothesis, due to dependence within the chain. Thus, one would require heavily thinned chains, which is wasteful of samples and computationally burdensome. Our experiments indicate that the wild bootstrap approach allows consistent tests directly on the chains, as it attains a desired number of false positives.

To assess performance of the wild bootstrap in determining MCMC convergence, we consider the situation where samples $\{X_i\}$ and $\{Y_i\}$ are bivariate, and both have the identical marginal distribution given by an elongated normal $P = \mathcal{N}\left(\begin{bmatrix} 0 & 0 \end{bmatrix}, \begin{bmatrix} 15.5 & 14.5 \\ 14.5 & 15.5 \end{bmatrix}\right)$. However, they could have arisen either as independent samples, or as outputs of the Gibbs sampler with stationary distribution P. Table 3.1 shows the *rejection rates* under the significance level

Table 3.1: Rejection rates for two-sample experiments. **M**CMC: sample size=500; a Gaussian kernel with bandwidth $\sigma = 1.7$ is used; every second Gibbs sample is kept (i.e., after a pass through both dimensions). Wild bootstrap uses blocksize of $l_n = 20$; averaged over at least 200 trials. The Type II error for all tests was zero

	experiment \ method	permutation	$\widehat{\text{MMD}}_{k,b}$	B_1	B_2
MCMC	i.i.d. vs i.i.d. ($\mathbf{H}_0$)	.040	.025	.012	.070
	i.i.d. vs Gibbs ($\mathbf{H}_0$)	.528	.100	.052	.105
	Gibbs vs Gibbs ($\mathbf{H}_0$)	.680	.110	.060	.100

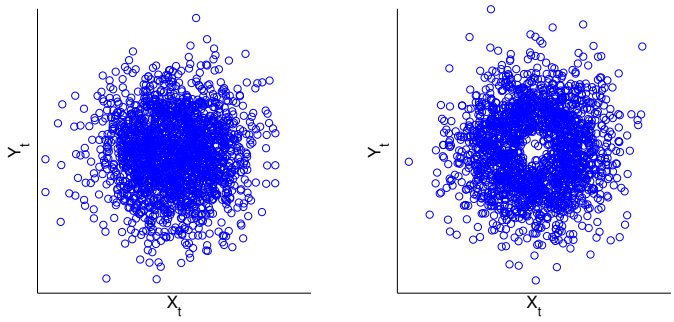

Figure 3.1: X_t and Y_t, described in the instantaneous independence experiment, with extinction rates 50% (left) and 99.8% (right), respectively.

$\alpha = 0.05$. It is clear that in the case where at least one of the samples is a Gibbs chain, the permutation-based test has a Type I error much larger than α. The wild bootstrap using B_1 (without artificial degeneration) yields the correct Type I error control in these cases. Consistent with findings in [67, Section 5], B_1 mimics the null distribution better than B_2. In the alternative scenario where $\{Y_i\}$ was taken from a distribution with the same covariance structure but with the mean set to $\mu = \begin{bmatrix} 2.5 & 0 \end{bmatrix}$, the Type II error for all tests was zero.

Instantaneous independence To examine instantaneous independence test performance, we compare it with the Shift-HSIC procedure. The shift procedure is a type of a block bootstrap, that we have proposed in the preliminary work, in which two times series are shifted with respect to each other. Shifting preserves most of the temporal dependence and removes some of the instantaneous defence, since X_t, Y_{t+k}, for large k, are likely to be almost independent. The quantiles approximated by calculating the the test statistic on series $\{X_t, Y_{(t+k)mod n}\}_{t=1}^n$ for k in some range $K_1 < K_2 < n$ where K_1 obviously must be quite large. We compare two procedures on the 'Extinct Gaussian'

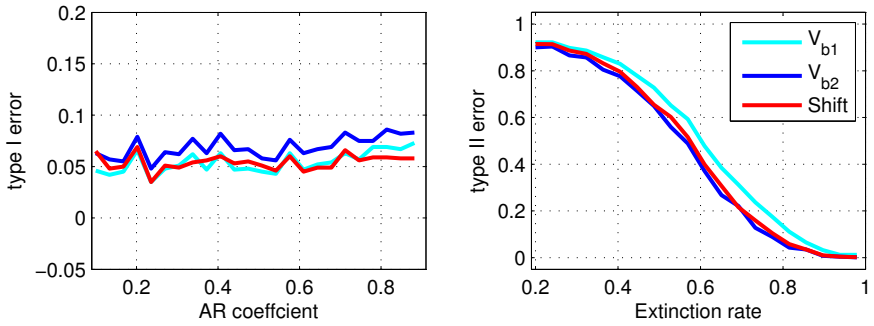

Figure 3.2: Comparison of Shift-HSIC (A) and tests based on B_1^* and B_2^*. The left panel shows the performance under the null hypothesis, where a larger AR coefficient implies a stronger temporal dependence. The right panel show the performance under the alternative hypothesis, where a larger extinction rate implies a greater dependence between processes. $n = 500$

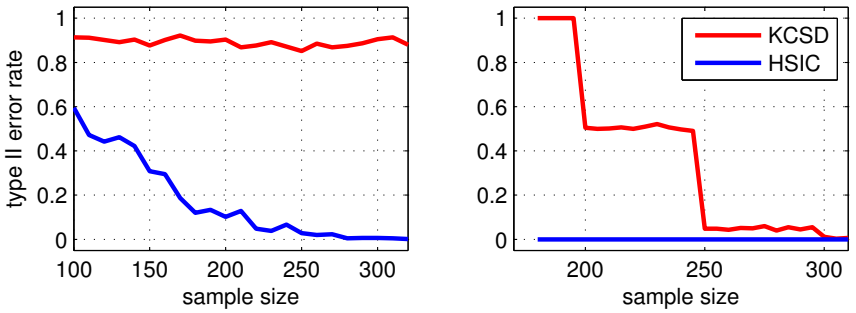

Figure 3.3: In both panel Type II error is plotted. The left panel presents the error of the lag-HSIC and KCSD algorithms for a process following dynamics given by the equation (3.13). The errors for a process with dynamics given by equations (3.14) and (3.15) are shown in the right panel. The X axis is indexed by the time series length, i.e., sample size. The Type I error was around 5%.

Algorithm 1 Generate innovations

Input: extinction rate $0 \leq p \leq 1$, radius r.
repeat
  Initialize η_t, ϵ_t to $N(0,1)$ and d to a number uniformly distributed on $[0,1]$
  .
  if $\eta_t^2 + \epsilon_t^2 > r^2$ **or** $d > p$ **then**
    return η_t, ϵ_t
  end if
until true

autoregressive process, specified by equations

$$X_t = aX_{t-1} + \epsilon_t \quad Y_t = aY_{t-1} + \eta_t, \tag{3.12}$$

with an autoregressive component a. The coupling of the processes is a result of the dependence in the innovations ϵ_t, η_t. These ϵ_t, η_t are drawn from an Extinct Gaussian distribution, defined in Algorithm 2 (top of this page). The parameter p controls how often a point drawn form a ball $B(0,r)$ dies off. According to Algorithm 2, the probability of seeing a point inside the ball $B(0,r)$ is different than for a two dimensional Gaussian $N(\mathbf{0}, Id)$. On the other hand, as p goes to zero, the Extinct Gaussian converges in distribution to $N(\mathbf{0}, Id)$. Figure 3.1 illustrates the joint distribution of X_t, Y_t. The left scatter plot in Figure 3.1 presents X_t and Y_t generated with an extinction rate of 50%, while the right hand plot is generated with an extinction rate of 99.87%. Processes used in this experiment had an autoregressive component of 0.2, and the radius of the innovation process was 1.

We compute type I error as a function of the temporal dependence and type II error as a function of extinction rate. Figure 3.2 shows that all three tests (Shift-HSIC and tests based on B_1 and B_2 (1)) perform similarly.

Lag-HSIC The KCSD Besserve et al. [14] is, to our knowledge, the only test procedure to reject the null hypothesis if there exist t,t' such that Z_t and $Z_{t'}$ are dependent. In the experiments, we compare lag-HSIC with KCSD on two kinds of processes: one inspired by econometrics and one from Besserve et al. [14].
In lag-HSIC, the number of lags under examination was equal to $\max\{10, \log n\}$, where n is the sample size. We used Gaussian kernels with widths estimated by the median heuristic. The cumulative distribution of the V-statistics was

approximated by samples from nB_2^*. To model the tail of this distribution, we have fitted the generalized Pareto distribution to the bootstrapped samples ([76] shows that for a large class of underlying distribution functions such an approximation is valid).

The first process is a pair of two time series which share a common variance,

$$\begin{aligned} X_t &= \epsilon_{1,t}\sigma_t^2, \\ Y_t &= \epsilon_{2,t}\sigma_t^2, \sigma_t^2 = 1 + 0.45(X_{t-1}^2 + Y_{t-1}^2), \\ \epsilon_{i,t} &\overset{i.i.d.}{\sim} \mathcal{N}(0,1), \quad i \in \{1,2\}. \end{aligned} \tag{3.13}$$

The above set of equations is an instance of the VEC dynamics [10] used in econometrics to model market volatility. The left panel of the Figure 3.3 presents the Type II error rate: for KCSD it remains at 90% while for lag-HSIC it gradually drops to zero. The Type I error, which we calculated by sampling two independent copies $(X_t^{(1)}, Y_t^{(1)})$ and $(X_t^{(2)}, Y_t^{(2)})$ of the process and performing the tests on the pair $(X_t^{(1)}, Y_t^{(2)})$, was around 5% for both of the tests. Our next experiment is a process sampled according to the dynamics proposed by [14],

$$X_t = \cos(\phi_{t,1}), \qquad \phi_{t,1} = \phi_{t-1,1} + 0.1\epsilon_{1,t} + 2\pi f_1 T_s \tag{3.14}$$

$$Y_t = [2 + C\sin(\phi_{t,1})]\cos(\phi_{t,2}), \quad \phi_{t,2} = \phi_{t-1,2} + 0.1\epsilon_{2,t} + 2\pi f_2 T_s \tag{3.15}$$

with parameters $C = .4$, $f_1 = 4Hz$,$f_2 = 20Hz$, and frequency $\frac{1}{T_s} = 100Hz$ and $\epsilon_{1,t}, \epsilon_{2,t} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. We compared performance of the KCSD algorithm, with parameters set to vales recommended in [14], and the lag-HSIC algorithm. The Type II error of lag-HSIC, presented in the right panel of the Figure 3.3, is substantially lower than that of KCSD. The Type I error ($C = 0$) is equal or lower than 5% for both procedures. Most oddly, KCSD error seems to converge to zero in steps. This may be due to the method relying on a spectral decomposition of the signals across a fixed set of bands. As the number of samples increases, the quality of the spectrogram will improve, and dependence will become apparent in bands where it was undetectable at shorter signal lengths.

Chapter 4

A Kernel Test of Goodness of Fit.

This chapter is based on Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *ICML*, 2016.

In this chapter we propose a nonparametric statistical test for goodness-of-fit: given a set of samples, the test determines how likely it is that these were generated from a target density function. The measure of goodness-of-fit is a divergence constructed via Stein's method using functions from a Reproducing Kernel Hilbert Space. Our test statistic is based on an empirical estimate of this divergence, taking the form of a V-statistic in terms of the log gradients of the target density and the kernel. We derive a statistical test, both for i.i.d. and non-i.i.d. samples, where we estimate the null distribution quantiles using a wild bootstrap procedure. We apply our test to quantifying convergence of approximate Markov Chain Monte Carlo methods, statistical model criticism, and evaluating quality of fit vs model complexity in nonparametric density estimation.

Test Definition: Statistic and Threshold

We begin with a high-level construction of our divergence discrepancy and the statistical test. While this section aims to outline the main ideas, we provide details and proofs in Section 4.2.

Stein Operator in RKHS

Our goal is to write the maximum discrepancy between target distribution p and observed sample distribution q in a RKHS. Denote by $\mathcal{F}$ the RKHS of real-valued functions on $\mathbb{R}^d$ with reproducing kernel k, and by $\mathcal{F}^d$ the product

RKHS consisting of elements $f := (f_1, \ldots, f_d)$ with $f_i \in \mathcal{F}$, and with a standard inner product $\langle f, g \rangle_{\mathcal{F}^d} = \sum_{i=1}^{d} \langle f_i, g_i \rangle_{\mathcal{F}}$. We further assume that all measures considered in this paper are supported on an open set, equal to zero on the border, and strictly positive[1] (so logarithms are well defined). Similarly to Stein [101], Gorham and Mackey [40], Oates et al. [74], we begin by defining a Stein operator T_p acting on $f \in \mathcal{F}^d$

$$(T_p f)(x) := \sum_{i=1}^{d} \left(\frac{\partial \log p(x)}{\partial x_i} f_i(x) + \frac{\partial f_i(x)}{\partial x_i} \right).$$

Suppose a random variable Z is distributed according to a measure[2] q and X is distributed according to the target measure p. As we will see, the operator can be expressed by defining a function that depends on gradients of the log-density and the kernel,

$$\xi_p(x, \cdot) := \left[\nabla \log p(x) k(x, \cdot) + \nabla k(x, \cdot) \right], \tag{4.1}$$

whose expected inner product with f gives exactly the expected value of the Stein operator

$$E_q T_p f(Z) = \langle f, E_q \xi_p(Z) \rangle_{\mathcal{F}^d} = \sum_{i=1}^{d} \langle f_i, E_q \xi_{p,i}(Z) \rangle_{\mathcal{F}},$$

where $\xi_{p,i}(x, \cdot)$ is the i-th component of $\xi_p(x, \cdot)$. For X from the target measure, we have $E_p(T_p f)(X) = 0$, which can be seen using integration by parts, c.f. Lemma 12 in the supplement. We can now define a Stein discrepancy and express it in the RKHS,

$$\begin{aligned} S_p(Z) &:= \sup_{\|f\|<1} E_q(T_p f)(Z) - E_p(T_p f)(X) \\ &= \sup_{\|f\|<1} E_q(T_p f)(Z) \\ &= \sup_{\|f\|<1} \langle f, E_q \xi_p(Z) \rangle_{\mathcal{F}^d} \\ &= \| E_q \xi_p(Z) \|_{\mathcal{F}^d}, \end{aligned}$$

[1] An example of such a space is the real line

[2] Throughout the article, all occurrences of Z, e.g. $Z', Z_i, Z_\heartsuit$, are understood to be distributed according to q.

This makes it clear why $E_p(T_p f)(X) = 0$ is a desirable property: we can compute $S_p(Z)$ by computing $\|E_q \xi_p(Z)\|$, without the need to access X in the form of samples from p. To state our first result we define

$$\begin{aligned} h_p(x,y) := &\nabla \log p(x)^\top \nabla \log p(y) k(x,y) \\ &+ \nabla \log p(y)^\top \nabla_x k(x,y) \\ &+ \nabla \log p(x)^\top \nabla_y k(x,y) \\ &+ \langle \nabla_x k(x,\cdot), \nabla_y k(\cdot,y) \rangle_{\mathcal{F}^d}, \end{aligned}$$

where the last term can be written as a sum $\sum_{i=1}^d \frac{\partial k(x,y)}{\partial x_i \partial y_i}$. The following theorem gives a simple closed form expression for $\|E_q \xi_p(Z)\|_{\mathcal{F}^d}$ in terms of h_p.

Theorem 5. *If $E h_p(Z,Z) < \infty$, then $S_p^2(Z) = \|E_q \xi_p(Z)\|_{\mathcal{F}^d}^2 = E_q h_p(Z,Z')$.*

The second main result states that the discrepancy $S_p(Z)$ can be used to distinguish two distributions.

Theorem 6. *Let q, p be probability measures and $Z \sim q$. If the kernel k is cc-universal [20, Definition 4.1], $E_q h_q(Z,Z) < \infty$ and $E_q \left\| \nabla \left(\log \frac{p(Z)}{q(Z)} \right) \right\|^2 < \infty$ then $S_p(Z) = 0$ if and only if $p = q$.*

Section 4.2 contains all necessary proofs. We now proceed to construct an estimator for $S(Z)^2$, and outline its asymptotic properties.

Wild Bootstrap Testing

It is straightforward to estimate the squared Stein discrepancy $S(Z)^2$ from samples $\{Z_i\}_{i=1}^n$: a quadratic time estimator is a V-Statistic, and takes the form

$$V_n = \frac{1}{n^2} \sum_{i,j=1}^n h_p(Z_i, Z_j).$$

The asymptotic null distribution of the normalised V-Statistic nV_n, however, has no computable closed form.

Furthermore, care has to be taken when the Z_i exhibit correlation structure, as the null distribution significantly changes, impacting test significance. To model temporal structure of the observations we use τ-mixing, discussed in Section 2.2. τ-mixing is a notion of dependence within the observations,

weak enough for most practical applications. Trivially, i.i.d. observations are τ-mixing. As for Markov chains, whose convergence we study in the experiments, the property of geometric ergodicity implies τ-mixing cf. Section 2.2 for more discussion. For this work we will assume a technical condition $\sum_{t=1}^{\infty} t^2 \sqrt{\tau(t)} \leq \infty$.

We will use the wild bootstrap technique to addresses both, lack of a close form expression for the quantiles of the null distribution and temporal dependence. First, it allows us to estimate quantiles of the null distribution in order to compute test thresholds. Second, it accounts for correlation structure in the Z_i by mimicking it with an auxiliary random process: a simple Markov chain taking values in $\{-1, 1\}$, starting from $W_{1,n} = 1$,

$$W_{t,n} = \mathbf{1}(U_t > a_n) W_{t-1,n} - \mathbf{1}(U_t < a_n) W_{t-1,n},$$

where the U_t are uniform $[0, 1]$ i.i.d. random variables and a_n is the probability of $W_{t,n}$ changing sign (for i.i.d. data we may set $a_n = 0.5$). This leads to a bootstrapped V-statistic (cf. Section 2.2).

$$B_n = \frac{1}{n^2} \sum_{i,j=1}^{n} W_{i,n} W_{j,n} h(Z_i, Z_j).$$

Proposition 2, based of the Theorems 1, 2, 3 from Chapter 3, establishes that, under the null hypothesis, nB_n is a good approximation of nV_n, so it is possible to approximate quantiles of the null distribution by sampling from it. Under the alternative, however, V_n dominates B_n – resulting in almost sure rejection of the null hypothesis.

Proposition 2. Suppose h is Lipschitz continuous and $Eh_p(Z, Z) < \infty$. Under the null hypothesis nV_n and nB_n have the same limiting distribution (in a weak sense). Under the alternative hypothesis, B_n converges to zero, while V_n converges to a positive constant.

We propose the following test procedure for testing the null hypothesis that the Z_i are distributed according to the target distribution p.

- Calculate the test statistic nV_n.
- Obtain wild bootstrap samples $\{nB_n\}_{i=1}^{D}$ and estimate the $1-\alpha$ empirical

quantile of these samples.

- If nV_n exceeds the quantile, reject.

Proofs of the Main Results

We now prove the claims made in the previous Section.

Stein Operator in RKHS

We show in Lemma 12 that the expected value of the Stein operator is zero on the target measure.

Lemma 12. *Oates et al. [74, Lemma 1] If a random variable X is distributed according to p, under conditions on the kernel*

$$0 = \oint_{\partial\mathcal{X}} k(x,x')q(x)n(x)dS(x'),$$
$$0 = \oint_{\partial\mathcal{X}} \nabla_x k(x,x')^\top n(x')q(x')dS(x'),$$

and then for all $f \in \mathcal{F}$, the expected value of T is zero, i.e. $E_p(Tf)(X) = 0$.

This result was proved on bounded domains $\mathcal{X} \subset \mathbb{R}^d$ by Oates et al. [74, Lemma 1], where $n(x)$ is the unit vector normal to the boundary at x, and $\oint_{\partial\mathcal{X}}$ is the surface integral over the boundary $\partial\mathcal{X}$. The case of unbounded domains was discussed by Oates et al. [74, Remark 2]. Here we provide an alternative, elementary proof for the latter case.

Proof. First we show that the function $p \cdot f_i$ vanishes at infinity, by which we mean that for all dimensions j

$$\lim_{x_j \to \infty} p(x_1, \cdots, x_d) \cdot f_i(x_1, \cdots, x_d) = 0.$$

The density function p vanishes at infinity. The function f is bounded, which is implied by Cauchy-Schwarz inequality, $|f(x)| \leq \|f\| \sqrt{k(x,x)}$. This implies that the function $p \cdot f_i$ vanishes at infinity. We check that the expected value

$E_q(T_q)f(Z)$ is zero. For all dimensions i

$$
\begin{aligned}
& E_p \xi_p(X) \\
& E_p \left(\frac{\partial \log p(X)}{\partial x_i} f_i(X) + \frac{\partial f_i(X)}{\partial x_i} \right) \\
&= \int_{R_d} \left[\frac{\partial \log p(x)}{\partial x_i} f_i(x) + \frac{\partial f_i(x)}{\partial x_i} \right] p(x) dx \\
&= \int_{R_d} \left[\frac{1}{p(x)} \frac{\partial p(x)}{\partial x_i} f(x) + \frac{\partial f(x)}{\partial x_i} \right] p(x) dx \\
&= \int_{R_d} \left[\frac{\partial p(x)}{\partial x_i} f_i(x) + \frac{\partial f_i(x)}{\partial x_i} p(x) \right] dx \\
&\overset{(a)}{=} \int_{R_{d-1}} \left(\lim_{R \to \infty} p(x) f_i(x) \Big|_{x_i=-R}^{x_i=R} \right) dx_1 \cdots dx_{i-1} \cdots dx_{i+1} \cdots dx_d \\
&= \int_{R_{d-1}} 0 dx_1 \cdots dx_{i-1} \cdots dx_{i+1} \cdots dx_d \\
&= 0.
\end{aligned}
$$

For the equation (a) we have used integration by parts, the fact that $p(x)f_i(x)$ vanishes at infinity and Fubini-Toneli theorem to show that we can do iterated integration. The sufficient condition for the Fubini-Toneli theorem is that $E_q \langle f, \xi_p(Z) \rangle^2 < \infty$. This is true since $E_p \|\xi_p(X)\|^2 \leq E_p h_p(X, X) < \infty$. □

Proof of Theorem 5. $\xi_p(x, \cdot)$ is an element of the reproducing kernel Hilbert space $\mathcal{F}^d$ – by Steinwart and Christmann [102, Lemma 4.34] $\nabla k(x, \cdot) \in \mathcal{F}$, and $\frac{\partial \log p(x)}{\partial x_i}$ is just a scalar. We first show that $h_p(x, y) = \langle \xi_p(x, \cdot), \xi_p(y, \cdot) \rangle$. Using notations

$$
\begin{aligned}
\nabla_x k(x, \cdot) &= \left(\frac{\partial k(x, \cdot)}{\partial x_1}, \cdots, \frac{\partial k(x, \cdot)}{\partial x_d} \right) \\
\nabla_y k(\cdot, y) &= \left(\frac{\partial k(\cdot, y)}{\partial y_1}, \cdots, \frac{\partial k(\cdot, y)}{\partial y_d} \right),
\end{aligned}
$$

we calculate

$$
\begin{aligned}
\langle \xi_p(x, \cdot), \xi_p(y, \cdot) \rangle = & \nabla \log p(x)^\top \nabla \log p(y) k(x, y) \\
& + \nabla \log p(y)^\top \nabla_x k(x, y) \\
& + \nabla \log p(x)^\top \nabla_y k(x, y) \\
& + \langle \nabla_x k(x, \cdot), \nabla_y k(\cdot, y) \rangle_{\mathcal{F}^d}.
\end{aligned}
$$

Next we show that $\xi_p(x,\cdot)$ is Bochner integrable (see [102, Definition A.5.20]), which allows us to change order of the integration

$$E_q\|\xi_p(Z)\|_{\mathcal{F}^d} \leq E_q\|\xi_p(Z)\|^2_{\mathcal{F}^d} = E_q h_p(Z,Z) < \infty.$$

We next relate the expected value of the Stein operator to the inner product of f and the expected value of $\xi_q(Z)$,

$$E_q T_p f(Z) = \langle f, E_q\xi_p(Z)\rangle_{\mathcal{F}^d} = \sum_{i=1}^{d}\langle f_i, E_q\xi_{p,i}(Z)\rangle_{\mathcal{F}}. \tag{4.2}$$

We check the claim for all dimensions

$$\begin{aligned}
&\langle f_i, E_q\xi_{p,i}(Z)\rangle_{\mathcal{F}} \\
&= \left\langle f_i, E_q\left[\frac{\partial \log p(Z)}{\partial x_i}k(Z,\cdot) + \frac{\partial k(Z,\cdot)}{\partial x_i}\right]\right\rangle_{\mathcal{F}} \\
&= E_q\left\langle f_i, \frac{\partial \log p(Z)}{\partial x_i}k(Z,\cdot) + \frac{\partial k(Z,\cdot)}{\partial x_i}\right\rangle_{\mathcal{F}} \\
&= E_q\left[\frac{\partial \log p(Z)}{\partial x_i}f_i(Z) + \frac{\partial f_i(Z,\cdot)}{\partial x_i}\right].
\end{aligned}$$

The second equality follows from the fact that a linear operator $\langle f_i,\cdot\rangle_{\mathcal{F}}$ can be interchanged with the Bochner integral, and the fact that ξ_p is Bochner integrable. Using definition of $S(Z)$, Lemma (12) and Equation (4.2) we have

$$\begin{aligned}
S_p(Z) &:= \sup_{\|f\|<1} E_q(T_p f)(Z) - E_p(T_p f)(X) \\
&= \sup_{\|f\|<1} E_q(T_p f)(Z) \\
&= \sup_{\|f\|<1} \langle f, E_q\xi_p(Z)\rangle_{\mathcal{F}^d} \\
&= \|E_q\xi_p(Z)\|_{\mathcal{F}^d}.
\end{aligned}$$

We now calculate closed form formula for $S_p^2(Z)$,

$$\begin{aligned}
S_p^2(Z) &= \langle E_q\xi_p(Z), E_q\xi_p(Z)\rangle_{\mathcal{F}^d} = E_q\langle \xi_p(Z), E_q\xi_p(Z)\rangle_{\mathcal{F}^d} \\
&= E_q\langle \xi_p(Z), \xi_p(Z')\rangle_{\mathcal{F}^d} = E_q h_p(Z,Z'),
\end{aligned}$$

where Z' is an independent copy of Z.

