B-test: A Non-parametric, Low Variance Kernel Two-sample Test

Wojciech Zaremba
Center for Visual Computing
École Centrale Paris
Châtenay-Malabry, France
woj.zaremba@gmail.com

Arthur Gretton
Gatsby Unit
University College London
United Kingdom
arthur.gretton@gmail.com

Matthew Blaschko
Center for Visual Computing
École Centrale Paris
Châtenay-Malabry, France
matthew.blaschko@inria.fr

Abstract

We propose a family of maximum mean discrepancy (MMD) kernel two-sample tests that have low sample complexity and are consistent. The test has a hyperparameter that allows one to control the tradeoff between sample complexity and computational time. Our family of tests, which we denote as B-tests, is both computationally and statistically efficient, combining favorable properties of previously proposed MMD two-sample tests. It does so by better leveraging samples to produce low variance estimates in the finite sample case, while avoiding a quadratic number of kernel evaluations and complex null-hypothesis approximation as would be required by tests relying on one sample U-statistics. The B-test uses a smaller than quadratic number of kernel evaluations and avoids completely the computational burden of complex null-hypothesis approximation, while maintaining consistency and probabilistically conservative thresholds on Type I error. Finally, recent results of combining multiple kernels transfer seamlessly to our hypothesis test, allowing a further increase in discriminative power and decrease in sample complexity.

1 Introduction

Given two samples $\{x_i\}_{i=1}^n$ where $x_i \sim P$ i.i.d., and $\{y_i\}_{i=1}^n$, where $y_i \sim Q$ i.i.d, the two sample problem consists in testing whether to accept the null hypothesis that $P = Q$. In its most general sense, where P and Q might be distributions over high dimensional data or structured objects, this problem has recently been addressed using measures of similarity computed in a reproducing kernel Hilbert space (RKHS). These include the maximum mean discrepancy [9, 6] (of which the energy distance is an example [15, 2, 19]), which is the distance between expected features of P and Q in the RHKS; the kernel Fisher discriminant [11], which is the distance between expected feature maps normalized by the feature space covariance; and density ratio estimates [20]. When used in testing, it is necessary to determine whether the empirical estimate of the relevant similarity measure is sufficiently large as to give the hypothesis $P = Q$ low probability (i.e., below a user-defined threshold α, denoted the test level).

The minimum variance unbiased estimator MMD_u of the maximum mean discrepancy, on the basis of n samples observed from each of P and Q, is a U-statistic, costing $O(n^2)$ to compute. Unfortunately, this statistic is degenerate under the null hypothesis $\mathcal{H}_0$ that $P = Q$, meaning its asymptotic distribution takes the form of an infinite weighted sum of independent χ^2 variables (it is asymptotically Gaussian under the alternative hypothesis $\mathcal{H}_A$ that $P \neq Q$). Two methods for empirically estimating the null distribution in a consistent way have been proposed: the bootstrap [9], and a method requiring an eigendecomposition of the kernel matrices across both samples [7]. Unfortunately, both procedures are computationally demanding: the former costs $O(n^2)$, with a large constant (the MMD must be computed repeatedly over random assignments of the pooled data); the latter costs $O(n^3)$, but with a smaller constant, hence can in practice be faster than the bootstrap. Another approach is to approximate the null distribution by a member of a simpler parametric family (for instance, a Pearson curve approximation), however this has no consistency guarantees.

More recently, an $O(n)$ estimate MMD_l of the maximum mean discrepancy has been proposed [9, Section 6], which is simply a running average over independent pairs of samples from P and Q. While this has much greater variance than the U-statistic, it also has a simpler null distribution: being an average over i.i.d. terms, the central limit theorem gives an asymptotically Gaussian distribution, under both $\mathcal{H}_0$ and $\mathcal{H}_A$. It is shown in [8] that this simple asymptotic distribution makes it easy to optimize the Hodges and Lehmann asymptotic relative efficiency [16] over the family of kernels that define the statistic: in other words, to choose the kernel which gives the lowest Type II error for a given Type I error. Kernel selection for the U-statistic is a much harder question due to the complex form of the null distribution, and remains an open problem.

It appears that MMD_u and MMD_l fall at two extremes of a spectrum: the former has the lowest variance of any n-sample estimator, and should be used in limited data regimes; the latter is the estimator requiring the least computation while still looking at each of the samples, and usually achieves better Type II error than MMD_u at a given computational cost, albeit by looking at much more data (the "limited time, unlimited data" scenario). A major reason