□

Next, we prove that the discrepancy S discriminates different probability measures.

Proof of Theorem 6. If $p = q$ then $S_p(Z)$ is 0 by Lemma (12). Suppose $p \neq q$, but $S_p(Z) = 0$. If $S_p(Z) = 0$ then, by Theorem 5, $E_q \xi_p(Z) = 0$. In the following we substitute $\log p(Z) = \log q(Y) + [\log p(Z) - \log q(Y)]$,

$$\begin{aligned}
&E_q \xi_p(Z) \\
&= E_q \left(\nabla \log p(Z) k(Z, \cdot) + \nabla k(Z, \cdot) \right) \\
&= E_q \xi_q(Z) + E_q \left(\nabla [\log p(Z) - \log q(Y)] k(Z, \cdot) \right) \\
&= E_q \left(\nabla [\log p(Z) - \log q(Y)] k(Z, \cdot) \right)
\end{aligned}$$

We have used Theorem 5 and Lemma (12) to see that $E_q \xi_q(Z) = 0$, since $\| E_q \xi_q(Z) \|^2 = S_q(Z) = 0$.

We recognise that the expected value of $\nabla(\log p(Z) - \log q(Z)) k(Z, \cdot)$ is the mean embedding of a function $g(y) = \nabla \left(\log \frac{p(y)}{q(y)} \right)$ with respect to the measure q. By the assumptions function g is square integrable, therefore, since the kernel k is cc-universal, by Carmeli et al. [20, Theorem 4.4 c] its embedding is zero if and only if $g = 0$. This implies that

$$\nabla \log \frac{p(y)}{q(y)} = (0, \cdots, 0).$$

A constant vector field of derivatives can only be generated by a constant function, so $\log \frac{p(y)}{q(y)} = C$, for some C, which implies that $p(y) = e^C q(y)$. Since p and q both integrate to one, $C = 0$ then $p = q$, which is a contradiction. □

Wild Bootstrap Testing

Proof of proposition 2. We show that, under the alternative hypothesis, B_n converges to zero. We use Theorem 3, the assumption $\tau(r) = o(r^{-4})$ is satisfied since $\sum_{t=1}^{\infty} t^2 \sqrt{\tau(t)} \leq \infty$. We check the assumption

$$\sup_n \sup_{i,j<n} E_q h_p(Z_i, Z_j)^2 < \infty.$$

We have $E_q h_p(Z, Z')^2 \leq \left(E_q \|\xi_p(Z)\|^2\right)^2 = \left(E_q h_p(Z, Z)\right)^2 < \infty$.

We show that, under the alternative hypothesis, V_n converges to a positive constant. We use Theorem 2, the zero component of h is positive since $S(Z)^2 > 0$. We checked the assumption $\sup_n \sup_{i \in N^2} Eh(Z_i)^2 < \infty$ above.

We show that, under the null hypothesis V_n and B_n have the same limiting distribution. We use Theorem 2, by the assumptions h is Lipschitz continuous. h is kernel since it can be written as $h(x, y) = \langle \xi(x), \xi(y) \rangle$ and it is one degenerate under the null, since

$$Eh(Z, y) = \langle E\xi(Z), \xi(y) \rangle = 0.$$

We checked the assumption $\sup_n \sup_{i,j<n} E_q h_p(Z_i, Z_j)^2 < \infty$. above.

Finally we check the bootstrap assumption (Definition 1) : $\{W_{t,n}\}_{1 \leq t \leq n}$ is a row-wise strictly stationary triangular array independent of all Z_t such that $EW_{t,n} = 0$ and $\sup_n E|W_{t,n}^{2+\sigma}| = 1 < \infty$ for some $\sigma > 0$. The auto-covariance of the process is given by $EW_{s,n}W_{t,n} = (1 - 2a_n)^{-|s-t|}$, so the function $\rho(x) = \exp(-x)$, and $l_n = \log(1 - 2a_n)^{-1}$. We verify that $\lim_{u \to 0} \rho(u) = 1$. If we set $a_n = w_n^{-1}$, such that $w_n = o(\sqrt{n})$ and $\lim_{n \to \infty} w_n = \infty$, then $l_n = O(w_n)$ and $\sum_{r=1}^{n-1} \rho(|r|/l_n) = \frac{1-(1-2a_n)^{n+1}}{p_n} = O(w_n) = O(l_n)$. □

As a consequence, if the null hypothesis is true, we can approximate any quantile; while under the alternative hypothesis, all quantiles of B_n collapse to zero while $P(V_n > 0) \to 1$.

Experiments

We provide a number of experimental applications for our test. We begin with a simple check to establish correct test calibration on non-i.i.d. data, followed by a demonstration of statistical model criticism for Gaussian process (GP) regression. We then apply the proposed test to quantify bias-variance trade-offs in MCMC, and demonstrate how to use the test to verify whether MCMC samples are drawn from the desired stationary distribution. In the final experiment, we move away from the MCMC setting, and use the test to evaluate the convergence of a nonparametric density estimator.

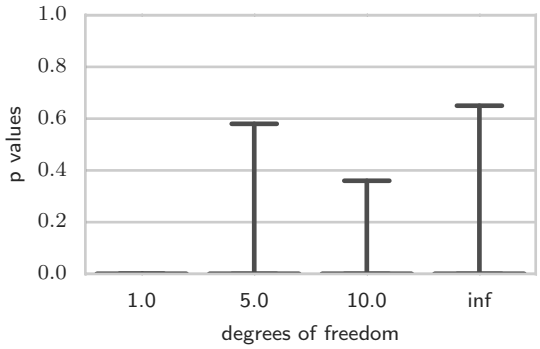

Figure 4.1: Large autocovariance, unsuitable bootstrap. The parameter a_n is too large and the bootstrapped V-statistics B_n are too low on average. Therefore, it is very likely that $V_n > B_n$ and the test is too conservative.

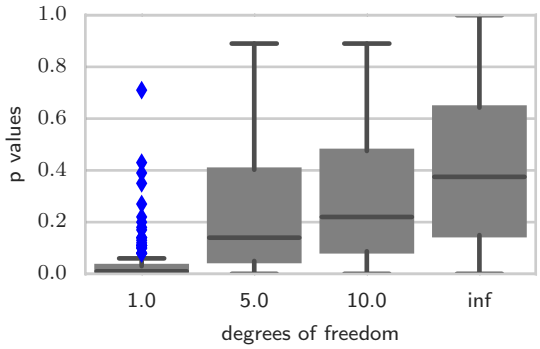

Figure 4.2: Large autocovariance, suitable bootstrap. The parameter $a_n$is chosen suitably, but due to a large autocorrelation within the samples, the power of the test is small (effective sample size is small).

Student's t vs Normal

In our first task, we modify experiment 4.1 from Gorham and Mackey 40. The null hypothesis is that the observed samples come from a standard normal distribution. We study the power of the test against samples from a Student's t distribution. We expect to observe low p-values when testing against a Student's t distribution with few degrees of freedom. We consider 1, 5, 10 or ∞ degrees of freedom, where ∞ is equivalent to sampling from a standard normal distribution. For a fixed number of degrees of freedom we draw 1400 samples and calculate the p-value. This procedure is repeated 100 times, and the bar plots of p-values are shown in Figures 4.1,4.2,4.3.

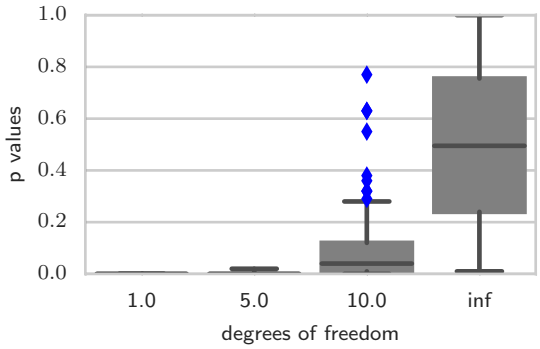

Figure 4.3: Thinned sample, suitable bootstrap. Most of the autocorrelation within the sample is canceled by thinning. To guarantee that the remaining autocorrelation is handled properly, the flip probability is set at 0.1.

Our twist on the original experiment 4.1 by Gorham and Mackey 40 is that the draws from the Student's t distribution exhibit temporal correlation. We generate samples using a Metropolis–Hastings algorithm, with a Gaussian random walk with variance 1/2. We emphasise the need for an appropriate choice of the wild bootstrap process parameter a_n. In Figure 4.1 we plot p-values for a_n being set to 0.5. Such a high value of a_n is suitable for i.i.d. observations, but results in p-values that are too conservative for temporally correlated observations. In Figure 4.2, we set $a_n = 0.02$, which gives a well calibrated distribution of the p-values under the null hypothesis, however, the test power is reduced. Indeed, p-values for five degrees of freedom are already large. The solution that we recommend is a mixture of thinning and adjusting a_n, as presented in the Figure 4.3. We thin the observations by a factor of 20 and set $a_n = 0.1$, thus preserving both good statistical power and correct calibration of p-values under the null hypothesis. In a general, we recommend to thin a chain so that $\text{Cor}(X_t, X_{t-1}) < 0.5$, set $a_n = 0.1/k$, and run test with at least $\max(500k, d100)$ data points, where $k < 10$.

Comparing to a parametric test in increasing dimensions.

In this experiment, we compare with the test proposed by [8], which essentially is an MMD test for normality, i.e. the null hypothesis is that Z is a d-dimensional standard normal random variable. We set the sample size to $n = 500, 1000$ and $a_n = 0.5$, generate

$$Z \sim \mathcal{N}(0, I_d) \qquad Y \sim U[0,1]$$

	d	**2**	**5**	**10**	**15**	**20**	**25**
B&H	$n = 500$	1	1	**1**	**0.86**	**0.29**	**0.24**
Stein		1	1	0.86	0.39	0.05	0.05
B&H	$n = 1000$	1	1	1	**1**	**0.87**	**0.62**
Stein		1	1	1	0.77	0.25	0.05

Table 4.1: Test power vs. sample size for the test by [8] (B&H) and our Stein based test.

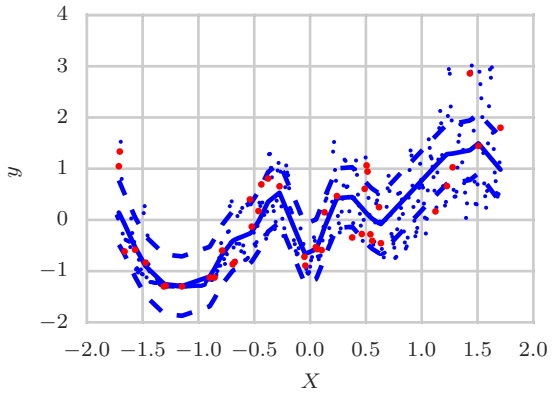

Figure 4.4: Fitted GP and data used to fit (blue) and to apply test (red).

and modify $Z_0 \leftarrow Z_0 + Y$. Table 4.1 shows the power as a function of the sample size. We observe that for higher dimensions, and where the expectation of the kernel exists in closed form, an MMD-type test like [8] is a better choice.

Statistical Model Criticism on Gaussian Processes

This experiment was conducted by co-author of the article on which this chapter is based, Heiko Strathmann.

We next apply our test to the problem of statistical model criticism for GP regression. Our presentation and approach are similar to the non i.i.d. case of Lloyd and Ghahramani [Section 6 70]. We use the `solar` dataset, consisting of a $d = 1$ regression problem with $N = 402$ pairs (X, y). We fit $N_{\text{train}} = 361$ data using a GP with an exponentiated quadratic kernel and a Gaussian noise model, and perform standard maximum likelihood II on the hyperparameters (length-scale, overall scale, noise-variance). We then apply our test to the remaining $N_{\text{test}} = 41$ data. The test attempts to falsify the null hypothesis that the `solar` dataset was generated from the plug-in predictive distribution (conditioned on

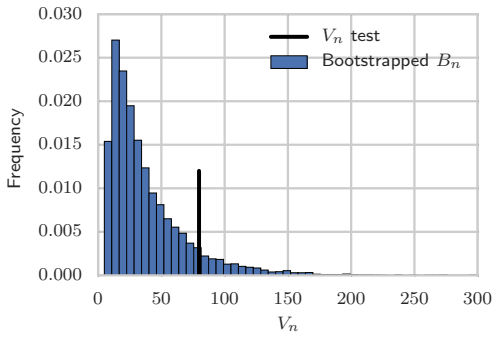

Figure 4.5: Bootstrapped B_n distribution with the test statistic V_n marked.

training data and predicted position) of the GP. Lloyd and Ghahramani refer to this setup as non i.i.d., since the predictive distribution is a different univariate Gaussian for every predicted point. Our particular $N_{\text{train}}, N_{\text{test}}$ were chosen to make sure the GP fit has stabilised, i.e. adding more data did not cause further model refinement.

Figure 4.4 shows training and testing data, and the fitted GP. Clearly, the Gaussian noise model is a poor fit for this particular dataset, e.g. around $X = -1$. Figure 4.5 shows the distribution over $D = 10000$ bootstrapped V-statistics B_n with $n = N_{\text{test}}$. The test statistic lies in an upper quantile of the bootstrapped null distribution, correctly indicating that it is unlikely the test points were generated by the fitted GP model, even for the low number of test data observed, $n = 41$.

In a second experiment, we compare against Lloyd and Ghahramani: we compute the MMD statistic between test data $(X_{\text{test}}, y_{\text{test}})$ and $(X_{\text{test}}, y_{\text{rep}})$, where y_{rep} are samples from the fitted GP. We draw 10000 samples from the null distribution by repeatedly sampling new $\tilde{y}_{\text{rep}}$ from the GP plug-in predictive posterior, and comparing $(X_{\text{test}}, \tilde{y}_{\text{rep}})$ to $(X_{\text{test}}, y_{\text{rep}})$. When averaged over 100 repetitions of randomly partitioning (X, y) for training and testing, our goodness of fit test produces a p-value that is statistically not significantly different from the MMD method ($p \approx 0.1$, note that this result is subject to $N_{\text{train}}, N_{\text{test}}$). We emphasise, however, that Lloyd and Ghahramani's test requires to sample from the fitted model (here 10000 null samples were required in order to achieve stable p-values). Our test *does not* sample from the GP at all and completely side-steps this more costly approach.

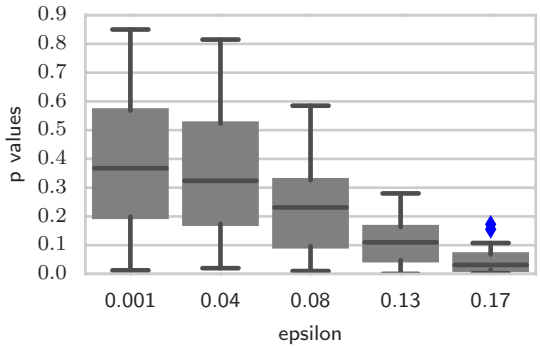

Figure 4.6: Distribution of p-values as a function of ϵ for austerity MCMC.

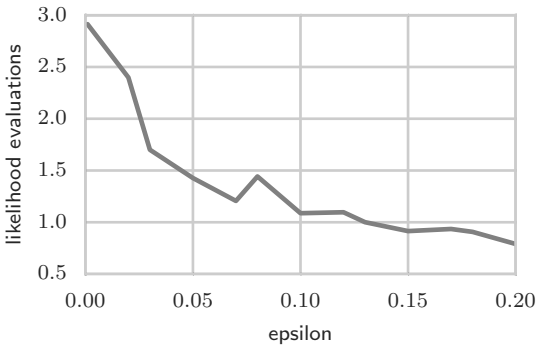

Figure 4.7: Average number of likelihood evaluations a function of ϵ for austerity MCMC (the y-axis is in millions of evaluations).

Bias quantification in Approximate MCMC

We now illustrate how to quantify bias-variance trade-offs in an approximate MCMC algorithm austerity MCMC [64]. For the purpose of illustration we use a simple generative model from Gorham and Mackey [40], Welling and Teh [112],

$$\theta_1 \sim \mathcal{N}(0,10); \theta_2 \sim \mathcal{N}(0,1)$$
$$X_i \sim \frac{1}{2}\mathcal{N}(\theta_1,4) + \frac{1}{2}\mathcal{N}(\theta_2,4).$$

Austerity MCMC is a Monte Carlo procedure designed to reduce the number of likelihood evaluation in the acceptance step of the Metropolis-Hastings algorithm. The crux of method is to look at only a subset of the data, and make an acceptance/rejection decision based on this subset. The probability of mak-

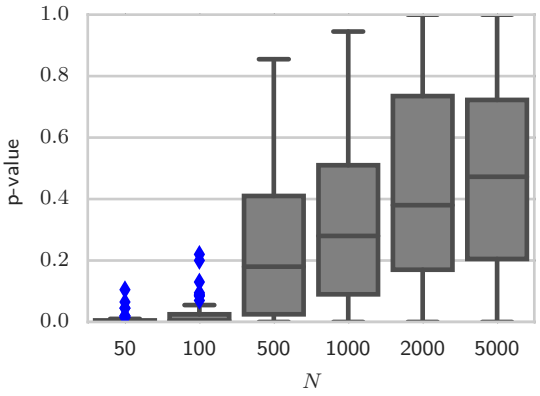

Figure 4.8: Density estimation: P-values for an increasing number of data N for the nonparametric model.

ing a wrong decision is proportional to a parameter $\epsilon \in [0,1]$. This parameter influences the time complexity of austerity MCMC: when ϵ is larger, i.e., when there is a greater tolerance for error, the expected computational cost is lower. We simulate $\{X_i\}_{1\leq i\leq 400}$ points from the model with $\theta_1 = 0$ and $\theta_2 = 1$. In our experiment, there are two modes in the posterior distribution: one at $(0,1)$ and the other at $(1,-1)$. We run the algorithm with ϵ varying over the range $[0.001, 0.2]$. For each ϵ we calculate an individual thinning factor, such that correlation between consecutive samples from the chains is smaller than 0.5 (greater ϵ generally required more thinning). For each ϵ we test the hypothesis that $\{\theta_i\}_{1\leq i\leq 500}$ is drawn from the true stationary posterior, using our goodness of fit test. We generate 100 p-values for each ϵ , as shown in Figure 4.6. $\epsilon = 0.4$ yields a good approximation of the true stationary distribution, while being parsimonious in terms of likelihood evaluations, as shown in Figure 4.7.

Convergence in nonparametric density estimation

In our final experiment, we apply our goodness of fit test to measuring quality-of-fit in nonparametric density estimation. We evaluate two density models: the infinite dimensional exponential family [100], and a recent approximation to this model using random Fourier features [106]. Our implementation of the model assumes the log density to take the form $f(x)$, where f lies in a RKHS induced by a Gaussian kernel with bandwidth 1. We fit the model using N observations drawn from a standard Gaussian, and perform our quadratic time test on a separate evaluation dataset of fixed size $n = 500$. Our goal is

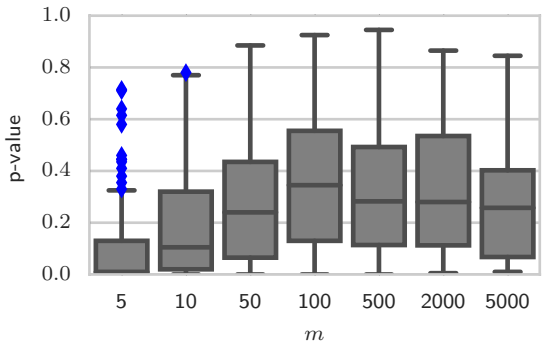

Figure 4.9: Approximate density estimation: P-values for an increasing number of random features m.

to identify N sufficiently large that the goodness of fit test does not reject the null hypothesis (i.e., the model has learned the density sufficiently well, bearning in mind that it is guaranteed to converge for sufficiently large N). Figure 4.8 shows how the distribution of p-values evolves as a function of N; this distribution is uniform for $N = 5000$, but at $N = 500$, the null hypothesis would very rarely be rejected.

We next consider the random Fourier feature approximation to this model, where the log pdf f, is approximated using a finite dictionary of random Fourier features [79]. The natural question when using this approximation is: "How many random features are needed?" Using the same test set size $n = 500$ as above, and a large number of available samples $N = 5 \cdot 10^4$, Figure 4.9 shows the distributions of p-values for an increasing number of random features m. From $m = 50$, the null hypothesis would rarely be rejected, given a reasonable choice of test level. Note, however, that the p-values do *not* have a uniform distribution, even for a large number of random features. This subtle effect is caused by over-smoothing due to the regularisation approach taken in [106, KMC finite], which would not otherwise have been detected.

Chapter 5

Fast Analytic Functions Based Two Sample Test.

This chapter is based on Kacper Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, pages 1972–1980, 2015.

In this chapter we propose a class of nonparametric two-sample tests with a cost linear in the sample size. Two tests are given, both based on an ensemble of distances between analytic functions representing each of the distributions. The first test uses smoothed empirical characteristic functions to represent the distributions, the second uses distribution embeddings in a reproducing kernel Hilbert space. Analyticity implies that differences in the distributions may be detected almost surely at a finite number of randomly chosen locations/frequencies. The new tests are consistent against a larger class of alternatives than the previous linear-time tests based on the (non-smoothed) empirical characteristic functions, while being much faster than the current state-of-the-art quadratic-time kernel-based or energy distance-based tests. Experiments on artificial benchmarks and on challenging real-world testing problems demonstrate that our tests give a better power/time tradeoff than competing approaches, and in some cases, better outright power than even the most expensive quadratic-time tests. This performance advantage is retained even in high dimensions, and in cases where the difference in distributions is not observable with low order statistics.

Analytic embeddings and distances

In this section we consider mappings from the space of probability measures into a sub-space of real valued analytic functions. We will show that evaluating these maps at J randomly selected points is almost surely injective for any $J > 0$. Using this result, we obtain a simple (randomized) metrization of the space of probability measures. This metrization is used in the next section to construct linear-time nonparametric two-sample tests.

Pseudometrics based on characteristic functions. A practical limitation when using the MMD in testing is that an empirical estimate is expensive to compute, this being the sum of two U-statistics and an empirical average, with cost quadratic in the sample size [45, Lemma 6]. We might instead consider a finite dimensional approximation to the MMD, achieved by estimating the integral (2.9), with the random variable

$$d^2_{\varphi,J}(P,Q) = \frac{1}{J}\sum_{j=1}^{J} |\varphi_P(T_j) - \varphi_Q(T_j)|^2, \tag{5.1}$$

where $\{T_j\}_{j=1}^J$ are sampled independently from the distribution with a density function $F^{-1}\kappa$. This type of approximation is applied to various kernel algorithms under the name of *random Fourier features* [79, 65]. In the statistical testing literature, the quantity $d_{\varphi,J}(P,Q)$ predates the MMD by a considerable time, and was studied in [54, 55, 31], and more recently revisited in [117]. Our first proposition is that $d^2_{\varphi,J}(P,Q)$ can be a poor choice of distance between probability measures, as it fails to distinguish a large class of measures. The following result is proved in the Appendix.

Proposition 3. *Let $J \in \mathbb{N}$ and let $\{T_j\}_{j=1}^J$ be a sequence of real valued i.i.d. random variables with a distribution which is absolutely continuous with respect to the Lebesgue measure. For any $0 < \epsilon < 1$, there exists an uncountable set $\mathcal{A}$ of mutually distinct probability measures (on the real line) such that for any $P, Q \in \mathcal{A}$, $\mathbb{P}\left(d^2_{\varphi,J}(P,Q) = 0\right) \geq 1 - \epsilon$.*

We are therefore motivated to find distances of the form (5.1) that can distinguish larger classes of distributions, yet remain efficient to compute. These distances are characterized as follows:

Definition 4 (Random Metric)**.** *A random process d with values in $\mathbf{R}$, indexed*

with pairs from the set of probability measures $\mathcal{M}$, i.e., $d = \{d(P,Q) : P,Q \in \mathcal{M}\}$, is said to be a random metric if it satisfies all the conditions for a metric with qualification 'almost surely'. Formally, for all $P,Q,R \in \mathcal{M}$, random variables $d(P,Q), d(P,R), d(R,Q)$ must satisfy

1. $d(P,Q) \geq 0$ *a.s.*
2. *if* $P = Q$, *then* $d(P,Q) = 0$ *a.s, if* $P \neq Q$ *then* $d(P,Q) \neq 0$ *a.s.*
3. $d(P,Q) = d(Q,P)$ *a.s.*
4. $d(P,Q) \leq d(P,R) + d(R,Q)$ *a.s.* [1]

From the statistical testing point of view, the coincidence axiom of a metric d, $d(P,Q) = 0$ if and only if $P = Q$, is key, as it ensures consistency against all alternatives. The quantity $d_{\varphi,J}(P,Q)$ in (5.1) violates the coincidence axiom, so it is only a random pseudometric (other axioms are trivially satisfied). We remedy this problem by replacing the characteristic functions by smooth characteristic functions:

Definition 5. *A smooth characteristic function $\phi_P(t)$ of a measure P is a characteristic function of P convolved with an analytic smoothing kernel l, i.e.*

$$\phi_P(t) = \int_{\mathbf{R}^d} \varphi_P(w) l(t-w) dw, \qquad t \in \mathbf{R}^d. \tag{5.2}$$

Proposition 5 shows that smooth characteristic function can be estimated in a linear time. The analogue of $d_{\varphi,J}(P,Q)$ for smooth characteristic functions is simply

$$d^2_{\phi,J}(P,Q) = \frac{1}{J} \sum_{j=1}^{J} |\phi_P(T_j) - \phi_Q(T_j)|^2,$$

where $\{T_j\}_{j=1}^J$ are sampled independently from the absolutely continuous distribution (returning to our earlier example, this might be $F^{-1}\kappa(t)$ if we believe this to be an informative choice). The following theorem, proved in the Appendix, demonstrates that the smoothing greatly increases the class of distributions we can distinguish.

Theorem 7. *Let l be an analytic, integrable kernel with an inverse Fourier transform that is non-zero almost everywhere. Then, for any $J > 0$, $d_{\phi,J}$ is a random*

[1] Note that this does not imply that realizations of d are distances on $\mathcal{M}$, but it does imply that they are almost surely distances for all arbitrary finite subsets of $\mathcal{M}$.

metric on the space of probability measures with integrable characteristic functions, and ϕ_P is an analytic function.

This result is primarily a consequence of analyticity of smooth characteristic functions and the fact that analytic functions are 'well behaved'. There is an additional, practical advantage to smoothing: when the variability in the difference of the characteristic functions is high, and these differences are local, smoothing distributes the difference in CFs more broadly in the frequency domain (a simple illustration is in Fig. 5.1), making it easier to find by measurement at a small number of randomly chosen points. This accounts for the observed improvements in test power in Section A.4, over differences in unsmoothed CFs.

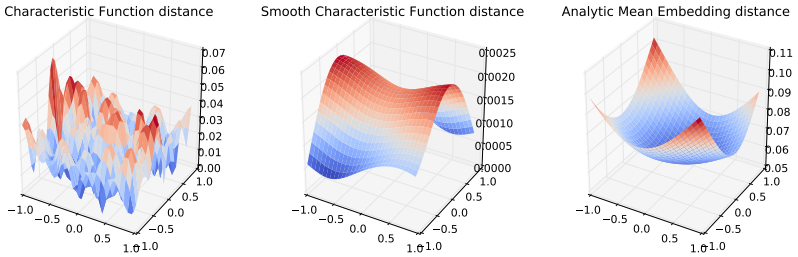

Figure 5.1: Smooth vs non-smooth. Left: pseudo-distance $d_{\varphi,1}(P,Q)$ which uses a single frequency $t \in \mathbf{R}^2$ as a function of this frequency; **Middle**: $d_{\phi,1}(P,Q)$ depicted in the same way; **Right**: $d_{\mu,1}(P,Q)$ which uses a single location $t \in \mathbf{R}^2$ as a function of this location. The measures P,Q used are illustrated in Figure 5.6 - these are grids of Gaussian distributions discussed in detail in Section 5.4.

Metrics based on mean embeddings. The key step which leads us to the construction of a random metric $d_{\phi,J}$ is the convolution of the original characteristic functions with an analytic smoothing kernel. This idea needs not be restricted to the representations of probability measures in the frequency domain. We may instead directly convolve the probability measure with a positive definite kernel k (that needs not be translation invariant), yielding its mean embedding into the associated RKHS,

$$\mu_P(t) = \int_E k(x,t)dP(x).$$

We say that a positive definite kernel $k : \mathbf{R}^D \times \mathbf{R}^D \to \mathbf{R}$ is analytic on its domain if for all $x \in \mathbf{R}^D$, the feature map $k(x,\cdot)$ is an analytic function on $\mathbf{R}^D$. By using embeddings with *characteristic and analytic* kernels, we obtain partic-

ularly useful representations of distributions. As for the smoothed CF case, we define

$$d^2_{\mu,J}(P,Q) = \frac{1}{J}\sum_{j=1}^{J}(\mu_P(T_j) - \mu_Q(T_j))^2.$$

The following theorem ensures that $d_{\mu,J}(P,Q)$ is also a random metric.

Theorem 8. *Let k be an analytic, integrable and characteristic kernel. Then for any $J > 0$, $d_{\mu,J}$ is a random metric on the space of probability measures (and μ_P is an analytic function).*

Note that this result is stronger than the one presented in Theorem 7, since it is not restricted to the class of probability measures with integrable characteristic functions. Indeed, the assumption that the characteristic function is integrable implies the existence and boundedness of a density. Recalling the representation of MMD in (2.11), we have proved that it is almost always sufficient to measure difference between μ_P and μ_Q at a finite number of points, provided our kernel is characteristic and analytic. In the next section, we will see that metrization of the space of probability measures using random metrics $d_{\mu,J}$, $d_{\phi,J}$ is very appealing from the computational point of view. It turns out that the statistical tests that arise from these metrics have linear time complexity (in the number of samples) and constant memory requirements.

Hypothesis Tests Based on Distances Between Analytic Functions

In this section, we provide two linear-time two-sample tests: first, a test based on analytic mean embeddings, and next a test based on smooth characteristic functions. We further describe the relation with competing alternatives. Proofs of all propositions are in Appendix 5.3.

Difference in analytic functions In the previous section we described the random metric based on a difference in analytic mean embeddings, $d^2_{\mu,J}(P,Q) = \frac{1}{J}\sum_{j=1}^{J}(\mu_P(T_j) - \mu_Q(T_j))^2$. If we replace μ_P with the empirical mean embedding $\hat{\mu}_P = \frac{1}{n}\sum_{i=1}^{n} k(X_i,\cdot)$ it can be shown that for any sequence of unique $\{t_j\}_{j=1}^{J}$, under the null hypothesis, as $n \to \infty$,

$$\sqrt{n}\sum_{j=1}^{J}(\hat{\mu}_P(t_j) - \hat{\mu}_Q(t_j))^2 \tag{5.3}$$

converges in distribution to a sum of correlated chi-squared variables. Even

for fixed $\{t_j\}_{j=1}^J$, it is very computationally costly to obtain quantiles of this distribution, since this requires a bootstrap or permutation procedure. We will follow a different approach based on Hotelling's T^2-statistic [59]. The Hotelling's T^2-squared statistic of a normally distributed, zero mean, Gaussian vector $W = (W_1, \cdots, W_J)$, with a covariance matrix Σ, is $T^2 = W\Sigma^{-1}W$. The compelling property of the statistic is that it is distributed as a χ^2-random variable with J degrees of freedom. To see a link between T^2 and equation (5.3), consider a random variable $\sum_{i=j}^J W_j^2$: this is also distributed as a sum of correlated chi-squared variables. In our case W is replaced with a difference of normalized empirical mean embeddings, and Σ is replaced with the empirical covariance of the difference of mean embeddings. Formally, let Z_i denote the vector of differences between kernels at tests points T_j,

$$Z_i = (k(X_i, T_1) - k(Y_i, T_1), \cdots, k(X_i, T_J) - k(Y_i, T_J)) \in \mathbf{R}^J.$$

We define the vector of mean empirical differences $W_n = \frac{1}{n}\sum_{i=1}^n Z_i$, and its covariance matrix $\Sigma_n = \frac{1}{n}\sum_i (Z_i - W_n)(Z_i - W_n)^T$. The test statistic is

$$S_n = nW_n\Sigma_n^{-1}W_n.$$

The computation of S_n requires inversion of a $J \times J$ matrix Σ_n, but this is fast and numerically stable: J will typically be small, and is less than 10 in our experiments. The next proposition demonstrates the use of S_n as a two-sample test statistic.

Proposition 4 (Asymptotic behavior of S_n)**.** *Let $d^2_{\mu,J}(P,Q) = 0$ a.s. and let $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ be i.i.d. samples from P and Q respectively. If Σ_n^{-1} exists for n large enough, then the statistic S_n is a.s. asymptotically distributed as a χ^2-random variable with J degrees of freedom (as $n \to \infty$ with d fixed). If $d^2_{\mu,J}(P,Q) > 0$ a.s., then a.s. for any fixed r, $\mathbb{P}(S_n > r) \to 1$ as $n \to \infty$.*

We now apply the above proposition to obtain a statistical test.

Test 1 (Analytic mean embedding)**.** *Calculate S_n. Choose a threshold r_α corresponding to the $1-\alpha$ quantile of a χ^2 distribution with J degrees of freedom, and reject the null hypothesis whenever S_n is larger than r_α.*

There are a number of valid sampling schemes for the test points $\{T_j\}_{j=1}^J$ to evaluate the differences in mean embeddings: see Section 5.5 for a discussion.

Difference in smooth characteristic functions From the convolution definition of a smooth characteristic function (5.2) it is not immediately obvious how to calculate its estimator in linear time. In the next proposition, however, we show that a smooth characteristic function is an expected value of some function (with respect to the given measure), which can be estimated in a linear time.

Proposition 5. *Let k be an integrable translation-invariant kernel and f its inverse Fourier transform. Then the smooth characteristic function of P can be written as $\phi_P(t) = \int_{\mathbf{R}^d} e^{it^\top x} f(x) dP(x)$.*

It is now clear that a test based on the smooth characteristic functions is similar to the test based on mean embeddings. The main difference is in the definition of the vector of differences Z_i:

$$Z_i = (f(X_i)\sin(X_i T_1) - f(Y_i)\sin(Y_i T_1), f(X_i)\cos(X_i T_1) - f(Y_i)\cos(Y_i T_1), \cdots) \in \mathbf{R}^{2J}$$

The imaginary and real part of the $e^{\sqrt{-1}T_j^\top X_i} f(X_i) - e^{\sqrt{-1}T_j^\top Y_i} f(Y_i)$ are stacked together, in order to ensure that W_n, Σ_n and S_n as all real-valued quantities.

Proposition 6. *Let $d^2_{\phi,J}(P,Q) = 0$ and let $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ be i.i.d. samples from P and Q respectively. Then the statistic S_n is almost surely asymptotically distributed as a χ^2-random variable with $2J$ degrees of freedom (as $n \to \infty$ with J fixed). If $d^2_{\phi,J}(P,Q) > 0$, then almost surely for any fixed r, $P(S_n > r) \to 1$ as $n \to \infty$.*

Other tests. The test [31] based on empirical characteristic functions was constructed originally for one test point and then generalized to many points - it is quite similar to our second test, but does not perform smoothing (it is also based on a T^2-Hotelling statistic). The block MMD [114] is a sub-quadratic test, which can be trivially linearized by fixing the block size, as presented in the Appendix. Finally, another alternative is the MMD, an inherently quadratic time test. We scale MMD to linear time by sub-sampling our data set, and choosing only $\sqrt{n}$ points, so that the MMD complexity becomes $O(n)$. Note, however, that the true complexity of MMD involves a permutation calculation of the null distribution at cost $O(b_n n)$, where the number of permutations b_n grows with n. See Appendix 5.4 for a detailed description of block MMD.

Proofs

Proof of Proposition 3

Proof. Since T_i are supported on the whole real line, for each ϵ there exists real number I such that $P(T_i \in [-I, I]) < 1 - (1-\epsilon)^{\frac{1}{J}}$. We now construct a family of triangle characteristic functions, centered around zero, that differ only on the interval $[-I, I]$.

Define a triangle function $f_w(t) = max(1 - w|t|, 0)$. By Polya's theorem, $\mathcal{A} = \{f_w\}_{w > \frac{1}{I}}$ is an uncountable family of different characteristic functions.

These functions are the same (equal to zero) on the complement of $[-I, I]$. The probability $P(T_i \notin [-I, I]) > (1-\epsilon)^{\frac{1}{J}}$. The probability that all T_i sit outside of the interval $[-I, I]$ is greater than $\left((1-\epsilon)^{\frac{1}{J}}\right)^J = 1 - \epsilon$. If all T_i sit outside of the interval $[-I, I]$, $S^2_{\varphi,J} = 0$. Therefore $P(S^2_{\varphi,J} = 0) > 1 - \epsilon$.

□

Proof of Theorem 8

First we give a proposition that characterizes limits of analytic functions.

Proposition 7 ([25, Proposition 3])**.** *If $\{f_n\}$ is a sequence of real valued, uniformly bounded analytic functions on $\mathbf{R}^d$ converging pointwise to f, then f is analytic.*

Now we characterize the RKHS of an analytic kernel. Similar results were proved for specific classes of kernels in [108, Theorem 1], [105, Corollary 3.5].

Lemma 13. *If k is a bounded, analytic kernel on $\mathbf{R}^d \times \mathbf{R}^d$, then all functions in the RKHS $\mathcal{H}_k$ associated with this kernel are analytic.*

Proof. Since $\mathbf{R}^d$ is separable then by [103, Lemma 4.33] Hilbert Space $\mathcal{H}_k$ is separable. Therefore, by Moore-Aronszajn Theorem [13] there exist a set H_0 of linear combinations of functions $k(\cdot, x), x \in \mathbf{R}^d$, which is dense in $\mathcal{H}_k$ and $\mathcal{H}_k$ is a set of functions which are limits of Cauchy sequences in H_0. For $f \in \mathcal{H}_k$ let $\{f_n\} \in \mathcal{H}_0$ be a sequence of functions converging in the RKHS norm to f. Since $\{f_n\}$ is convergent there exists N such that $\forall n > N \ \|f_n - f\| \leq 1$. For all

n there exist a uniform bound $\max(1, \max_{1\leq i\leq N} \|f_i\|) + \|f\|$ on norm of f_n

$$\|f_n\| = \|f_n - f + f\| \leq \|f_n - f\| + \|f\| \leq \max(1, \max_{1\leq i\leq N} \|f_i\|) + \|f\|.$$

Since k is bounded, by the [103, Lemma 4.33], there exists C such that for any $f \in \mathcal{H}_k$, $\|f\|_\infty \leq C\|f\|$. Therefore for all n

$$\|f_n\|_\infty \leq C \max(1, \max_{1\leq i\leq N} \|f_i\|) + C\|f\|.$$

Finally, using Proposition 7 we conclude that f is analytic. We check assumptions – convergence in norm in RKHS implies pointwise convergence, the sequence is uniformly bounded by $C \max(1, \max_{1\leq i\leq N} \|f_i\|) + C\|f\|$ and each element of the sequence f_n is analytic as a linear combinations of analytic functions $k(\cdot, x), x \in \mathbf{R}^d$. □

Next, we show that analytic functions are 'well behaved'.

Lemma 14. *Let μ be absolutely continuous measure on $\mathbf{R}^d$ (wrt. the Lebesgue measure). Non-zero, analytic function f can be zero at most at the set of measure 0, with respect to the measure μ.*

Proof. If f is zero at the set with a limit point then it is zero everywhere. Therefore f can be zero at most at a set A without a limit point, which by definition is a discrete set (distance between any two points in A is greater then some $\epsilon > 0$). Discrete sets have zero Lebesgue measure (as a countable union of points with zero measure). Since P is absolutely continuous then $\mu(A)$ is zero as well. □

Next, we show how to construct random distances.

Lemma 15. *Let Λ be an injective mapping from the space of the probability measures into a space of analytic functions on $\mathbf{R}^d$. Define*

$$d^2_{\Lambda,J}(P,Q) = \sum_{j=1}^{J} \left| [\Lambda P](T_j) - [\Lambda Q](T_j) \right|^2$$

where $\{T_j\}_{j=1}^J$ are real valued i.i.d. random variables from a distribution which is absolutely continuous with respect to the Lebesgue measure. Then, $d^2_{\Lambda,J}(P,Q)$ is a random metric.

Proof. Let ΛP and ΛQ be images of measures P and Q respectively. We want to apply Lemma 14 to the analytic function $f = \Lambda P - \Lambda Q$, with the measure $\mu = \mu_{T_i}$, to see that if $P \neq Q$ then $f \neq 0$ a.e. To do so, we need to show that $P \neq Q$ implies that f is non-zero. Since mapping to Λ is injective, there must exists at least one point o where f is non-zero. By continuity of f, there exists a ball around o in which f is non-zero. Now we use Lemma 14 to infer that $f \neq 0$ a.e.

$f \neq 0$ a.e. implies that $d_{\Lambda,J}(P,Q) > 0$ a.s. If $P = Q$ then $f = 0$ and $d_{\Lambda,J}(P,Q) = 0$.

By the construction $d_{\Lambda,J}(P,Q) = d_{\Lambda,J}(Q,P)$ and for any measure U, $d_{\Lambda,J}(P,Q) \leq d_{\Lambda,J}(P,U) + d_{\Lambda,J}(U,Q)$ a.s. since the triangle inequality holds for any vectors in $\mathbf{R}^J$. □

We are ready to prove Theorem 8.

Proof of Theorem 8. Since k is characteristic, the mapping $\Lambda : P \to \mu_P$ is injective. μ_P is an element of RKHS associated with k. Since k is a bounded, analytic kernel on $\mathbf{R}^d \times \mathbf{R}^d$, Lemma 13 guarantees that μ_P is analytic, hence the image of Λ is a subset of analytic functions. Therefore, we can use Lemma 15 to see that $d_{\Lambda,J}(P,Q)^2 = d_{\mu,J}(P,Q)^2$ is a random metric. □

Proof of Theorem 7

We first show that smooth characteristic functions are unique to distributions.

Lemma 16. *If l is an analytic, integrable, translation invariant kernel with an inverse Fourier transform strictly greater then zero and P has integrable characteristic function, then the mapping*

$$\Lambda : P \to \phi_P$$

is injective and ϕ_P is element of the RKHS $\mathcal{H}_l$ associated with l.

Proof. For the integrable characteristic function φ we define a functional $L : \mathcal{H}_l \to R$ given by formula

$$Lf = \int_{\mathbf{R}^d} \varphi(x) f(x) dx$$

Since $L(f+g) = L(f) + L(g)$, L is linear. We check that L is bounded; let $B = \{f \in \mathcal{H}_l : \| f \| \leq 1\}$ be a unit ball in the Hilbert Space.

$$\sup_{f \in B} |Lf| \leq \sup_{f \in B} \int_{\mathbf{R}^d} \varphi(x) f(x) dx \leq \sup_{f \in B} \int_{\mathbf{R}^d} \varphi(x) \|f\| l(x,x) dx = \int_{\mathbf{R}^d} \varphi(x) l(x,x) dx < \infty$$

By Riesz representation Theorem there exist $\phi \in H$ such that $\langle \phi, f \rangle = \int_{\mathbf{R}^d} \varphi(x) f(x) dx$. By reproducing property ϕ is given by equation $\phi(x) = \langle \phi, l(t,) \rangle = \int_{\mathbf{R}^d} l(x,t) \varphi(x) dx$. With each probability measure P with an integrable characteristic function φ_P we associate the smooth characteristic function with

$$P \to \phi_P(x) = \int_{\mathbf{R}^d} l(x,t) \varphi_P(x) dx$$

We will prove that $P \to \phi_P$ is injective. We will show that , $\forall_x \phi_Q(x) = \phi_P(x)$ implies $P = Q$.

$$\phi_Q = \phi_P \Rightarrow \int_{\mathbf{R}^d} l(x-t) \varphi_P(x) dx = \int_{\mathbf{R}^d} l(x-t) \varphi_Q(x) dx. \tag{5.4}$$

We apply inverse Fourier transform to this convolution and get

$$g(x) f_X(x) = f_Y(x) g(x)$$

Where $Tg = l$, $Tf_Y = \varphi_Q$ and $Tf_X = \varphi_P$. Since inverse Fourier transform is injective on the space of continuous, integrable functions, and both ϕ_Q, ϕ_P are continuous and integrable, then application of the inverse Fourier transform does not enlarge the null space of Eq. (5.4). Since $g(x) > 0$, $f_X(x) = f_Y(x)$ everywhere, implying that the mapping $P \to \phi_P$ is injective.

□

Next, we show that smooth characteristic function is analytic.

Lemma 17. *If l is an analytic, integrable kernel with an inverse Fourier transform strictly greater then zero and P has an integrable characteristic function then the smooth characteristic function ϕ_P is analytic.*

Proof. By lemma 5, all functions in the RKHS associated with l are analytic, and by Lemma 16 ϕ_P is an element of this RKHS. □

We are ready to prove Theorem 7.

Proof of Theorem 7. Since l is an analytic, integrable kernel with an inverse Fourier transform strictly greater then zero then by the Lemma 16 the mapping $\Lambda : P \to \phi_P$ is injective and $\Lambda(P)$ is an element of the RKHS associated with l. Lemma 17 shows that μ_P is analytic. Therefore we can use Lemma 15 to see that $d_{\Lambda,J}(P,Q)^2 = d_{\phi,J}(P,Q)^2$ is a random metric. □

Proof of Lemma 5

Proof. By Fubini's theorem we get

$$\begin{aligned}
\phi_P(t) &= \int_{\mathbf{R}^d} \varphi_P(t-w) f(w) dw \\
&= \int_{\mathbf{R}^d} \left(\int_{\mathbf{R}^d} e^{i(t-w)^\top x} dP(x) \right) f(w) dw \\
&= \int_{\mathbf{R}^d} e^{it^\top x} \left(\int_{\mathbf{R}^d} e^{-iw^\top x} f(w) dw \right) dP(x) \\
&= E[e^{it^\top X} F f(X)].
\end{aligned}$$

Use of Fubini's theorem is justified, since the iterated integral is finite [86][Theorem 8.8 b] i.e.

$$\begin{aligned}
&\int_{\mathbf{R}^d} \int_{\mathbf{R}^d} |e^{i(t-w)^\top x} f(w)| dP(x) dw \\
&\quad = \int_{\mathbf{R}^d} |f(w)| \int_{\mathbf{R}^d} 1 dP(x) dw < \infty.
\end{aligned}$$

□

Proof of Proposition 4

Proof. The probability spaces of random variables $\{T_j\}_{1 \le j \le J}$ and $\{X_i\}_{1 \le i \le n}$ are $(\Omega_1, \mathcal{F}_1, P_1)$ and $(\Omega_2, \mathcal{F}_2, P_2)$, respectively. We will show that for almost all

$\omega \in \Omega_1$, S_n converges to χ^2 distribution with J degrees of freedom. We define

$$
\begin{aligned}
Z_i^\omega &= (k(X_i,T_1(\omega)) - k(Y_i,T_1(\omega)), \cdots, k(X_i,T_J(\omega)) - k(Y_i,T_J(\omega))) \in R^J, \\
W_n^\omega &= \frac{1}{n}\sum_{i=1}^{n} Z_i^\omega \\
\Sigma_n^\omega &= \frac{1}{n}\sum_{i} (Z_i^\omega - W_n^\omega)(Z_i^\omega - W_n^\omega)^T \\
S_n^\omega &= n W_n^\omega \Sigma_n^{-1} W_n^\omega.
\end{aligned}
$$

If there exists $a \neq b$, such that $T_a(\omega) = T_b(\omega)$, then we redefine $Z_i^\omega = N(0, Id_J)$.

Suppose $d^\omega_{\mu,J}(P,Q) = 0$. Then, by theorem 8, for all j, $\mu_P(T_j(\omega)) = \mu_Q(T_j(\omega)$. This implies that $EZ_i^\omega = 0$, which in turn implies, by [2][5.2.3], that S_n^ω is asymptotically χ^2 distributed with J degrees of freedom.

If $EZ_i^\omega \neq 0$ then

$$
P(S_n^\omega > r) = P\left((W_n^\omega)^\top (\Sigma_n^{-1})^\omega W_n^\omega - \frac{r}{n} > 0\right) \to 1.
$$

To see that, first we show that $(\Sigma_n^{-1})^\omega$ converges in probability to the positive definite matrix $(\Sigma^{-1})^\omega$. Indeed, each entry of the matrix Σ_n^ω converges to the matrix Σ^ω, hence entires of the matrix $(\Sigma^{-1})^\omega$, given by a continuous function of the entries of Σ^ω, are limit of the sequence $(\Sigma_n^{-1})^\omega$. Similarly W_n^ω converges in probability to the vector W^ω. Since $(W^\omega)^\top (\Sigma^{-1})^\omega W^\omega = a^\omega > 0$ ($(\Sigma^{-1})^\omega$ is positive definite), then $(W_n^\omega)^\top (\Sigma_n^{-1})^\omega W_n^\omega$, being a continuous function of the entries of W_n^ω and $(\Sigma_n^{-1})^\omega$, converges to a^ω. On the other hand $\frac{r}{n}$ converges to zero and the proposition follows. Finally since $d^\omega_{\mu,J}(P,Q) > 0$ almost surely then $EZ_i^\omega \neq 0$ for almost all $\omega \in \Omega_1$.

We have showed that the proposition holds for almost all ω, and thus S_n^ω converges for almost all ω. Indeed it does not hold if it happens that for some $a \neq b$, $T_a(\omega) = T_b(\omega)$ or $d^\omega_{\mu,J}(P,Q) = 0$ for $P \neq Q$. But both those events have zero measure.

□

Proof of Proposition 6 The poof is analogue to the proof of the Proposition 4.

Experiments

In this section we compare two-sample tests on both artificial benchmark data and on real-world data. We denote the smooth characteristic function test as 'Smooth CF', and the test based on the analytic mean embeddings as 'Mean Embedding'. We compare against several alternative testing approaches: block MMD ('Block MMD'), a characteristic functions based test ('CF'), a sub-sampling MMD test ('MMD($\sqrt{n}$)'), and the quadratic-time MMD test ('MMD(n)'). We discuss Block MMD below.

Block MMD

An alternative to the quadratic-time MMD test is a B-test (block-based test): the idea is to break the data into blocks, compute a quadratic-time statistic on each block, and average these quantities to obtain the test statistic. More specifically, for an individual block, laying on the main diagonal and starting at position $(i-1)B+1$, the statistic $\eta(i)$ is calculated as

$$\eta(i) = \frac{1}{\binom{B}{2}} \sum_{a=(i-1)B+1}^{iB} \sum_{b=(i-1)B+1 \neq a}^{iB} h(X_a, X_b, Y_a, Y_b).$$

The overall test statistic is then

$$\eta = \frac{B}{n} \sum_{i=1}^{\frac{n}{B}} \eta(i). \tag{5.5}$$

The choice of B determines computation time - at one extreme is the linear-time MMD suggested by [45, 48] where we have $n/2$ blocks of size $B = 2$, and at the other extreme is the usual full MMD with 1 block of size n, which requires calculating the test statistic on the whole kernel matrix in quadratic time. In our case, the size of the block remains constant as n increases, and is greater than 2. This is very similar to the case proposed by [114], and the consistency of the test is not affected.

B-test of [114] assumes that $B \to \infty$ together with n, which implies that the statistic $\hat{\eta}$ defined in (5.5) under the null distribution satisfies

$$\sqrt{nB}\hat{\eta} \xrightarrow{D} \mathcal{N}\left(0, 4\sigma_0^2\right), \tag{5.6}$$

for asymptotic variance $\sigma_0^2 = \mathbb{E}_{XX'} k^2(X,X') + (\mathbb{E}_{XX'} k(X,X'))^2 - 2\mathbb{E}_X \left[(\mathbb{E}_{X'} k(X,X'))^2 \right]$ that can easily be estimated directly or by considering the empirical variance of the statistics computed within each of the blocks. Note that the same asymptotic variance σ_0^2 is obtained in the case of a quadratic-time statistic [45] – albeit convergence rate being a faster $O(1/n)$ in that case. Indeed, (5.6) is obtained directly from the leading term of the variance of each block-based statistic being $\frac{4\sigma_0^2}{B^2}$. Therefore, the p-value for B-test is approximated as $\Phi\left(-\frac{\sqrt{nB}\hat{\eta}}{2\hat{\sigma}_0}\right)$, where Φ is the standard normal cdf. When B remains constant as n increases, it can be shown that the variance of each block-based statistic is exactly $\frac{4\sigma_0^2}{B(B-1)}$, and thus we obtain by CLT that

$$\sqrt{n}\hat{\eta} \xrightarrow{D} \mathcal{N}\left(0, \frac{4\sigma_0^2}{B-1}\right).$$

Therefore, a slight change to p-value needs to be applied when σ_0^2 is estimated directly: $\Phi\left(-\frac{\sqrt{n(B-1)}\hat{\eta}}{2\hat{\sigma}_0}\right)$. If, however, one simply uses the empirical variance of the individual statistics computed within each block, the procedure is unaffected.

Experimental setup. For all the experiments, D is the dimensionality of samples in a dataset, n is a number of samples in the dataset (sample size) and J is number of test frequencies. Parameter selection is required for all the tests. The table summarizes the main choices of the parameters made for the experiments. The first parameter is the test function, used to calculate the particular statistic. The scalar γ represents the length-scale of the observed data. Notice that for the kernel tests we recover the standard parameterization $\exp(-\|\frac{x}{\gamma} - \frac{y}{\gamma}\|^2) = \exp(-\frac{\|x-y\|^2}{\gamma^2})$. The original CF test was proposed without any parameters, hence we added γ to ensure a fair comparison - for this test varying γ is equivalent to adjusting the variance of the distribution of frequencies T_j. For all tests, the value of the scaling parameter γ was chosen so as to minimize a p-value estimate on a held-out training set: details are described in Appendix 5.5. We chose not to optimize the sampling scheme for the Mean Embedding and Smooth CF tests, since this would give them an unfair advantage over the Block MMD, MMD($\sqrt{n}$) and CF tests. The block size in the Block MMD test and the number of test frequencies in the Mean Embedding, Smooth CF, and CF tests, were always set to the same value (not greater than 10) to maintain exactly the same time complexity. Note that we did *not* use the popular median heuristic for kernel bandwidth choice (MMD and B-test), since it gives

poor results for the Blobs and AM Audio datasets [48]. We do not run MMD(n) test for 'Simulation 1' or 'Amplitude Modulated Music', since the sample size is 10000, and too large for a quadratic-time test with permutation sampling for the test critical value.

It is important to verify that Type I error is indeed at the design level, set at $\alpha = 0.05$ in this paper. This is verified in the Figure 5.2. Also shown in the plots is the 95% percent confidence intervals for the results, as averaged over 4000 runs.

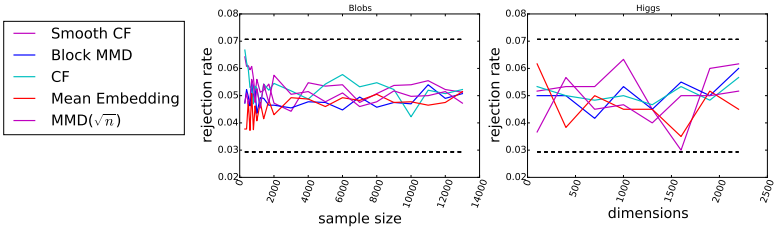

Figure 5.2: Type I error of the blobs dataset **(left)** and the dimensions dataset **(right)**. The dashed line is the 99% Wald interval $\alpha \pm 2.57\sqrt{\alpha(1-\alpha)/4000}$ (4000 is number of repetitions) around the design test size of $\alpha = 0.05$.

Test	Test Function	Parameters
Mean Embedding	$\exp(-\|\gamma^{-1}(x-t)\|^2)$	$T_j \sim N(0_D, I_D)$ J - no. of test frequencies
Smooth CF	$\exp(it^\top\gamma^{-1}x - \|\gamma^{-1}x - t\|^2)$	$T_j \sim N(0_D, I_D)$ J - no. of test frequencies
MMD(n),MMD($\sqrt{n}$)	$\exp(-\|\gamma^{-1}(x-t)\|^2)$	b -bootstraps
Block MMD	$\exp(-\|\gamma^{-1}(x-t)\|^2)$	B-block size
CF	$\exp(it^\top\gamma^{-1}x)$	$T_j \sim N(0_D, I_D)$ J - no. of test frequencies

Real Data 1: Higgs dataset, $D = 4$, *n varies,* $J = 10$. The first experiment we consider is on the UCI Higgs dataset [68] described in [5] - the task is to distinguish signatures of processes that produce Higgs bosons from background processes that do not. We consider a two-sample test on certain extremely low-level features in the dataset - kinematic properties measured by the particle detectors, i.e., the joint distributions of the azimuthal angular momenta φ for four particle jets. We denote by P the jet φ-momenta distribution of the background process (no Higgs bosons), and by Q the corresponding distribution for the process that produces Higgs bosons (both are distributions on $\mathbf{R}^4$). As discussed in [5, Fig. 2], φ-momenta, unlike transverse momenta p_T, carry very

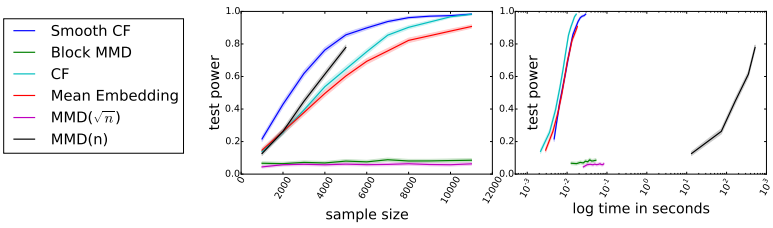

Figure 5.3: Higgs dataset. **Left:** Test power vs. sample size. **Right:** Test power vs. execution time.

little discriminating information for recognizing whether Higgs bosons were produced. Therefore, we would like to test the null hypothesis that the distributions of angular momenta P (no Higgs boson observed) and Q (Higgs boson observed) might yet be rejected. The results for different algorithms are presented in the Figure 5.3. We observe that the joint distribution of the angular momenta is in fact discriminative. Sample size varies from 1000 to 12000. The Smooth CF test has significantly higher power than the other tests, including the quadratic-time MMD, which we could only run on up to 5100 samples due to computational limitations. The leading performance of the Smooth CF test is especially remarkable given it is several orders of magnitude faster than the quadratic-time MMD(n), even though we used the fastest quadratic-time MMD implementation, where the asymptotic distribution is approximated by a Gamma density .

Real Data 2: Amplitude Modulated Music, $D = 1000$*,* $n = 10000$*,* $J = 10$*.* Amplitude modulation is the earliest technique used to transmit voice over the radio. In the following experiment observations were one thousand dimensional samples of carrier signals that were modulated with two different input audio signals from the same album, song P and song Q (further details of these data are described in [48, Section 5]). To increase the difficulty of the testing problem, independent Gaussian noise of increasing variance (in the range 1 to 4.0) was added to the signals. The results are presented in the Figure 5.4. Compared to the other tests, the Mean Embedding and Smooth CF tests are more robust to the moderate noise contamination.

Simulation 1: High Dimensions, D *varies,* $n = 10000$*,* $J = 3$*.* It has recently been shown, in theory and in practice, that the two-sample problem gets more difficult for an increasing number of dimensions increases on which the distributions do not differ [80, 81]. In the following experiment, we study the

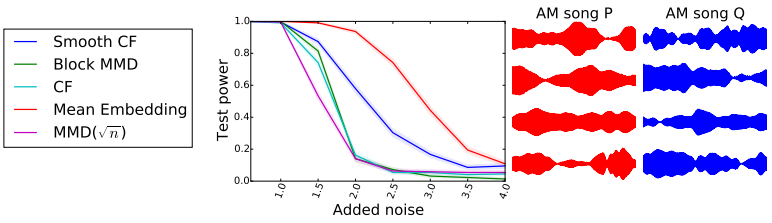

Figure 5.4: Music Dataset.**Left:** Test power vs. added noise. **Right:** four samples from P and Q.

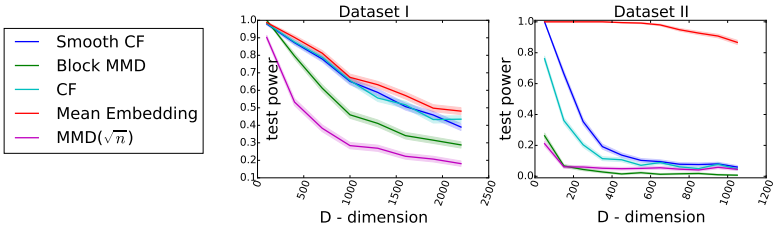

Figure 5.5: Power vs. redundant dimensions comparison for tests on high dimensional data.

power of the two-sample tests as a function of dimension of the samples. We run two-sample tests on two datasets of Gaussian random vectors which differ *only* in the first dimension,

$$\begin{aligned} \text{Dataset I:} \quad & P = N(0_D, I_D) \quad vs. \quad Q = N((1,0,\cdots,0), I_D) \\ \text{Dataset II:} \quad & P = N(0_D, I_D) \quad vs. \quad Q = N(0_D, \text{diag}((2,1,\cdots,1))), \end{aligned}$$

where 0_d is a D-dimensional vector of zeros, I_D is a D-dimensional identity matrix, and $\text{diag}(v)$ is a diagonal matrix with v on the diagonal. The number of dimensions (D) varies from 50 to 2500 (Dataset I) and from 50 to 1200 (Dataset II). The power of the different two-sample tests is presented in Figure 5.5. The Mean Embedding test yields best performance for both datasets, where the advantage is especially large for differences in variance.

Simulation 2: Blobs, $D = 2$*,* n *varies,* $J = 5$*.* The Blobs dataset is a grid of two dimensional Gaussian distributions (see Figure 5.6), which is known to be a challenging two-sample testing task. The difficulty arises from the fact that the difference in distributions is encoded at a much smaller length-scale than the overall data. In this experiment both P and Q are four by four grids of Gaussians, where P has unit covariance matrix in each mixture component, while each component of Q has direction of the largest variance ro-

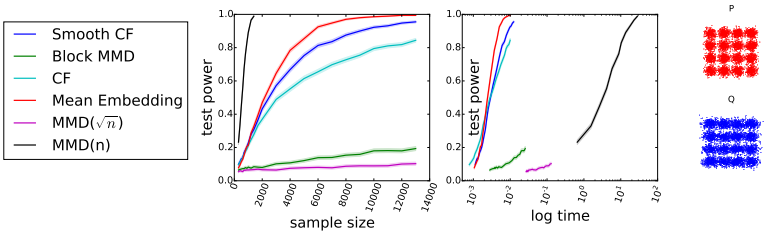

Figure 5.6: Blobs Dataset. **Left:** test power vs. sample size. **Center:** test power vs. execution time. **Right:** illustration of the blob dataset.

tated by $\pi/4$ and amplified to 4. It was demonstrated by [48] that a good choice of kernel is crucial for this task. Figure 5.6 presents the results of two-sample tests on the Blobs dataset. The number of samples varies from 50 to 14000 (MMD(n) reached test power one with $n = 1400$). We found that the MMD(n) test has the best power as function of the sample size, but the worst power/computation tradeoff. By contrast, random distance based tests have the best power/computation tradeoff.

Parameters Choice

We split our data into disjoint training and testing sets, and optimized parameters on the training set. To evaluate different data scalings λ, we plotted the associated p-values of tests on the training data. Figure 5.7 presents such a plot for three different tests. The p-values were obtained by running the test several times (20 to 50) for each λ. In the case of simulated data, we generated a new training dataset for each repetition at a given λ. For the amplitude modulated audio dataset, we added different independent noise to the training samples for each repetition (note that this was in retrospect not an ideal choice: a better approach would have been to draw bootstrap samples from the training data, possibly using additional tracks from the CD to provide sufficient training samples). For the Higgs dataset, we had an abundance of training data, hence we were able to use bootstrap samples without repetition from the training set. This last approach is the recommended strategy for real-life data. We emphasize that p-value optimization is a successful heuristic, but is *not* a substitute for a choice of parameters that optimizes test power. Better parameter choice might be accomplished following a strategy analogous to [48].

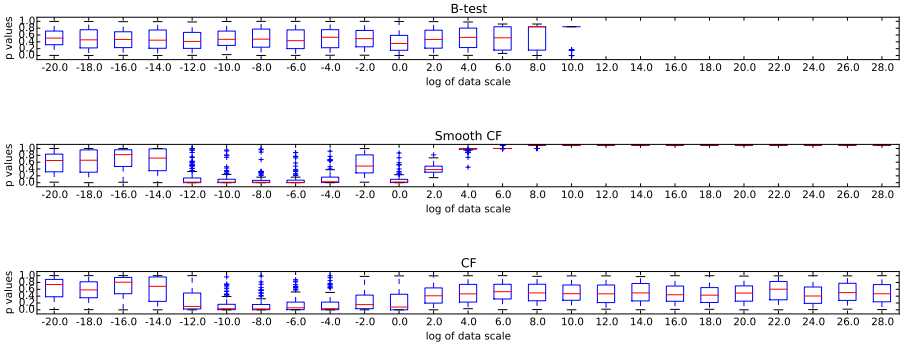

Figure 5.7: Box plot of p-values used for parameter selection. The X axis shows the binary logarithm of the scaling parameter applied to data. We chose the scaling with the smallest median p-value. If the medians were similar we used a scaling that had few outliers and was surrounded by other scalings with small p-values. In this example we chose scalings $\lambda = 2^0 = 1$ for the B-test, $\lambda = 2^{-8}$ for the Smoothed CF test, and $\lambda = 2^{-10}$ for the CF test.

Acknowledgement. We would like thank Bharath Sriperumbudur and Wittawat Jitkrittum for insightful comments.

Chapter 6

Conclusions and Future Work

In this thesis we discuss some topics in nonparametric, kernel based, statistical hypothesis testing. We showed that the wild bootstrap is a valid bootstrap method for degenerate kernel tests. We applied wild bootstrap method to construct Maximum Mean Discrepancy two-sample test and Hilbert Schmidt Independence Criterion test for time series. We constructed a kernel goodness of fit test that requires knowledge of density only up to a normalizing constant, which is a common setting in machine learning. We used the wild bootstrap method to adopt the kernel goodness of fit test for MCMC convergence diagnostics. Finally, we constructed a linear time, almost surely consistent, kernel two sample test using properties of homomorphic, injective mean embeddings and smoothed characteristic functions.

Existing Extensions

Some results presented in this thesis were used by other researchers in their work. *Wild Bootstrap for Degenerate Kernel Tests.* Methods developed in Chapter 3 were used by Rubenstein et al. [85] to adopt Lancaster three-variable interaction test by Sejdinovic et al. [87] for time-series. The test is based on a Lancaster interaction measure, which kernel mean embedding is equal to zero if the measure of the three random variables factorizes in any way. This property makes Lancaster test useful is discovering so called V-structures or triples of random variables that are totally independent. The analysis of a degenerate V-statistic presented in that work is similar to the one that we present for HSIC test in Section 3.3 (since the core is a function of 10 arguments and a naive implementation would results in time complexity $O(n^{10})$).

Fast Analytic Functions Based Two Sample Test. Jitkrittum et al. [61] propose

an objective function for optimizing test locations, and show that maximizing this objective corresponds to maximizing the test power of the test that we have proposed in Chapter 5. They derive a finite-sample probabilistic bound guaranteeing the convergence of the objective to the corresponding population quantity. Benchmarks confirm that the approach significantly increases the test power, while keeping the false positive rate to the specified significance level.

Extensions and Future work.

Reduction for HSIC. It might be possible to reduce the problem of the HSIC degenerate V-statistic convergence to a simple application of CLT for real valued random variables.

In Section 3.3 we represent HSIC test statistic as a squared norm of normalized sum of features in RKHS, recall definition of T_n from Equation 3.7

$$T_n = \frac{1}{\sqrt{n}} \sum_{i \in N} W_i (\phi(X_i) - \mu_X) \otimes (\phi(Y_i) - \mu_Y).$$

and further define

$$H_n = \frac{1}{\sqrt{n}} \sum_{i \in N} (\phi(X_i) - \mu_X) \otimes (\phi(Y_i) - \mu_Y).$$

We showed in Lemma 9 that the squared norm of T_n is equal to bootstraped V-statistic of h_2, $B_n(h_2)$. Similarly, squared norm of H_n is equal to $V_n(h_2)$. Under the null hypothesis

$$E\langle f \otimes g, T_n \rangle = E\langle f \otimes g, H_n \rangle = 0.$$

where $\mathcal{H}$ RKHS associated with the kernel given by the feature map $\phi \otimes \phi$. This means that both T_n and H_n must converge weakly to the same distribution, provided they converge at all (it is sufficient to check expected values along linear bounded maps in RKHS). Using Continuous Mapping Theorem $\|T_n\|^2$ and $\|H_n\|^2$ must converge to the same distribution. Sufficient conditions for H_n and T_n to converge are that

$$\forall f \otimes g \in \mathcal{H} \quad \frac{1}{\sqrt{n}} \sum_{i \in N} (f(X_i) - Ef(X_i))(g(X_i) - Eg(X_i))$$

converges weakly and tightness (For each $\epsilon > 0$, there exists a compact set $K \subset \mathcal{H}$ such that $P(T_n \in K) \geq 1 - \epsilon$). Note that the rest of the reasoning form Section 3.3 must be accordingly adjusted.

A Kernel Test of Goodness of Fit. It would be interesting to use Stein based discrepancy to generate points from the stationary distribution. Currently, we use Stein discrepancy to verify if a set of points is drawn form the target distribution. It seems that if points are not drawn from the target distribution, moving them around as to minimize the Stein divergence, can produce a set of points that resembles points drawn form the target distribution. This could be implemented using an algorithm similar to a particle filter. Another approach would be to use kernel herding. The feature map associated with goodness of fit test (4.1) is

$$\xi_p(x, \cdot) := \left[\nabla \log p(x) k(x, \cdot) + \nabla k(x, \cdot)\right],$$

The kernel herding algorithm is in this case very simple

$$\begin{aligned} x_{t+1} &= argmax_{x \in X} \langle \xi_p(x, \cdot), w_t \rangle \\ w_{t+1} &= w_t - \xi(x_{t+1}), \end{aligned}$$

since $E_{X \sim p} \xi_p(X) = 0$!

Linear Time Kernel Test of Goodness of Fit. It would be interesting to modify the kernel goodness of fit test in spirit of Chapter 5. For a fixed location y and a random variable X, we define $s(X, y)$ to be

$$s_p(X, y) = \nabla \log p(X) g(X, y) - \nabla g(X, y).$$

For a number of random locations $Y_1, \cdots, Y_J$ and a dataset $\{X_i\}_{1 \leq i \leq n}$ we define random vector Z_i

$$Z_i = (s_p(X_i, Y_1), \cdots, s_p(X_i, Y_J)).$$

Let W_n be a mean of Z_i's, $W_n = \frac{1}{n} \sum_{i=1}^{n} Z_i$, and Σ_n its covariance matrix $\Sigma_n = \frac{1}{n} Z Z^T$. Consider the test statistic

$$S_n = n W_n \Sigma_n^{-1} W_n.$$

The computation of S_n requires inversion of a $J \times J$ matrix Σ_n, but this is fast and numerically stable for small J. It is easy to see that if $X_i \sim p$ (under the null hypothesis) S_n is asymptotically distributed as ξ-squared random variable with J degrees of freedom. In this way one could even go outside the class of RKHS functions and use e.g. neural networks to construct test functions. Those functions can be also optimized similarly to what was done in [61].

Bibliography

[1] N. Anderson, P. Hall, and D. 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.

[2] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, July 2003. ISBN 978-0-471-36091-9.

[3] M. Arcones. The law of large numbers for u-statistics under absolute regularity. *Electron. Comm. Probab*, 3:13–19, 1998.

[4] M.A. Arcones. The law of large numbers for U-statistics under absolute regularity. *Electron. Comm. Probab*, 3:13–19, 1998.

[5] P. Baldi, P. Sadowski, and D. Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5, 2014.

[6] R. Bardenet, A. Doucet, and C. Holmes. Towards scaling up Markov Chain Monte Carlo: an adaptive subsampling approach. In *ICML*, pages 405–413, 2014.

[7] L Baringhaus and C Franz. On a new multivariate two-sample test. 88 (1):190–206, 2004.

[8] L. Baringhaus and N. Henze. A consistent test for multivariate normality based on the empirical characteristic function. *Metrika*, 35:339–348, 1988.

[9] A. R. Barron. Uniformly powerful goodness of fit tests. *The Annals of Statistics*, 17:107–124, 1989.

[10] L. Bauwens, S. Laurent, and J.V.K. Rombouts. Multivariate GARCH models: a survey. *J. Appl. Econ.*, 21(1):79–109, January 2006. ISSN 1099-1255. doi: 10.1002/jae.842. URL `http://onlinelibrary.wiley.com/doi/10.1002/jae.842/abstract`.

[11] J. Beirlant, L. Györfi, and G. Lugosi. On the asymptotic normality of the l_1- and l_2-errors in histogram density estimation. *Canadian Journal of Statistics*, 22:309–318, 1994.

[12] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

[13] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*, volume 3. Kluwer Academic Boston, 2004.

[14] M. Besserve, N. Logothetis, and B. Schölkopf. Statistical analysis of coupled time series with kernel cross-spectral density operators. In *NIPS*, pages 2535–2543, 2013.

[15] I. Borisov and N. Volodko. Orthogonal series and limit theorems for canonical U- and V-statistics of stationary connected observations. *Siberian Advances in Mathematics*, 18(4):242–257, 2008.

[16] A.W. Bowman and P.J. Foster. Adaptive smoothing and density based tests of multivariate normality. *J. Amer. Statist. Assoc*, 88:529–537, 1993.

[17] R. Bradley et al. Basic properties of strong mixing conditions. a survey and some open questions. *Probability surveys*, 2(107-44):37, 2005.

[18] Richard C Bradley. On the central limit question under absolute regularity. *The Annals of Probability*, pages 1314–1325, 1985.

[19] W. Broock, J. Scheinkman, D. Dechert, and B. LeBaron. A test for independence based on the correlation dimension. *Econometric Reviews*, 15 (3):197–235, 1996.

[20] Claudio Carmeli, Ernesto De Vito, Alessandro Toigo, and Veronica Umanitá. Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

[21] Kacper Chwialkowski and Arthur Gretton. A kernel independence test for random processes. In *ICML*, 2014.

[22] Kacper Chwialkowski, Dino Sejdinovic, and Arthur Gretton. A wild bootstrap for degenerate kernel tests. In *Advances in Neural Information Processing Systems 27*, pages 3608–3616. Curran Associates, Inc., 2014.

[23] Kacper Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, pages 1972–1980, 2015.

[24] Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *ICML*, 2016.

[25] K. R. Davidson. Pointwise limits of analytic functions. *Am math mon*, pages 391–394, 1983.

[26] J. Dedecker, P. Doukhan, G. Lang, S. Louhichi, and C. Prieur. *Weak dependence: with examples and applications*, volume 190. Springer, 2007.

[27] Jérôme Dedecker and Clémentine Prieur. New dependence coefficients. examples and applications to statistics. *Probability Theory and Related Fields*, 132(2):203–236, 2005.

[28] M. Denker and G. Keller. On U-statistics and v. Mises statistics for weakly dependent processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 64(4):505–522, 1983.

[29] C. Diks and V. Panchenko. Nonparametric tests for serial independence based on quadratic forms. Technical report, Tinbergen Institute Discussion Paper, 2005.

[30] P. Doukhan. Mixing. properties and examples. In *Mixing*, number 85 in Lect. Notes in Stat., pages 87–109. Springer, January 1994. ISBN 978-0-387-94214-8, 978-1-4612-2642-0. URL `http://link.springer.com/chapter/10.1007/978-1-4612-2642-0_9`.

[31] T.W. Epps and K.J. Singleton. An omnibus test for the two-sample problem using the empirical characteristic function. *Journal of Statistical Computation and Simulation.*, 26(3-4):177–203, 1986.

[32] B. Fadlallah, A. Brockmeier, S. Seth, L. Li, A. Keil, and J. Principe. An association framework to analyze dependence structure in time series. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 6176–6179. IEEE, 2012.

[33] Yanqin Fan and Aman Ullah. On goodness-of-fit tests for weakly dependent processes using kernel method. *Journal of Nonparametric Statistics*, 11(1-3):337–360, 1999.

[34] A. Feuerverger. A consistent test for bivariate dependence. *International Statistical Review/Revue Internationale de Statistique*, pages 419–433, 1993.

[35] Magalie Fromont and Batrice Laurent. Adaptive goodness-of-fit tests in a density model. *Ann. Statist.*, 34(2):680–720, 04 2006. doi: 10.1214/009053606000000119. URL `http://dx.doi.org/10.1214/009053606000000119`.

[36] Magalie Fromont, Batrice Laurent, and Patricia Reynaud-Bouret. The two-sample problem for poisson processes: Adaptive tests with a nonasymptotic wild bootstrap approach. *Ann. Statist.*, 41(3):1431–1461, 06 2013. doi: 10.1214/13-AOS1114. URL `http://dx.doi.org/10.1214/13-AOS1114`.

[37] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems*, volume 20, pages 489–496, 2007.

[38] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *NIPS 20*, pages 489–496, 2008.

[39] A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472, 1992.

[40] J. Gorham and L. Mackey. Measuring sample quality with stein's method. In *NIPS*, pages 226–234, 2015.

[41] C. Granger, E. Maasoumi, and J. Racine. A dependence metric for possibly nonlinear processes. *Journal of Time Series Analysis*, 25(5):649–669, 2004.

[42] A. Gretton and L. Gyorfi. Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, 11:1391–1423, 2010.

[43] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic learning theory*, pages 63–77. Springer, 2005.

[44] A. Gretton, K. Fukumizu, C Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In *NIPS*, volume 20, pages 585–592, 2007.

[45] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. 13:723–773, 2012.

[46] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *JMLR*, 13:723–773, 2012.

[47] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal Machine Learning Research*, 13:723–773, 2012.

[48] A. Gretton, B. Sriperumbudur, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, and K. Fukumizu. Optimal kernel choice for large-scale two-sample tests. 2012.

[49] Arthur Gretton, Kenji Fukumizu, Choon H Teo, Le Song, Bernhard Schölkopf, and Alex J Smola. A kernel statistical test of independence. In *Advances in neural information processing systems*, pages 585–592, 2007.

[50] L. Györfi and I. Vajda. Asymptotic distributions for goodness of fit statistics in a sequence of multinomial models. *Statistics and Probability Letters*, 56:57–67, 2002.

[51] L. Györfi and E. C. van der Meulen. A consistent goodness of fit test based on the total variation distance. In G. Roussas, editor, *Nonparametric Functional Estimation and Related Topics*, pages 631–645. Kluwer, Dordrecht, 1990.

[52] Z. Harchaoui, F. Bach, and E. Moulines. Testing for homogeneity with kernel Fisher discriminant analysis. In *Advances in Neural Information Processing Systems*. 2008.

[53] L. Haugh. Checking the independence of two covariance-stationary time series: a univariate residual cross-correlation approach. *Journal of the American Statistical Association*, 71(354):378–385, 1976.

[54] CE Heathcote. A test of goodness of fit for symmetric random variables. *Aust J stat*, 14(2):172–181, 1972.

[55] CR Heathcote. The integrated squared error estimation of parameters. *Biometrika*, 64(2):255–264, 1977.

[56] H.-C. Ho and G. Shieh. Two-stage U-statistics for hypothesis testing. *Scandinavian Journal of Statistics*, 33(4):861–873, 2006.

[57] Y. Hong. Testing for independence between two covariance stationary time series. *Biometrika*, 83(3):615–625, 1996.

[58] Y. Hong and H. White. Asymptotic distribution theory for nonparametric entropy measures of serial dependence. *Econometrica*, 73(3):837–901, 2005.

[59] H. Hotelling. The generalization of student's ratio. *Ann. Math. Statist.*, 2(3):360–378, 1931. doi: 10.1214/aoms/1177732979. URL `http://dx.doi.org/10.1214/aoms/1177732979`.

[60] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

[61] Wittawat Jitkrittum, Zoltan Szabo, and Arthur Gretton. A fast interpretable nonparametric two-sample test. *arXiv*, 2016.

[62] A. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *G. Ist. Ital. Attuari*, 4:83–91, 1933.

[63] A. Korattikara, Y. Chen, and M. Welling. Austerity in MCMC Land: Cutting the Metropolis-Hastings Budget. In *ICML*, pages 181–189, 2014.

[64] Anoop Korattikara, Yutian Chen, and Max Welling. Austerity in mcmc land: Cutting the metropolis-hastings budget. *arXiv preprint arXiv:1304.5299*, 2013.

[65] Q. Le, T. Sarlos, and A. Smola. Fastfood - computing Hilbert space expansions in loglinear time. In *ICML*, volume 28, pages 244–252, 2013.

[66] A. Leucht. Degenerate U- and V-statistics under weak dependence: Asymptotic theory and bootstrap consistency. *Bernoulli*, 18(2):552–585, 2012.

[67] A. Leucht and M.H. Neumann. Dependent wild bootstrap for degenerate U- and V-statistics. *Journal of Multivariate Analysis*, 117:257–280, 2013. ISSN 0047-259X. doi: 10.1016/j.jmva.2013.03.003. URL `http://www.sciencedirect.com/science/article/pii/S0047259X13000304`.

[68] M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

[69] Q. Liu, J. Lee, and M. I. Jordan. A kernelized stein discrepancy for goodness-of-fit tests and model evaluation. Technical report, ArXiv, 2016.

[70] James R Lloyd and Zoubin Ghahramani. Statistical model criticism using kernel two sample tests. In *NIPS*, pages 829–837, 2015.

[71] R. Lyons. Distance covariance in metric spaces. *The Annals of Probability*, 41(5):3051–3696, 2013.

[72] Abdelkader Mokkadem. Mixing properties of arma processes. *Stochastic processes and their applications*, 29(2):309–315, 1988.

[73] A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[74] C. Oates, M. Girolami, and N. Chopin. Control functionals for monte carlo integration. Technical Report arXiv:1410.2392v4, ArXiv, 2015.

[75] J. Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press, 2000.

[76] J. Pickands III. Statistical inference using extreme order statistics. *Annals of Statistics*, pages 119–131, 1975.

[77] J. Pinkse. A consistent nonparametric test for serial independence. *Journal of Econometrics*, 84(2):205–231, 1998.

[78] J. Racine and E. Maasoumi. A versatile and robust metric entropy test of time-reversibility, and other hypotheses. *Journal of Econometrics*, 138 (2):547–567, 2007.

[79] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *NIPS*, pages 1177–1184, 2007.

[80] A. Ramdas, S. Reddi, B. Póczos, A. Singh, and L. Wasserman. On the decreasing power of kernel- and distance-based nonparametric hypothesis tests in high dimensions. *AAAI*, 2015.

[81] S. Reddi, A. Ramdas, B. Póczos, A. Singh, and L. Wasserman. On the high-dimensional power of linear-time kernel two-sample testing under mean-difference alternatives. *AISTATS*, 2015.

[82] M. L. Rizzo. New goodness-of-fit tests for pareto distributions. *ASTIN Bulletin: Journal of the International Association of Actuaries*, 39(2):691–715, 2009.

[83] P. Robinson. Consistent nonparametric entropy-based testing. *The Review of Economic Studies*, 58(3):437–453, 1991.

[84] M. Rosenblatt and B. Wahlen. A nonparametric measure of independence under a hypothesis of independent components. *Statistics & probability letters*, 15(3):245–252, 1992.

[85] Paul K Rubenstein, Kacper P Chwialkowski, and Arthur Gretton. A kernel test for three-variable interactions with random processes. *arXiv preprint arXiv:1603.00929*, 2016.

[86] Walter Rudin. *Real and complex analysis*. Tata McGraw-Hill Education, 1987.

[87] D. Sejdinovic, A. Gretton, and W. Bergsma. A kernel test for three-variable interactions. In *Advances in Neural Information Processing Systems*, pages 1124–1132, 2013.

[88] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, 41(5):2263–2291, 2013.

[89] D. Sejdinovic, H. Strathmann, M. Lomeli Garcia, C. Andrieu, and A. Gretton. Kernel Adaptive Metropolis-Hastings. In *ICML*, 2014.

[90] R. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 1980.

[91] R. Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2002.

[92] X. Shao. A generalized portmanteau test for independence between two stationary time series. *Econometric Theory*, 25(01):195–210, 2009.

[93] X. Shao. The dependent wild bootstrap. *J. Amer. Statist. Assoc.*, 105 (489):218–235, 2010.

[94] N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, 19:279–281, 1948.

[95] A. J Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory*, volume LNAI4754, pages 13–31, Berlin/Heidelberg, 2007. Springer-Verlag.

[96] L Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization. *JMLR*, 13:1393–1434, 2012.

[97] B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Hilbert space embeddings and metrics on probability measures. *JMLR*, 11:1517–1561, 2010.

[98] B. Sriperumbudur, K. Fukumizu, and G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *JMLR*, 12:2389–2410, 2011.

[99] B. Sriperumbudur, K. Fukumizu, and G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. 12:2389–2410, 2011.

[100] B. Sriperumbudur, K. Fukumizu, R. Kumar, A. Gretton, and A. Hyvärinen. Density Estimation in Infinite Dimensional Exponential Families. *arXiv preprint arXiv:1312.3516*, 2014.

[101] Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, pages 583–602, Berkeley, Calif., 1972. University of California Press. URL `http://projecteuclid.org/euclid.bsmsp/1200514239`.

[102] I. Steinwart and A. Christmann. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008. ISBN 978-0-387-77241-7.

[103] I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[104] I. Steinwart and C. Scovel. Mercers theorem on general domains: on the interaction between measures, kernels, and rkhss. *Constructive Approximation*, 35(3):363–417, 2012.

[105] I. Steinwart, D. Hush, and C. Scovel. An explicit description of the reproducing kernel hilbert spaces of gaussian rbf kernels. *Information Theory, IEEE Transactions on*, 52(10):4635–4643, 2006.

[106] H. Strathmann, D. Sejdinovic, S. Livingstone, Z. Szabo, and A. Gretton. Gradient-free Hamiltonian Monte Carlo with Efficient Kernel Exponential Families. In *NIPS*, 2015.

[107] M. Sugiyama, T. Suzuki, Y. Itoh, T. Kanamori, and M. Kimura. Least-squares two-sample test. *Neural Networks*, 24(7):735–751, 2011.

[108] Hong-Wei Sun and Ding-Xuan Zhou. Reproducing kernel hilbert spaces associated with analytic translation-invariant mercer kernels. *Journal of Fourier Analysis and Applications*, 14(1):89–101, 2008.

[109] G. J. Székely and M. L. Rizzo. A new test for multivariate normality. *J. Multivariate Analysis*, 93(1):58–80, 2005.

[110] G. J Székely, M. L Rizzo, et al. Brownian distance covariance. *The annals of applied statistics*, 3(4):1236–1265, 2009.

[111] GJ Székely. E-statistics: The energy of statistical samples. Technical report, 2003.

[112] M. Welling and Y.W. Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *ICML*, pages 681–688, 2011.

[113] K. Yoshihara. Limiting behavior of u-statistics for stationary, absolutely regular processes. *Probability Theory and Related Fields*, 35(3):237–252, 1976.

[114] W. Zaremba, A. Gretton, and M. Blaschko. B-test: A non-parametric, low variance kernel two-sample test. 2013.

[115] K. Zhang, J. Peters, D. Janzing, B., and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *UAI*, pages 804–813, 2011.

[116] X. Zhang, L. Song, A. Gretton, and A. Smola. Kernel measures of independence for non-iid data. In *NIPS*, volume 22, 2008.

[117] Ji Zhao and Deyu Meng. FastMMD: Ensemble of circular discrepancy for efficient two-sample test. *Neural computation*, (27):1345–1372, 2015.

[118] AA Zinger, AV Kakosyan, and LB Klebanov. A characterization of distributions by mean values of statistics and certain probabilistic metrics. *Journal of Mathematical Sciences*, 59(4):914–920, 1992.

Appendix A

Preliminary article on HSIC for time series

Introduction

Measures of statistical dependence between pairs of random variables (X, Y) are well established, and have been applied in a wide variety of areas, including fitting causal networks [75], discovering features which have significant dependence on a label set [96], and independent component analysis [60]. Where pairs of observations are independent and identically distributed, a number of non-parametric tests of independence have been developed [34, 44, 110, 42], which determine whether the dependence measure value is statistically significant. These non-parametric tests are consistent against any fixed alternative - they make no assumptions as to the nature of the dependence.

For many data analysis tasks, however, the observations being tested are drawn from a time series: each observation is dependent on its past values. Examples include audio signals, financial data, and brain activity. Given two such random processes, we propose a hypothesis test of instantaneous dependence, of whether the two signals are dependent at a particular time t. Our test satisfies two important properties: it is consistent against any fixed alternatives, and it is non-parametric - we do not assume the dependence takes a particular form (such as linear correlation), nor do we require parametric models of the time series. We further avoid making use of a density estimate as an intermediate step, so as to avoid the assumption that the distributions have densities (for instance, when dealing with text or other structured data).

We use as our test statistic the Hilbert-Schmidt Independence Criterion (HSIC) [43, 44], which can be interpreted as the distance between embeddings of the joint distribution and the product of the marginals in a reproducing kernel

Hilbert space (RKHS) [46, Section 7]. When characteristic RKHSs are used, the HSIC is zero iff the variables are independent [97]. Under the null hypothesis of independence, $P_{XY} = P_X P_Y$, the minimum variance estimate of HSIC is a degenerate U-statistic. The distribution of the empirical HSIC under the null is an infinite sum of independent χ^2 variables [44], which follows directly from e.g. [91, Ch. 5]. In practice, given a sample $(x_i, y_i)_{i=1}^n$ of pairs of variables drawn from P_{XY}, the null distribution is approximated by a bootstrap procedure, where a histogram is obtained by computing the test statistic on many different permutations $\{x_i, y_{\pi(i)}\}_{i=1}^n$, to decouple X and Y.

In the case where the samples $Z_t = (X_t, Y_t)$ are drawn from a random process, the analysis of the asymptotic behaviour of HSIC requires substantially more effort than in the i.i.d. case. As our main contribution, we obtain both the null and alternative distributions of HSIC for random processes, where the null distribution is defined as X_t being independent of Y_t at time t. Such a test may be used for rejecting causal effects (i.e., whether one signal is not dependent on the values of another signal at a particular delay) or instant coupling (see our first experiment in Section A.4).[1] The null distribution is again an infinite weighted sum of χ^2 variables, however these are now correlated, rather than independent. Under the alternative hypothesis, the statistic has an asymptotically normal distribution.

For the test to be used in practice, we require an empirical estimate of the null distribution, which gives the correct test threshold when $Z_t = (X_t, Y_t)$ is a random process. Evidently, the bootstrap procedure used in the i.i.d. case is incorrect, as the temporal dependence structure within the Y_t will be removed. This turns out to cause severe problems in practice, since the permutation procedure will give an increasing rate of false positives as the temporal dependence of the Y_t increases (i.e., dependence will be detected between X_t and Y_t, even though none exists, this is also known as a Type I error). Instead, our null estimate is obtained by making shifts of one signal relative to the other, so as to retain the dependence structure within each signal. Consequently, we are able to keep the Type I error at the designed level $\alpha = 0.05$. In our experiments, we address three examples: one artificial case consisting of two signals which are dependent but have no correlation, and two real-world examples on forex data. HSIC for random processes reveals dependencies that classical

[1] We distinguish our case from the problem of ensuring time series are independent simultaneously across all time lags, e.g the null will hold even if $X_t = Y_{t-1}$ where Y_t is white noise.

approaches fail to detect. Moreover, our new approach gives the correct Type I error rate, whereas a bootstrap-based approach designed for i.i.d. signals returns too many false positives.

Related work Prior work on testing independence in time series may be categorized in two branches: testing serial dependence within a single time series, and testing dependence between one time series and another. The case of serial dependence turns out to be relatively straightforward, as under the null hypothesis, the samples become independent: thus, the analysis reduces to the i.i.d. case. Pinkse [77], Diks and Panchenko [29] provide a quadratic forms function-based serial dependence test which employs the same statistic as HSIC. Due to the simple form of the null hypothesis, the analysis of [91, Ch. 5] applies. Further work in the context of the serial dependency testing includes simple approaches based on rank statistics e.g. Spearman's correlation or Kendall's tau, correlation integrals e.g. [19]; criteria based on integrated squared distance between densities e.g [84]; KL-divergence based criteria e.g. [83, 58]; and generalizations of KL-divergence to so called q-class entropies e.g. [41, 78].

In most of the tests of independence of two time series, specific conditions have been enforced, e.g that processes follow a moving average specification or the dependence is linear. Prior work in the context of dependency tests of two time series includes cross covariance based tests e.g. [53, 57, 92]; and a Generalized Association Measure based criterion [32]. Some work has been undertaken in the non-parametric case, however. A non-parametric measure of independence for time series, based on the Hilbert-Schmidt Independence Criterion, was proposed by Zhang et al. [116]. While this work established the convergence in probability of the statistic to its population value, no asymptotic distributions were obtained, and the statistic was not used in hypothesis testing. To our knowledge, the only non-parametric independence test for pairs of time series is due to Besserve et al. [14], which addresses the harder problem of testing independence across all time lags simultaneously. [2] The procedure is to compute the Hilbert-Schmidt norm of a cross-spectral density operator (the Fourier transform of the covariance operator at each time lag). The resulting statistic is a function of frequency, and must be zero at all frequencies for independence, so a correction for multiple hypothesis testing is required. It is not clear how

[2] Let X_t follow a MA(2) model and put $Y_t = X_{t-20}$. This is a case addressed by Besserve et al. [14], who will reject their null hypothesis, whereas our null is accepted

the asymptotic analysis used in the present work would apply to this statistic, and this remains an interesting topic of future study.

The remaining material is organized as follows. In Section A.2 we provide a brief introduction to random processes and various mixing conditions, and an expression for our independence statistic, HSIC. In Section A.3, we characterize the asymptotic behaviour of HSIC for random variables with temporal dependence, under the null and alternative hypotheses, and establish the test consistency. We propose an empirical procedure for constructing a statistical test, and demonstrate that the earlier bootstrap approach will not work for our case. Section A.4 provides experiments on synthetic and real data.

Background

In this section we introduce necessary definitions referring to random processes. We then go on to define a V-statistic estimate of the Hilbert-Schmidt Independence Criterion, which applies in the i.i.d. case.

Random process. First, we introduce the probabilistic tools needed for pairs of time series. Let $(Z_t, \mathcal{F}_t)_{t\in\mathbb{N}}$ be a strictly stationary sequence of random variables defined on a probability space Ω with a probability measure P and natural filtration $\mathcal{F}_t$. Assume that Z_t denotes a pair of random variables i.e. $Z_t = (X_t, Y_t)$, where X_t is defined on $\mathcal{X}$, and Y_t on $\mathcal{Y}$. Each Z_t takes values in a measurable Polish space $(\mathbf{Z}, \mathcal{B}(\mathbf{Z}), P_{\mathbf{Z}})$. The space $\mathbf{Z}$ is a Cartesian product of two Polish spaces $\mathbf{X}$ and $\mathbf{Y}$, endowed with a natural Borel sigma field and a probability measure.

We introduce a sequence of independent copies of Z_0, i.e., $(Z_t^*)_{t\in\mathbb{N}}$. Since Z_t is stationary, Z_t^* retains the dependence between random variables X_t and Y_t, but breaks the temporal dependence.

Next, we formalize a concept of memory of a process. A process is called absolutely regular (β-mixing) if $\beta(m) \to 0$, where

$$\beta(m) = \frac{1}{2} \sup_n \sup \sum_{i=1}^{I} \sum_{j=1}^{J} |P(A_i \cap B_j) - P(A_i)P(B_j)|.$$

The second supremum in the $\beta(m)$ definition is taken over all pairs of finite partitions $\{A_1, \cdots, A_I\}$ and $\{B_1, \cdots, B_J\}$ of the sample space such that $A_i \in \mathcal{A}_1^n$ and $B_j \in \mathcal{A}_{n+m}^{\infty}$, and $\mathcal{A}_b^c$ is a sigma field spanned by a subsequence, $\mathcal{A}_b^c =$

$\sigma(Z_b, Z_{b+1}, ..., Z_c)$. A process is called uniform mixing (ϕ-mixing) if $\phi(m) \to 0$, where

$$\phi(m) = \sup_n \sup_{A \in \mathcal{A}_1^n} \sup_{B \in \mathcal{A}_{n+m}^\infty} |P(B|A) - P(B)|.$$

Uniform mixing implies absolute regularity, i.e. $\beta(m) \leq \phi(m)$ [17]. Under technical assumptions, Autoregressive Moving Average processes — or more generally Markov Chains — are absolutely regular or uniformly mixing [30].

Hilbert-Schmidt Independence Criterion Let k, l be positive definite kernels associated with respective reproducing kernel Hilbert spaces $\mathcal{H}_\mathcal{X}$ on $\mathcal{X}$, and $\mathcal{H}_\mathcal{Y}$ on $\mathcal{Y}$. We assume that k and l are bounded and continuous. We associate to the random variable X a mean embedding $\mu_X(x) := \mathcal{E}_X k(X, x)$, such that $\forall f \in \mathcal{H}_\mathcal{X}$, $\langle f, \mu_X \rangle_{\mathcal{H}_\mathcal{X}} = \mathcal{E}_X(f(X))$ [12, 95]. We assume k, l are characteristic kernels, meaning the mappings μ_X and $\mu_Y(y) := \mathcal{E}_Y l(Y, y)$ are injective embeddings of the probability measures to the corresponding RKHSs; i.e., distributions have unique embeddings [38, 97].

We next recall a measure of statistical dependence, the Hilbert-Schmidt Independence Criterion (HSIC), which can be expressed in terms of expectations of RKHS kernels [43, 44]. Denote a group of permutations over 4 elements by S_4, with π one of its elements, i.e., a permutation of four elements. We define

$$h(z_1, z_2, z_3, z_4) = \frac{1}{4!} \sum_{\pi \in S_4} k(x_{\pi(1)}, x_{\pi(2)})[l(y_{\pi(1)}, y_{\pi(2)}) + l(y_{\pi(3)}, y_{\pi(4)}) - 2l(y_{\pi(2)}, y_{\pi(3)})].$$

Lemma 18. *Let γ be an expected value of the function h, $\gamma = Eh(Z_1^*, Z_2^*, Z_3^*, Z_4^*)$. This expectation corresponds to HSIC, computed using a function symmetric in its arguments. For k and l characteristic, continuous, translation invariant, and vanishing at infinity, γ is equal to zero if and only if the null hypothesis holds (see [71, Lemma 3.8], applying [98, Proposition 2], and the note at the end of Section A.5).*

The value of γ corresponds to a distance between embeddings of (X_1^*, Y_2^*) and (X_1^*, Y_1^*) to an RKHS with the product kernel $\kappa = k \cdot l$ [46, Section 7]. A biased empirical estimate of the Hilbert-Schmidt Independence Criterion can be expressed as a V-statistic (the unbiased estimate is a U-statistic, however the difference will be accounted for when constructing a hypothesis test, through

an appropriate null distribution).

V statistics. A V-statistic of a k-argument, symmetric function f is written

$$V(f,Z) = \frac{1}{n^k} \sum_{1 \leq i_1, \cdots, i_k \leq n} f(Z_{i_1}, ..., Z_{i_k}). \tag{A.1}$$

Gretton et al. [43] show that the biased estimator of γ is $V(h,Z)$. The asymptotic behaviour of this statistic depends on the degeneracy of the function that defines it. We say that a k-argument, symmetric function f is j-degenerate ($j < k$) if for each $z_1, \cdots, z_j \in \mathbb{Z}$,

$$Ef(z_1, \cdots, z_j, Z^*_{j+1}, \cdots, Z^*_k) = 0.$$

If $j = k-1$ we say that the function is canonical. We refer to a normalized V statistic as a V-statistic multiplied by the sample size, $n \cdot V$.

HSIC for random processes

In this section we construct the Hilbert-Schmidt Independence Criterion for random processes, and define its asymptotic behaviour. We then introduce an independence testing procedure for time series.

We introduce two hypotheses: the null hypothesis $\mathbf{H_0}$ that X_t and Y_t are independent, and the alternative hypothesis $\mathbf{H_1}$ that they are dependent. To build a statistical test based on $n \cdot V(h,Z)$ we need two main results. First, if null hypothesis holds, we show $n \cdot V(h,Z)$ converges to a random variable. Second, if the null hypothesis does not hold, the $n \cdot V(h,Z)$ estimator diverges to infinity. Following these results, the Type I error (the probability of mistakenly rejecting the null hypothesis) will stabilize at the design parameter α, and the Type II error (the probability of mistakenly accepting the null hypothesis when the variables are dependent) will drop to zero, as the sample size increases.

We begin by introducing an auxiliary kernel function s, and characterize the normalized V-statistic distribution of s using a CLT introduced by [15]. We then show that the normalized V-statistic associated with the function s has the same asymptotic distribution as the $n \cdot V(h,Z)$ distribution.

Let s be an auxiliary function $s(z_1, z_2) = \tilde{k}(x_1, x_2)\tilde{l}(y_1, y_2)$, where

$$\begin{aligned}\tilde{k}(x_1, x_2) =& k(x_1, x_2) - Ek(x_1, X_2) \\ & - Ek(X_1^*, x_2) + Ek(X_1^*, X_2^*),\end{aligned}$$

and $\tilde{l}$ is defined similarly.

Both $\tilde{k}$ and $\tilde{l}$ are kernels, meaning that they are dot products between features centred in their respective RKHSs [12]. Therefore $s = \tilde{k} \cdot \tilde{l}$ defines a kernel on a product space of pairs Z_t. Using Mercer's Theorem we obtain an expansion for s.

Statement 2. *By Steinwart and Scovel [104] Corollary 3.5, the bounded, continuous kernel s has a representation*[3]

$$s(z_a, z_b) = \sum_{i=1}^{\infty} \lambda_i e_i(z_a) e_i(z_b), \tag{A.2}$$

where $(e_i)_{i \in \mathbb{N}^+}$ denotes an orthonormal basis of $L^2(\mathbf{Z}, \mathcal{B}(\mathbf{Z}), P_{\mathbf{Z}})$. The series $(\sum_{i=1}^{N} \lambda_i \, e_i(z_a) e_i(z_b))$ converges absolutely and uniformly. e_i are eigenfunctions of s and λ_i are eigenvalues of s.

We will henceforth assume that for every collection of pairwise distinct subscripts (t_1, t_2), the distribution of (Z_{t_1}, Z_{t_2}) is absolutely continuous with respect to the $(Z_{t_1}^*, Z_{t_2}^*)$ distribution. This assumption prevents the occurrence of degenerate cases, such that all Z_t being the same. The following results are proved in Section A.5.

Lemma 19. *Let the process Z_t have a mixing coefficient smaller than m^{-3} $(\beta(m), \phi(m) \leq m^{-3})$ and satisfy either of the following conditions:*

A *Z_t is ϕ-mixing.*

B *Z_t is β-mixing. For some $\epsilon > 0$ and for an even number $c \geq 2$, the following holds*

1. *$\sup_i E|e_i(X_1)|^{2+\epsilon} \leq \infty$, where e_i is the basis introduced in the Statement 2 and $|\cdot|$ denotes an absolute value.*
2. *$\sum_{m=1}^{\infty} \beta^{\epsilon/(2+\epsilon)}(m) < \infty$.*

[3] A bounded kernel is compactly embedded into $L^2(\mathbf{Z}, \mathcal{B}(\mathbf{Z}), P_{\mathbf{Z}})$ [104].

If the null hypothesis holds, then s is a canonical function and a kernel. What is more,

$$\lim_{n\to\infty} n \cdot V(s,Z) \stackrel{D}{=} \sum_{j}^{\infty} \lambda_j \tau_j^2,$$

where τ_j is a centred Gaussian sequence with the covariance matrix

$$\begin{aligned} E\tau_a\tau_b = Ee_a(Z_1)e_b(Z_1) + \\ + \sum_{j=1}^{\infty} \left[Ee_a(Z_1)e_b(Z_{j+1}) + Ee_b(Z_1)e_a(Z_{j+1})\right]. \end{aligned}$$

We now characterize the asymptotics of $V(h,Z)$.

Theorem 9. *Under assumptions of Lemma 19, if $\mathbf{H_0}$ holds, then the asymptotic distribution of the empirical HSIC (with scaling n) is the same as that of $n \cdot V(s,Z)$,*

$$\lim_{n\to\infty} n \cdot V(h,Z) \stackrel{D}{=} \lim_{n\to\infty} n \cdot V(s,Z).$$

Theorem 10. *Under assumptions of the Lemma 19, if $\mathbf{H_1}$ holds, then $\gamma > 0$ and $\sqrt{n}(V(h,Z) - \gamma)$ has asymptotically normal distribution with mean zero and finite variance.*

Consequently, if the null hypothesis does not hold then $P(n \cdot V(h,Z) > C) = P(V(h,Z) > \frac{C}{n}) \to 1$ for any fixed C. Finally, we show that the γ estimator is easy to compute. According to Gretton et al. [44, equation 4], $V(h,Z) = n^{-2} trHKHL$, where $K_{ab} = k(X_a, X_b)$, $L_{ab} = l(Y_a, Y_b)$,$H_{ij} = \delta_{ij} - n^{-1}$ and n is a sample size.

Testing procedure We begin by showing that the H_0 distribution of the γ estimator obtained via the bootstrap approach of [29, 44] gives an incorrect p-value estimate when used with independent random processes. In fact, the null hypothesis obtained by permutation corresponds to the processes being *both* i.i.d. *and* independent from each other. Recall the covariance structure of the γ estimator from Theorem 9,

$$\begin{aligned} E\tau_a\tau_b = Ee_a(Z_1)e_b(Z_1) + \\ + \sum_{j=1}^{\infty} \left[Ee_a(Z_1)e_b(Z_{j+1}) + Ee_b(Z_1)e_a(Z_{j+1})\right]. \end{aligned} \tag{A.3}$$

We can represent e_a and e_b as $e_a(z) = e_u^X(x)e_o^Y(y)$, $e_b(z) = e_i^X(x)e_p^Y(y)$, as a decomposition of the $\mathbf{Z}$ basis into bases of $\mathbf{X}$ and $\mathbf{Y}$, respectively. Consider a partial sum T_n of series from the above equation (A.3), with X_t replaced with its permutation $X_{\pi(t)}$,

$$T_n = \sum_{j=1}^{n} Ee_u^X(X_{\pi(1)})e_i^X(X_{\pi(j+1)})Ee_o^Y(Y_1)e_p^Y(Y_{j+1}). \tag{A.4}$$

Using covariance inequalities from [30, Section 1.2.2] we conclude that $Ee_o^Y(Y_1)e_p^Y(Y_{j+1}) = O(\Lambda(j)^{\frac{1}{2}})$ and $Ee_u^X(X_{\pi(1)})e_i^X(X_{\pi(j+1)}) = O(\Lambda(|\pi(j) - \pi(1)|)^{\frac{1}{2}})$ where Λ is an appropriate mixing coefficient (β or ϕ). Recall that $0 < \Lambda(j) < Cj^{-3}$.

We can therefore reduce the problem to the convergence of a random variable

$$S_n = \sum_{j=1}^{n} \Lambda(j)^{\frac{1}{2}}\Lambda(|\pi(j) - \pi(1)|)^{\frac{1}{2}}, \tag{A.5}$$

where π is a random permutation drawn from the uniform distribution over the set of n-element permutations. In the supplementary material we show that this sum converges in probability to zero.

Since $S_n > T_n > 0$, then T_n converges to zero in probability, and consequently the covariance matrix entry $E\tau_a\tau_b$ converges to unity for $a = b$, and to zero otherwise. Indeed, the expected value $Ee_a((X_{\pi(1)}, Y_1))e_b((X_{\pi(1)}, Y_1)) = 0$ if $a \neq b$ and is equal to one otherwise. Note that this is the covariance matrix described by Gretton et al. [44].

A correct approach to approximating the asymptotic null distribution of $n \cdot V(h, Z)$ under $\mathbf{H}_0$ is by *shifting* of one time series relative to the other. Define the shifted process $S_t^c = Y_{t+c \bmod n}$ for an integer c, $0 \leq c \leq n$ and $0 \leq t \leq n$. If we let c vary over $0 \leq A \leq B \leq n$ for A such that the dependence between Y_{t+A} and X_t is negligible, then we can approximate the null distribution with an empirical distribution calculated on points $(V(h, Z^k))_{A \leq k \leq B}$, where $Z_t^k = (X_t, S_t^k)$. This is due to the fact that the shifted process S_t^c retains most of the dependence, since it does not scramble the time index.[4] We call this method Shift HSIC. In the supplementary material we show that Shift HSIC samples

[4] As a illustration, consider $W_t = Y_{t-10}$. If Y_t is stationary then the dependence structure of (W_{t_1}, W_{t_2}) and (Y_{t_1}, Y_{t_2}) is the same. If we set $W_t = Y_{\pi(t)}$ this property does not hold.

from the correct null distribution.

Experiments

In the experiments we compare Shift HSIC with the Bootstrap HSIC of Gretton et al. [44]. We investigate three cases: an artificial dataset, where two time series are coupled non-linearly; and two forex datasets, where in one case we seek residual dependence after one time series has been used to linearly predict another, and in the other case, we reveal strong dependencies between signals that are not seen via linear correlation.

Artificial data

Non-linear dependence. We investigate two dependent, autoregressive random processes X_t,Y_t, specified by

$$X_t = aX_{t-1} + \epsilon_t \quad Y_t = aY_{t-1} + \eta_t, \tag{A.6}$$

with an autoregressive component a. The coupling of the processes is a result of the dependence in the innovations ϵ_t, η_t. These ϵ_t, η_t are drawn from an Extinct Gaussian distribution, defined in Algorithm 2. The parameter p (called extinction rate) controls how often a point drawn form a ball $B(0,r)$ dies off. According to Algorithm 2, the probability of seeing a point inside the ball $B(0,r)$ is different than for a two dimensional Gaussian $N(\mathbf{0}, Id)$. On the other hand, as p goes to zero, the Extinct Gaussian converges in distribution to $N(\mathbf{0}, Id)$. Figure A.1 illustrates the joint distribution of X_t, Y_t. The left scatter plot in Figure A.1 presents X_t and Y_t generated with an extinction rate of 50%, while the right hand plot is generated with an extinction rate of 99.87%. Processes used in this experiment had an autoregressive component of 0.2, and the radius of the innovation process was 1.

Figure A.2 compares the power of the Shift HSIC test and the correlation test. The X axis represents an extinction rate, while the Y axis shows the true positive rate. Shift HSIC is capable of detecting non-linear dependence between X_t and Y_t, which is missed by linear correlation. The red star depicts performance of the KCSD algorithm developed by Besserve et al. [14], with parameters tuned by its authors: note that this result required using four times as many data points as HSIC.

Algorithm 2 Generate innovations

Input: extinction rate $0 \leq p \leq 1$, radius r.
repeat
  Initialize η_t, ϵ_t to $N(0,1)$ and d to a number uniformly distributed on $[0,1]$.
  if $\eta_t^2 + \epsilon_t^2 > r^2$ **or** $d > p$ **then**
    return η_t, ϵ_t
  end if
until true

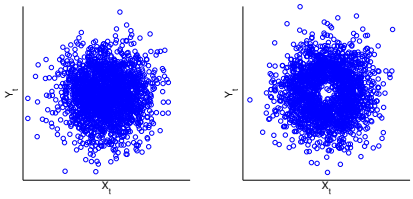

Figure A.1: X_t and Y_t, described in the Experiment A.4, with extinction rates 50% (left) and 99.8% (right), respectively.

False positive rates. We next investigate the rate of false positives for Shift HSIC and Bootstrap HSIC on independent copies of the $AR(1)$ processes used in the previous experiment. To generate independent processes, we first sampled two pairs (X_t, Y_t), (X'_t, Y'_t) of time series using (A.6), and then constructed Z by taking X from the first pair and Y from the second, i.e., $Z_t = (X_t, Y'_t)$. We set an extinction rate to 50%. [5] The AR component a in the model (A.6) controls the memory of a processes - the larger this component, the longer the memory. We performed the Shift HSIC and the Bootstrap HSIC tests on Z_t generated under $\mathbf{H_0}$ with different AR components. Figure A.3 illustrates the results of this experiment. The X axis is indexed by the AR component and Y axis shows the FP rate. As the temporal dependence increases, the Bootstrap HSIC incorrectly gives an increasing number of false positives: thus, it cannot be relied on to detect dependence in time series. The Shift HSIC false positive rate remains at the targeted 5% p-value level.

Forex data

We use Foreign Exchange Market quotes to evaluate Shift HSIC performance on the real life data. Practitioners point out that forex time series are noisy

[5] As a reviewer pointed out, the example for the FP rates can be simplified, however we decided to be consistent with the marginal distribution of X_t,Y_t across the experiments.

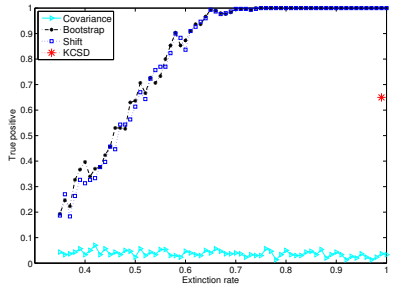

Figure A.2: True positive rate for the Shift HSIC, the Bootstrap HSIC and correlation based test: sample size 1200, results averaged over 300 repetitions. The red star shows KCSD performance at $4\times$ the HSIC sample size; see Section A.4 for details.

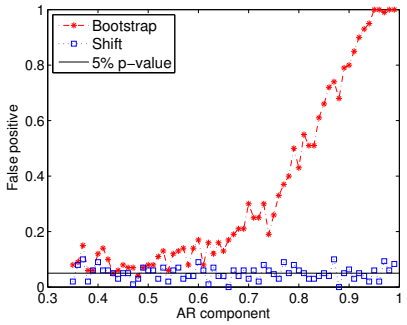

Figure A.3: False positive rate for the Shift HSIC and the Bootstrap HSIC. The sample size was 1200, and results were averaged over 300 repetitions.

and hard to handle, especially at low granulations (smaller then 15 minutes). We decided to work with forex time series to show that Shift HSIC can detect dependence even on such a difficult dataset. The forex time series were granulated to obtain two minute sampling (the granulation function returned the last price in the two minute window). Using the test of Diks and Panchenko [29], we checked that serial dependence of the differentiated time series decays fast enough to satisfy the assumed mixing conditions (by a differentiated time series, we refer to $(X_t - X_{t-1})_{t\in\mathbf{N}}$). The choice of the pairs and trading day (21st January 2013) were arbitrary.

Instantaneous coupling and causal effect. Having one Australian dollar we may obtain a quantity of Yen in two ways, either by using AUD/JPY exchange rate explicitly or by buying Canadian dollars and then selling them at the CAD/JPY rate. Let X_t be a differentiated AUD/JPY exchange rate and Y_t be

a differentiated product of exchange rates AUD/CAD×CAD/JPY. We will investigate the relation between these two. Common sense dictates that Y_t should behave similarly to X_t. After examining the cross-correlation of X_t and Y_t, we propose a simple regression model to describe the interaction between the signals,

$$\hat{Y}_t = a_0 X_t + a_1 X_{t-1} + \cdots + a_6 X_{t-6}.$$

We fit the model and see that $a_0 = 0.97$, and the remaining coefficients are not bigger then 0.06 in absolute value. This suggest that most of the dependence is explained by an instantaneous coupling. We further investigate the cross-correlation between residuals $R_t = Y_t - \hat{Y}_t$ and X_t. We observe no significant correlations in the first 30 lags.

Next we investigate dependence of residuals with lagged values of the explanatory variables, i.e., R_t with X_{t-k} for $k \in (0, \cdots, 30)$. After calculating p-values using the Bootstrap HSIC and the Shift HSIC, we discover dependence only at lags 4, 5, 9, 13 and 29, as presented in the Figure A.4. Lack of the dependence at lag zero suggests that the linear model for coupling is reasonable. However, both the Bootstrap HSIC and the Shift HSIC support the hypothesis that there is a strong relation at lag 5, which is not explained well by the linear model.

The questions remains whether test statistics at lags 4, 9, 13 and 29 indicate further model misspecification. Under $\mathbf{H_0}$, at a significance level 94%, we expect 1.8 out of 30 statistics to be higher than the 94th percentile. Excluding the statistic at lag 5, the Shift HSIC test reports two statistics above this percentile, while Bootstrap HSIC reports four. Should the statistics at the different lags be independent from each other, the probabilities of seeing two and four statistics above the percentile are respectively 25% and 6%. Shift HSIC indicates that the model fits the data well, while the Bootstrap HSIC suggests that some non-linear dependencies remain unexplained.

Dependence structure. The data are five currency pairs. A correlation based independence test, and the Shift HSIC test, were performed on each pair of currencies. The dependencies revealed by these tests are depicted in Figure A.5 - nodes represent the time series and edges represent dependence. Shift HSIC reveals a strong coupling between EUR/RUB and USD/JPY, HKD/JPY and XAU/USD that was not found by simple correlation. All edges revealed by Shift HSIC have p-values at most at level 0.03 - clearly, the Shift HSIC managed to

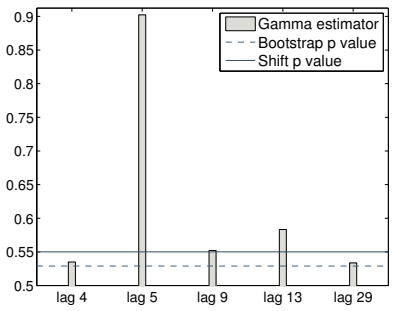

Figure A.4: Instantaneous coupling. Results for 720 samples, null threshold of Shift HSIC used 300 lags in range $100 - 400$.

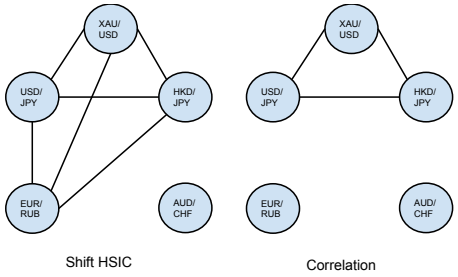

Figure A.5: Differences between the dependence structure on the forex revealed by the Shift HSIC and covariance. Parameter settings are as in Figure A.4.

find a strong non-linear dependence. Note that the obtained graphs are cliques.

Proofs

A U-statistic of a k-argument, symmetric function f, is written

$$U(f,Z) = \binom{n}{k}^{-1} \sum_{1 \le i_1 < \cdots < i_k \le n} f(Z_{i_1},...,Z_{i_k}).$$

A decomposition due to Hoeffding allows us to decompose this U-statistic into a sum of U-statistics of canonical functions, $U(h,Z) = \sum_{k=1}^{l} \binom{l}{k} U(h_k,Z)$, where $h_k(z_1,...,z_l)$ are components of the decomposition. According to Serfling [91, section 5.1.5], each of h_1,h_2,h_3,h_4 is symmetric and canonical. Note that h_k is defined using independent samples Z^* - this is because the CLT or LLN state that U-statistics or V-statistics of mixing processes converge to their expected value taken with respect to independent copies, i.e., Z^*. Under $\mathbf{H}_0$, h_1 is equal to zero everywhere and $h_2 = \frac{1}{6}s$, where these results were obtained by Gretton

et al. [44].[6] See supplementary material for details.

In order to characterize $U(h, Z)$, we show that under null hypothesis $U(h_2, Z)$ converges to a random variable, and both $U(h_3, Z)$,$U(h_4, Z)$ converge to zero in a probability (the latter proof can be found in the supplementary material). Bellow we characterise $U(h_2, Z)$ convergence.

Lemma 20. *Under assumptions of Lemma 19,*

$$\lim_{n\to\infty} n \cdot U(h_2, Z) \stackrel{D}{=} \frac{1}{6}\sum_{i_1}^{\infty} \lambda_{i_1}(\tau_{i_1}^2 - 1).$$

Proof. First recall that under null hypothesis $h_2 = \frac{1}{6}s$. We will check the conditions of [15, Theorem 1] (also available in the supplementary).
First, from Mercer's Theorem [104, Corollary 3.5], we deduce that the h_2 coefficients in $L_2(\mathbf{Z}, \mathcal{B}_{\mathbf{Z}}, P_{\mathbf{Z}})$ are absolutely summable. In the supplementary material we show that $Ee_i(Z_1^*) = 0$.
Recall the assumptions of Lemma 19. If **A** holds then $\sum_{k=1}^{\infty} \phi(k)^{\frac{1}{2}} < \infty$ and $\sup_i E|e_i(X_1)|^2 = 1 < \infty$ (e_i is an orthonormal eigenfunction). Finally, if **B** holds then the process Z_t is α-mixing. The remaining assumptions concerning uniform mixing in Borisov and Volodko [15] are exactly the same as in this lemma.

□

Main body proofs

Proof. (Lemma 19) We use the fact that h_2 is equal to s up to scaling ($6U(h_2, Z) = U(s, Z)$), and Lemma 20, to see that $nU(s, Z) \stackrel{D}{\to} \sum_i^{\infty} \lambda_i(\tau_i^2 - 1)$. Since $Es(Z_t, Z_t) = E\sum_{i=1}^{\infty} \lambda_i e_i(Z_t)^2 = \sum_{i=1}^{\infty} \lambda_i$, then by the LLN for mixing processes,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i}^{n} s(Z_i, Z_i) \stackrel{P}{=} \sum_{i=1}^{\infty} \lambda_i. \tag{A.7}$$

We use a relationship between U and V statistics,

$$\begin{aligned}\lim_{n\to\infty} nV(s, Z) &\stackrel{D}{=} \lim_{n\to\infty} nU(s, Z) + \lim_{n\to\infty} \frac{1}{n}\sum_{i}^{n} s(Z_i, Z_i) \\ &\stackrel{D}{=} \sum_{i=1}^{\infty} \lambda_i + \sum_{i}^{\infty} \lambda_i(\tau_i^2 - 1) \stackrel{D}{=} \sum_{i}^{\infty} \lambda_i \tau_i^2.\end{aligned}$$

[6] The second result is hard to locate - it is in appendix A.2, text between equations 12 and 13

□

Proof. (Theorem 9) We operate under the null hypothesis. Recall that $U(h,Z)$ can be decomposed as $U(h,Z)=\sum_{k=1}^{4}\binom{4}{k}U(h_k,Z)$. Here $h_1\equiv 0$. We show in the supplementary material that $U(h_3,Z)$ and $U(h_4,Z)$ tend to zero in probability. From Lemma 20,

$$\lim_{n\to\infty} nU(h,Z)\overset{D}{=}\lim_{n\to\infty} nU(s,Z)\overset{D}{=}\sum_{i}^{\infty}\lambda_i(\tau_i^2-1). \tag{A.8}$$

We define an auxiliary symmetric function w,

$$\begin{aligned} w(z_1,z_2,z_3) = {} & h(z_1,z_1,z_2,z_3)+h(z_1,z_2,z_2,z_3)\\ & +h(z_1,z_2,z_3,z_3)+h(z_1,z_1,z_3,z_2)\\ & +h(z_3,z_2,z_2,z_1)+h(z_2,z_1,z_3,z_3). \end{aligned}$$

It is obvious that $Ew(Z_1^*,Z_2^*,Z_3^*)=6Eh(Z_1^*,Z_1^*,Z_2^*,Z_3^*)$. We consider the difference between the unnormalized V and U statistics,

$$S_n=\sum_{1\le i_1,i_2,i_3,i_4\le n} h(Z_{i_1},...,Z_{i_4})-\sum_{i\in C_4} h(Z_{i_1},...,Z_{i_4}),$$

where $\sum_{i\in C_m}$ denotes summation over all $\binom{n}{m}$ combinations of m distinct elements $\{i_1,\cdots,i_m\}$ from $\{1,\cdots,n\}$. The difference is equal to the sum over 4-tuples with at least one pair of equal elements. We can choose such tuples in $\binom{4}{2}=6$ ways. Observe that w covers the choice of all these six tuples. Since for any $z_1,z_2\in\mathbf{Z}$, $h(z_1,z_1,z_1,z_2)=0$, then w is zero whenever more than two indices are equal. Therefore we can sum w over distinct indices z_1,z_2,z_3,

$$S_n=\sum_{i\in C_3} w(Z_{i_1},Z_{i_2},Z_{i_3}).$$

We see that S_n is almost a U-statistic ($U(w,Z)$). By the CLT for U-statistics from Denker and Keller [28], Theorem 1(c), we obtain

$$\lim_{n\to\infty}\frac{1}{n(n-1)(n-2)}S_n\overset{P}{=}6Eh(Z_1^*,Z_1^*,Z_2^*,Z_3^*).$$

On the other hand, via the relation $h_2=\frac{1}{6}s$ and the h_2 definition, we get

$Es(Z_1^*, Z_1^*) = 6Eh(Z_1^*, Z_1^*, Z_2^*, Z_3^*)$, and therefore

$$\lim_{n\to\infty} \frac{1}{n(n-1)(n-2)} S_n \stackrel{P}{=} \lim_{n\to\infty} \frac{1}{n} \sum_{i}^{n} s(Z_i, Z_i). \tag{A.9}$$

Finally, we rewrite S_n as

$$\sum_{1\le i_1,i_2,i_3,i_4\le n} h(Z_{i_1},...,Z_{i_4}) = S_n + \sum_{i\in C_4} h(Z_{i_1},...,Z_{i_4}).$$

We normalize by $\frac{1}{n(n-1)(n-2)}$, and take the limit in n,

$$\begin{aligned}\lim_{n\to\infty} \frac{n^4}{n(n-1)(n-2)} V(h,Z) &\stackrel{D}{=} \\ &= \lim_{n\to\infty} \left(\frac{1}{n(n-1)(n-2)} S_n + (n-4)U(h,Z) \right).\end{aligned}$$

We substitute (A.9) and (A.8) on the right hand side, and use equation (A.7) from Lemma 19 to replace $\lim_{n\to\infty} \frac{1}{n} \sum_i^n s(Z_i, Z_i)$ with $\sum_{i=1}^{\infty} \lambda_i$, yielding

$$\begin{aligned}\lim_{n\to\infty} n \cdot V(h,Z) &\stackrel{D}{=} \\ &\stackrel{D}{=} \lim_{n\to\infty} \frac{1}{n} \sum_{i}^{n} s(Z_i, Z_i) + \lim_{n\to\infty} \frac{1}{n} \sum_{i,j}^{n} s(Z_i, Z_j) \stackrel{D}{=} \\ &\stackrel{D}{=} \sum_{i=1}^{\infty} \lambda_i + \sum_{i}^{\infty} \lambda_i(\tau_i^2 - 1) \stackrel{D}{=} \sum_{i}^{\infty} \lambda_i \tau_i^2.\end{aligned}$$

□

Proof. (Theorem 10) If the null hypothesis does not hold, then $\gamma > 0$ [43]. In this case h is nondegenerate, and we can use Denker and Keller [28, Theorem 1(c)] to see that $\frac{\sqrt{n}}{4\sqrt{\sigma}}(V(h,Z) - \gamma) \sim N(0,1)$, where σ is finite (see the note below Theorem 1 of [28], stating that in case (c) σ^2 is finite, and the note above Theorem 1 stating that $\sigma^2 = \lim_{n\to\infty} n^{-1}\sigma_n^2$). □

Proof. (Lemma 18) We use Lemma 1 and Theorem 4 from Gretton et al. [43] to show that $Eh(Z_1^*, Z_2^*, Z_3^*, Z_4^*) = 0$ iff (X_1^*, Y_1^*) has a product distribution. Since $Z_1^* \stackrel{D}{=} Z_1$ and $Z_t \stackrel{D}{=} Z_1$, we infer that X_t is independent from Y_t iff $Eh(Z_1^*, Z_2^*, Z_3^*, Z_4^*) = 0$. □

Acknowledgements The authors thank the reviewers and colleagues for helpful feedback, especially M. Skomra, D. Toczydlowska, and A. Zaremba.

A Kernel Independence Test for Random Processes - Supplementary

The sections in the supplementary material are in the same order those in the article. In particular, the n-th reference to the supplementary in the article is n-th subsection in the supplementary material.

The arXiv version of the report and supplementary may be found at: `http://arxiv.org/abs/1402.4501`

Before we start, we cite [113, Lemma 1], which will be used below.

Lemma 21. *[113] Let $(Z_t)_{t\in\mathbb{N}+}$ be an absolutely regular process with a mixing coefficient $(\beta(n))_{n\in\mathbb{N}+}$. Let $(t_1, t_2, \cdot, t_l)$ be a non-decreasing l-tuple, and let j be an integer such that $2 \leq j \leq l$. Finally, let $g : \mathbb{R}^l \to \mathbb{R}$ be a measurable function satisfying*

$$\left(E|g(Z_{t_1}, \cdots, Z_{t_l})|^{1+\delta}\right) \leq M$$

for some $\delta > 0, M > 0$. Then

$$\left|Eg(Z_{t_1}, \cdots, Z_{t_l}) - Eg(Z_{t_1}, \cdots, Z_{t_{j-1}}, Z^*_{t_j}, \cdots, Z^*_{t_l})\right| \leq 4M^{\frac{1}{1+\delta}}\beta(t_j - t_{j-1})^{\frac{\delta}{1+\delta}}.$$

Note that if a function g is symmetric, then we can always reorder its arguments if necessary.

Testing procedure - convergence of S_n from equation (5).

Let π be a permutation drawn from a uniform distribution over the set of n-element permutations. We will prove that the random variable

$$Q_n = \sum_{i=1}^{n} \frac{1}{i^{\frac{3}{2}}} \frac{1}{|\pi(1) - \pi(i)|^{\frac{3}{2}}}$$

converges to zero in probability at rate $O(n^{-1})$. Since $0 \leq S_n \leq Q_n$, then S_n converges to zero in probability at the same rate.

Lemma 22. $E|\pi(1) - \pi(i)|^{-\frac{3}{2}} = O(n^{-1})$.

Proof. Let j be a positive integer smaller than n. Observe that the sum $\sum_i^n |j-i|^{-\frac{3}{2}}$ is finite,

$$\sum_{i}^{n} |j-i|^{-\frac{3}{2}} \leq 2\sum_{i}^{n} i^{-\frac{3}{2}} \leq 2\zeta\left(\frac{3}{2}\right), \tag{A.10}$$

where $\zeta(\cdot)$ is the Riemann zeta function. Now expand the expected value $E|\pi(1)-\pi(i)|^{-\frac{3}{2}}$ using a conditional expected value,

$$E|\pi(1)-\pi(i)|^{-\frac{3}{2}} = E(E|j-\pi(i)|^{-\frac{3}{2}}|\pi(1)=j) = \sum_{j=1}^{n}\frac{1}{n}(E|j-\pi(i)|^{-\frac{3}{2}}|\pi(1)=j) =$$
$$= \sum_{j=1}^{n}\frac{1}{n}\sum_{j\neq 1}^{n}\frac{1}{n-1}|j-i|^{-\frac{3}{2}} \leq \frac{1}{n(n-1)}\sum_{j=1}^{n}2\zeta\left(\frac{3}{2}\right) = 2\zeta\left(\frac{3}{2}\right)\frac{1}{n-1}. \tag{A.11}$$

□

Lemma 23. *If $k \neq j$ are positive integers smaller than n, then*

$$E|\pi(k)-\pi(1)|^{-\frac{3}{2}}|\pi(j)-\pi(1)|^{-\frac{3}{2}} = O\left(\frac{1}{n^2}\right)$$

Proof. We use the inequality (A.10) and properties of a conditional expected value.

$$\begin{aligned}
&E|\pi(k)-\pi(1)|^{-\frac{3}{2}}|\pi(j)-\pi(1)|^{-\frac{3}{2}} = E\left(E|\pi(k)-a|^{-\frac{3}{2}}|\pi(j)-a|^{-\frac{3}{2}}\big|\pi(1)=a\right)\\
&= \frac{1}{n}\sum_{a=1}^{n}\left(E|\pi(k)-a|^{-\frac{3}{2}}|\pi(j)-a|^{-\frac{3}{2}}\big|\pi(1)=a\right)\\
&= \frac{6}{n(n-1)(n-2)}\sum_{a\neq b, a\neq c, b\neq c}^{n}\frac{1}{|b-a|^{\frac{3}{2}}}\frac{1}{|c-a|^{\frac{3}{2}}}\\
&\leq \frac{1}{n(n-1)(n-2)}\sum_{a\neq b}^{n}\frac{2\zeta\left(\frac{3}{2}\right)}{|b-a|^{\frac{3}{2}}}\\
&\leq \frac{1}{n(n-1)(n-2)}\sum_{a}^{n}4\zeta\left(\frac{3}{2}\right)^2\\
&= \frac{1}{(n-1)(n-2)}4\zeta\left(\frac{3}{2}\right)^2\\
&= O\left(\frac{1}{n^2}\right)
\end{aligned} \tag{A.12}$$

□

Lemma 24. *Q_n converges to zero in probability. The convergence rate is $\frac{1}{n}$.*

Proof. First, using Lemma 22 , we compute the expected value of Q_n

$$EQ_n = E\sum_{i=1}^{n}\frac{1}{i^{\frac{3}{2}}}\frac{1}{|\pi(1)-\pi(i)|^{\frac{3}{2}}} = \sum_{i=1}^{n}\frac{1}{i^{\frac{3}{2}}}E\frac{1}{|\pi(1)-\pi(i)|^{\frac{3}{2}}} \leq \sum_{i=1}^{n}\frac{1}{i^{\frac{3}{2}}}\frac{1}{n}C \leq \frac{1}{n}C\zeta\left(\frac{3}{2}\right) = O\left(\frac{1}{n}\right).$$

Next, using Lemma 23, we compute the second moment

$$\begin{aligned} & E\left(\sum_{k=1}^{n}\frac{1}{k^{\frac{3}{2}}}\frac{1}{|\pi(k)-\pi(1)|^{\frac{3}{2}}}\right)\left(\sum_{j=1}^{n}\frac{1}{j^{\frac{3}{2}}}\frac{1}{|\pi(j)-\pi(1)|^{\frac{3}{2}}}\right) \\ & \leq E\left(C\frac{1}{n^2}\sum_{k\neq j}^{n}\frac{1}{k^{\frac{3}{2}}}\frac{1}{j^{\frac{3}{2}}}+\sum_{k=1}^{n}\frac{1}{k^3}\frac{1}{|\pi(k)-\pi(1)|^3}\right) \\ & \leq C\frac{1}{n^2}\zeta\left(\frac{3}{2}\right)^2+C'\frac{1}{n}\zeta(3)=O\left(\frac{1}{n}\right). \end{aligned} \tag{A.13}$$

Using the Chebyshev's inequality we obtain the required result. □

Testing procedure - Shift HSIC samples from the right distribution

We will investigate the value of the V-statistic for a shifted process i.e. $nV(h, Z^k)$.

Null hypothesis holds. If the null hypothesis holds, then X_t and Y_{t+k} are independent for any k. To see this, suppose that there exists k for which X_t and Y_{t+k} are dependent (the processes are stationary, so this is true for all t). The observation X_t depends on its past values: in particular, X_{t-k} is a parent of X_t. If in addition $X_{t-k} \to Y_t$, then Y_t and X_t will be dependent, as they share a parent.

We will use this fact to show that the $nV(h, Z^k)$ has the same distribution as the $nV(h, Z)$. Recall the covariance structure of $nV(h, Z)$ from Theorem 9,

$$E\tau_a\tau_b = Ee_a(Z_1)e_b(Z_1)+\sum_{j=1}^{\infty}\left[Ee_a(Z_1)e_b(Z_{j+1})+Ee_b(Z_1)e_a(Z_{j+1})\right]. \tag{A.14}$$

We represent e_a and e_b as $e_a(z) = e_u^X(x)e_o^Y(y)$, $e_b(z) = e_i^X(x)e_p^Y(y)$. This represents a decomposition of the basis of $\mathbf{Z}$ into basis of $\mathbf{X}, \mathbf{Y}$, respectively. Consider one of the above infinite sums with Y_t replaced with the shifted process S_t^k,

$$T_n = \sum_{j=1}^{n} Ee_a(Z_1^k)e_b(Z_{j+1}^k) = \sum_{j=1}^{n} Ee_u^X(X_1)e_i^X(X_{j+1})e_o^Y(S_1^k)e_p^Y(S_{j+1}^k). \tag{A.15}$$

We obtain the following covariance structure for $nV(h, Z^k)$,

$$
\begin{aligned}
T_n &= \sum_{j=1}^{n} Ee_u^X(X_1)e_i^X(X_{j+1})Ee_o^Y(S_1^k)e_p^Y(S_{j+1}^k) \\
&= \sum_{j=1}^{n-k-1} Ee_u^X(X_1)e_i^X(X_{j+1})e_o^Y(Y_{1+k})e_p^Y(Y_{j+1+k}) + \sum_{j=n-k}^{n} Ee_u^X(X_1)e_i^X(X_{j+1})e_o^Y(Y_{1+k})e_p^Y(Y_{1+n-k-j}) \\
&= \sum_{j=1}^{n-k-1} Ee_u^X(X_1)e_i^X(X_{j+1})Ee_o^Y(Y_{1+k})e_p^Y(Y_{j+1+k}) + \sum_{j=n-k}^{n} Ee_u^X(X_1)e_i^X(X_{j+1})Ee_o^Y(Y_{1+k})e_p^Y(Y_{1+n-k-j}) \\
&\leq \sum_{j=1}^{n-k} Ee_u^X(X_1)e_i^X(X_{j+1})e_o^Y(Y_1)e_p^Y(Y_{j+1}) + O(k(n-k)^{-\frac{3}{2}}).
\end{aligned}
$$

We have used the fact that Y_t is stationary, $Ee_o^Y(Y_{1+k})e_p^Y(Y_{j+1+k}) = e_o^Y(Y_1)e_p^Y(Y_{j+1})$ and that the pairs (X_1, X_{1+j}), (Y_{1+k}, Y_{j+1+k}) are independent (because X_t and Y_{t+k} are independent for all shifts k). For the second term,

$$
\sum_{j=n-k}^{n} Ee_u^X(X_1)e_i^X(X_{j+1})Ee_o^Y(Y_{1+k})e_p^Y(Y_{1+n-k-j}),
$$

we have used covariance inequalities from Doukhan [30, section 1.2.2] and our bounds on mixing coefficients to obtain that when $j \geq n-k$, then $E|e_u^X(X_1)e_i^X(X_{j+1})| \leq (n-k)^{-\frac{3}{2}}$ (and by e.g. Holders inequality, $Ee_o^Y(Y_{1+k})e_p^Y(Y_{1+(n-j)})$ is finite). The first component takes the form

$$
\sum_{j=1}^{n-k} Ee_u^X(X_1)e_i^X(X_{j+1})e_o^Y(Y_1)e_p^Y(Y_{j+1}) = \sum_{j=1}^{n-k} Ee_a(Z_1)e_b(Z_{j+1}).
$$

Here $\sum_{j=1}^{n-k} Ee_a(Z_1)e_b(Z_{j+1})$ converges to $\sum_{j=1}^{\infty} Ee_a(Z_1)e_b(Z_{j+1})$ from equation (A.14). Since $Ee_a(Z_1^k)e_b(Z_1^k) = Ee_a(Z_1)e_b(Z_1)$, the covariance structure from equation (A.14) is recovered.

Null hypothesis does not hold. In this case, the dependence between X_t and Y_{t+k} decreases as k increases, since the mixing coefficients for each of the time series converges to zero. In the limit of large k and n, the normalized V-statistic will converge to the null distribution, where X_t and Y_t are independent random processes. The proof of this result under the assumed mixing conditions, with suitable conditions on the increase of k with n, is a topic of future work (the next two sections give an outline of the results that would need to be established for the shifted process).

Proofs - Hoeffding decomposition

The Hoeffding decomposition [e.g. 91] allows us to decompose U-statistics into a sum of simpler U-statistics that can be easier to analyse. In the following section we will perform a Hoeffding decomposition of $U(h, Z)$ and investigate some of its properties. In the sequel we assume that k and l are bounded kernels. For the U statistic $U(h, Z)$, we call the function h a *core*.

Any U-statistic can be written as a sum of V-statistics with degenerate cores. To show this, we define the auxiliary functions

$$g_c(z_1, ...z_c) = Eh(z_1, ..., z_c, Z^*_{c+1}, ..., Z^*_m)$$

for each $c = 1, ..., m-1$ and put $g_m = h$.

We assume the expected value of the core with respect to starred $\{Z_t\}$ is zero, i.e., $Eh(Z^*_1, \cdots, Z^*_m) = 0$. The canonical functions that enable the core decomposition are

$$\begin{aligned}
h_1(z_1) &= g_1(z_1), \\
h_2(z_1, z_2) &= g_2(z_1, z_2) - h_1(z_1) - h_1(z_2), \\
h_3(z_1, z_2, z_3) &= g_3(z_1, z_2, z_3) - \sum_{1 \le i < j \le 3} h_2(z_i, z_j) - \sum_{1 \le i \le 3} h_1(z_i), \\
&\vdots \\
h_m(z_1, ..., z_m) &= g_m(z_1, ..., z_m) - \sum_{1 \le i_1 < ... < i_{m-1} \le m} h_{m-1}(z_{i_1}, ..., z_{i_{m-1}}) \\
&\quad - ... - \sum_{1 \le i_1 < i_2 \le m} h_2(z_{i_1}, z_{i_2}) - \sum_{1 \le i \le m} h_1(z_i).
\end{aligned}$$

We call these functions components of a core.

Statement 3. *The U-statistic of a core function h can be written as a sum of U-statistics with degenerate cores,*

$$U(h, Z) = U(h_m, Z) + \binom{m}{1} U(h_{m-1}, Z) + ... + \binom{m}{m-2} U(h_2, Z) + \binom{m}{m-1} U(h_1, Z).$$

Proof. Recall that $\sum_{i \in C_m}$ denotes summation over all $\binom{n}{m}$ combinations of m distinct

elements $\{i_1, \cdots, i_m\}$ from $\{1, \cdots, n\}$.

$$
\begin{aligned}
U(h,Z) &= \frac{1}{n^m} \sum_{i \in C_m} h(Z_{i_1}, ..., Z_{i_m}) \\
&= \frac{1}{n^m} \sum_{i \in C_m} \left(h_m(Z_1, ..., Z_m) + \sum_{1 \le j_1 < ... < j_{m-1} \le m} h_{m-1}(Z_{j_{i_1}}, ..., Z_{i_{j_{m-1}}}) \right. \\
&\quad \left. + ... + \sum_{1 \le j_1 < j_2 \le m} h_2(Z_{i_{j_1}}, Z_{i_{j_2}}) + \sum_{1 \le j \le m} h_1(Z_{i_j}) \right) \\
&= \frac{1}{n^m} \sum_{i \in C_m} h_m(Z_1, ..., Z_m) + \binom{m}{1} \frac{1}{n^{m-1}} \sum_{i \in N^{m-1}} h_{m-1}(Z_{i_1}, ..., Z_{i_{m-1}}) + \\
&\quad + ... + \binom{m}{m-2} \frac{1}{n^2} \sum_{i \in N^2} h_2(Z_{i_1}, Z_{i_2}) + \binom{m}{m-1} \frac{1}{n} \sum_{i \in N} h_1(Z_i) \\
&= U(h_m, Z) + \binom{m}{1} U(h_{m-1}, Z) + ... + \binom{m}{m-2} U(h_2, Z) + \binom{m}{m-1} U(h_1, Z).
\end{aligned}
$$

□

Lemma 25. *Under* $\mathbf{H_0}$, $\forall z \in \mathbf{Z}\ h_1(z) = 0$.

Proof. We use the shorthand notation $k(a,b) \equiv k(x_a, x_b)$, $l(a,b) \equiv l(y_a, y_b)$, such that

$$
h(z_a, z_b, z_c, z_d) = \frac{1}{4!} \sum_{\pi \in S_4} k(\pi_1, \pi_2) \left[l(\pi_1, \pi_2) + l(\pi_3, \pi_4) - 2l(\pi_2, \pi_3) \right].
$$

Let us expand this expression. By using the symmetry of k and l, and writing the

arguments in lexicographical order, we obtain

$$
\begin{aligned}
&h(z_a, z_b, z_c, z_d) = \\
&k(a,b)\,(l(a,b)+l(c,d)-2l(b,c))+k(a,b)\,(l(a,b)+l(c,d)-2l(b,d)) \\
&k(a,c)\,(l(a,c)+l(b,d)-2l(b,c))+k(a,c)\,(l(a,c)+l(b,d)-2l(c,d))+ \\
&k(a,d)\,(l(a,d)+l(b,c)-2l(b,d))+k(a,d)\,(l(a,d)+l(b,c)-2l(c,d))+ \\
&k(a,b)\,(l(a,b)+l(c,d)-2l(a,c))+k(a,b)\,(l(a,b)+l(c,d)-2l(a,d))+ \\
&k(b,c)\,(l(b,c)+l(a,d)-2l(a,c))+k(b,c)\,(l(b,c)+l(a,d)-2l(c,d))+ \\
&k(b,d)\,(l(b,d)+l(a,c)-2l(a,d))+k(b,d)\,(l(b,d)+l(a,c)-2l(c,d))+ \\
&k(a,c)\,(l(a,c)+l(b,d)-2l(a,b))+k(a,c)\,(l(a,c)+l(b,d)-2l(a,d))+ \\
&k(b,c)\,(l(b,c)+l(a,d)-2l(a,b))+k(b,c)\,(l(b,c)+l(a,d)-2l(b,d))+ \\
&k(c,d)\,(l(c,d)+l(a,b)-2l(a,d))+k(c,d)\,(l(c,d)+l(a,b)-2l(b,d))+ \\
&k(a,d)\,(l(a,d)+l(b,c)-2l(a,b))+k(a,d)\,(l(a,d)+l(b,c)-2l(a,c))+ \\
&k(b,d)\,(l(b,d)+l(a,c)-2l(a,b))+k(b,d)\,(l(b,d)+l(a,c)-2l(b,c))+ \\
&k(c,d)\,(l(c,d)+l(a,b)-2l(a,c))+k(c,d)\,(l(c,d)+l(a,b)-2l(b,c))\,.
\end{aligned}
$$

By grouping brackets we obtain

$$
\begin{aligned}
&h(z_a, z_b, z_c, z_d) = \\
&k(a,b)\,(2l(a,b)+2l(c,d)-2l(b,c)-2l(b,d)) \\
&k(a,c)\,(2l(a,c)+2l(b,d)-2l(b,c)-2l(c,d))+ \\
&k(a,d)\,(2l(a,d)+2l(b,c)-2l(b,d)-2l(c,d))+ \\
&k(a,b)\,(2l(a,b)+2l(c,d)-2l(a,c)-2l(a,d))+ \\
&k(b,c)\,(2l(b,c)+2l(a,d)-2l(a,c)-2l(c,d))+ \\
&k(b,d)\,(2l(b,d)+2l(a,c)-2l(a,d)-2l(c,d))+ \\
&k(a,c)\,(2l(a,c)+2l(b,d)-2l(a,b)-2l(a,d))+ \\
&k(b,c)\,(2l(b,c)+2l(a,d)-2l(a,b)-2l(b,d))+ \\
&k(c,d)\,(2l(c,d)+2l(a,b)-2l(a,d)-2l(b,d))+ \\
&k(a,d)\,(2l(a,d)+2l(b,c)-2l(a,b)-2l(a,c))+ \\
&k(b,d)\,(2l(b,d)+2l(a,c)-2l(a,b)-2l(b,c))+ \\
&k(c,d)\,(2l(c,d)+2l(a,b)-2l(a,c)-2l(b,c))\,.
\end{aligned}
$$

Finally we introduce colours to picture grouping of terms that will cancel each other

during integration.

$$
\begin{aligned}
&h(z_a, z_b, z_c, z_d) = \\
&\big[k(a,b)\,(4l(a,b) + 4l(c,d)) + k(a,c)\,(4l(a,c) + 4l(b,d)) + \\
&k(a,d)\,(4l(a,d) + 4l(b,c)) + k(b,c)\,(4l(b,c) + 4l(a,d)) + \\
&k(b,d)\,(4l(b,d) + 4l(a,c)) + k(c,d)\,(4l(c,d) + 4l(a,b))\big] + \\
&\big[k(a,b)(-2l(a,d) - 2l(a,c)) + k(a,b)(-2l(b,d) - 2l(b,c)) + \\
&k(a,c)(-2l(a,d) - 2l(a,b)) + k(a,c)(-2l(c,d) - 2l(b,c)) + \\
&k(a,d)(-2l(a,c) - 2l(a,b)) + k(a,d)(-2l(c,d) - 2l(b,d)) + \\
&k(b,c)(-2l(a,c) - 2l(a,b)) + k(b,c)(-2l(c,d) - 2l(b,d)) + \\
&k(b,d)(-2l(a,b) - 2l(a,d)) + k(b,d)(-2l(b,c) - 2l(c,d)) + \\
&k(c,d)(-2l(a,d) - 2l(a,c)) + k(c,d)(-2l(b,d) - 2l(b,c))\big]
\end{aligned}
\tag{A.16}
$$

We will show that brown terms of equation (A.16) cancel each other. Recall that $h_1(z_1) = Eh(z_1, Z_2^*, Z_3^*, Z_4^*)$. Without loss of generality we may assume that we integrate with respect to all variables but x_a and y_a. Observe that

$$
\begin{aligned}
Ek(x_a, X_b^*) = Ek(x_a, X_c^*) = Ek(x_a, X_d^*) \\
El(y_a, Y_b^*) = El(y_a, Y_c^*) = El(y_a, Y_d^*)
\end{aligned}
$$

Define $q = Ek(x_a, X_b^*)$, $p = El(y_a, Y_b^*)$. Therefore, after integration, the brown terms of the equation can be written as

$$
q4p + q4p + q4p + q(-2p - 2p) + q(-2p - 2p) + q(-2p - 2p) = 0
$$

Similar reasoning shows that red, green and violet terms cancel out. □

Statement 4. *A component of a core function is a canonical core.*

Proof. We will use induction by components' index to show that h_c is degenerate. The expected value of the first component is zero, indeed $Eh_1(Z_1^*) = Eh(Z_1^*, ..., Z_m^*) = 0$. Suppose that for all c' smaller then c degeneracy holds. Using component symmetry it is enough to show that the expected value $Eh_c(z_1, ..., Z_c^*)$ is equal to zero. We can write

$$
\sum_{1 \le i_1 < ... < i_{c'} \le c} h_{c'}(z_{i_1}, ..., z_{i_{c'}}) = \sum_{1 \le i_1 < ... < i_{c'} \le c-1} h_{c'}(z_{i_1}, ..., z_{i_{c'}}) + \sum_{1 \le i_1 < ... < i_{c'-1} < c} h_{c'}(z_{i_1}, ..., z_c).
$$

Now the first sum $\sum_{1 \le i_1 < ... < i_{c'} \le c-1} h_{c'}(z_{i_1}, ..., z_{i_{c'}})$ does not contain term z_c so integration with respect to Z_c^* does not affect it. On the other hand, by induction assumption $E \sum_{1 \le i_1 < ... < i_{c'-1} < c} h_{c'}(z_{i_1}, ..., Z_c^*) = 0$. Obviously $Eg_c(z_1, ..., Z_c^*) =$

$g_{c-1}(z_1,...,z_{c-1})$. Using these observations we obtain

$$\begin{aligned} Eh_c(z_1,...,Z_c^*) = g_{c-1}(z_1,...,z_{c-1}) - \sum_{1\le i_1<...<i_{c-1}\le c-1} h_{c-1}(z_{i_1},...,z_{i_{c-1}}) \\ - ... - \sum_{1\le i_1<i_2\le c-1} h_2(z_{i_1},z_{i_2}) - \sum_{i=1}^{c-1} h_1(z_i) \end{aligned} \tag{A.17}$$

Since the set $\{1 \le i_1 < ... < i_{c-1} \le c-1\}$ contains only one sequence,

$$\begin{aligned} Eh_c(z_1,...,Z_c^*) = -h_{c-1}(z_{i_1},...,z_{i_{c-1}}) + [g_{c-1}(z_1,...,z_{c-1}) \\ - ... - \sum_{1\le i_1<i_2\le c-1} h_2(z_{i_1},z_{i_2}) - \sum_{i=1}^{c-1} h_1(z_i)] = 0. \end{aligned} \tag{A.18}$$

For this nice simplification we have used definition of the component h_{c-1}. □

Lemma 26. *Under* $\mathbf{H_0}$*,*

$$h_2(z_1,z_2) = \frac{1}{6}\tilde{k}(x_1,x_2)\tilde{l}(y_1,y_2)$$

where

$$\begin{aligned} \tilde{k}(x_1,x_2) &= k(x_1,x_2) - Ek(x_1,X_2^*) - Ek(X_1^*,x_2) + Ek(X_1^*,X_2^*), \\ \tilde{l}(y_1,y_2) &= l(y_1,y_2) - El(y_1,Y_2^*) - El(Y_1^*,y_2) + El(Y_1^*,Y_2^*) \end{aligned}$$

Proof. We use that h_2 is canonial, and the exact form of $Eh(z_1,z_2,Z_3^*,Z_4^*)$ from [44], Section A.2, text between equation 12 and 13. □

Corollary 1. *Under* $\mathbf{H_0}$*,* $h_2 = \frac{1}{6}s$.

Proofs - $U(h_4,Z)$ and $U(h_3,Z)$ converge to zero

Lemma 27. *If* $(Z_t)_{t\in\mathbb{N}+}$ *is an absolutely regular process with mixing coefficient decaying faster than* n^{-3} *(*$\beta(n),\theta(n) \le n^{-3}$*), then* $n\cdot U(h_4,Z))$ *and* $n\cdot U(h_3,Z))$ *converge to zero in probability.*

Proof. Let $N := \{1,\cdots,n\}$, and let B be a set of all strictly increasing 4-tuples, $B \subset N^4$. A U-statistic can be expressed as sum over elements of B,

$$n\cdot U(h_4,Z)) = \left[\frac{1}{n^4}\binom{n}{4}^{-1}\right]\frac{1}{n^3}\sum_{\mathbf{b}\in B} h_4(Z_b).$$

If the variance of this random variable goes to zero,

$$\lim_{n\to 0} E\left(\frac{1}{n^3}\sum_{\mathbf{b}\in B} h_4(Z_b)\right)^2 \stackrel{P}{=} 0,$$

then using Chebyshev's inequality we can conclude that it converges to a constant in probability. To show this, we use Lemma 3 from Arcones [3]. We see that the first condition of Theorem 1 from Arcones [3] is met, since h_4 is bounded and the mixing coefficient converges to zero. Therefore, by the fact that h_4 is canonical, we can use Lemma 3 from Arcones [3], which states that

$$E\left(\sum_{\mathbf{b}\in B} h_4(Z_b)\right)^2 \leq Cn^4 M\left(1+\sum_{m=1}^{n-1} m^3\beta(m)^{(p-2)/p}\right)$$

for some $p>2$ and $M = \| h \|_\infty$. Take p such that $\frac{3(p-2)}{p} = 2.5$ and use inequality $\beta(m) \leq m^{-3}$ to obtain

$$\sum_{m=1}^{n-1} m^3\beta(m)^{(p-2)/p} \leq \sum_{m=1}^{n-1}\sqrt{m} = O(n^{1.5}).$$

Therefore

$$\lim_{n\to 0} E\left(\frac{1}{n^3}\sum_{\mathbf{b}\in B} h_4(Z_b)\right)^2 \stackrel{P}{=} \lim_{n\to 0}\frac{n^{5.5}}{n^6} \stackrel{P}{=} 0.$$

We now need to show that $EnU(h_4, Z)$ converges to zero. We will use Lemma 21 with $\delta = 2$, and that $\beta(k)^{\frac{2}{3}} \leq k^{-2}$,

$$\begin{aligned} EnU(h_4,Z) &= \frac{n}{n(n-1)(n-2)(n-3)} E \sum_{1\leq a<b<c<d\leq n} h_4(Z_a,Z_b,Z_c,Z_d) \\ &\leq \frac{n}{n(n-1)(n-2)(n-3)} \sum_{1\leq a<b<c<d\leq n} M^{\frac{1}{3}} \frac{1}{\max(b-a,c-b,d-c)^2}. \end{aligned} \tag{A.19}$$

for some constant M as in Lemma 21. Next

$$\begin{aligned} \sum_{1\leq a<b<c<d\leq n} \frac{1}{\max(b-a,c-b,d-c)^2} &= \sum_{a=1}^{n-3}\sum_{d=a+3}^{n}\sum_{a<b<c<d} \frac{1}{\max(b-a,c-b,d-c)^2} \\ &\leq \sum_{a=1}^{n-3}\sum_{d=a+3}^{n}\frac{3^2}{(d-a)^2} \leq 9\sum_{a=1}^{n-3} 2\zeta(2) \leq Cn. \end{aligned} \tag{A.20}$$

We have used the fact that $\sum_{d=a+3}^{n}\frac{1}{(d-a)^2} \leq 2\zeta(2)$.

The reasoning for $U(h_3, Z)$ is similar. □

Proofs - Borisov and Volodko [15, Theorem 1]

Theorem 11. *Let m be the number of arguments of a symmetric kernel f. Let one of the following two sets of conditions be fulfilled:*

1. *The stationary sequence X_i satisfies θ-mixing and*

   1.1. $\sum_{k=1}^{\infty} \phi(k)^{\frac{1}{2}} < \infty$,

   1.2. $\sup_i E|e_i(X_1)|^2 < \infty$.

2. *The stationary sequence X_i satisfies α-mixing. For some $\epsilon > 0$ and for an even number $c \geq 2$ the following holds:*

   2.1. $\sup_i E|e_i(X_1)|^{2+\epsilon} \leq \infty$,

   2.2. $\sum_{k=1}^{\infty} k^{c-2} \alpha^{\epsilon/(c+\epsilon)}(k) < \infty$

where $e_i(X_1)$ are a basis of $L_2(X, F)$. Then, for any degenerate kernel $f(t_1, ..., t_m) \in L_2(X_m, F_m)$, under conditions

- $\sum_{i_1,...,i_m}^{\infty} |f_{i_1,...,i_m}| < \infty$, *where $f_{i_1,...,i_m}$ are the coefficient of f in $L_2(X_m, F_m)$,*
- *for every collection of pairwise distinct subscripts $(j_1, ..., j_m)$, the distribution of $(X_{j_1}, ..., X_{j_m})$ is absolutely continuous with respect to the distribution of $(X_1^*, ..., X_m^*)$, where X_i* is an independent copy of X_1,*
- $e_0 = 1$ *or* $Ee_i(Z_j) = 0$ *for all i,*

the following assertion holds:

$$n^{\frac{m}{2}} U(f, Z) \to \sum_{i_1,...,i_m}^{\infty} f_{i_1,...,i_m} \prod_{j=1}^{\infty} H_{\nu_j(i_1,...,i_m)}(\tau_j),$$

where τ_j is a centred Gaussian sequence with the covariance matrix

$$E\tau_k\tau_l = Ee_k(X_1)e_l(X_1) + \sum_{j=1}^{\infty} \left[Ee_k(X_1)e_l(X_{j+1}) + Ee_l(X_1)e_k(X_{j+1}) \right],$$

$\nu_j(i_1, ..., i_m) := \sum_{r=1}^{m} \delta_{j,i_r}$, *and $H_k(x)$ are the Hermite polynomials,*

$$H_k(x) = (-1)^k e^{(x^2/2)} \frac{d^k}{dx^k}(e^{-x^2/2})$$

Proofs - Expected value of the eigenfunctions

From the eigenvalue equation $\lambda_i E e_i(z) = E h_2(z, Z_2^*) e_i(Z_2^*)$, h_2 degeneracy, and the independence of Z_1^* and Z_2^*, we conclude that

$$E e_i(Z_1^*) = \frac{1}{\lambda_i} E h_2(Z_1^*, Z_2^*) e_i(Z_2^*) = \frac{1}{\lambda_i} E[e_i(Z_2^*) E(h_2(Z_1^*, Z_2^*) | Z_2^* = z_2)] = \frac{1}{\lambda_i} E[e_i(Z_2^*) \cdot 0] = 0.$$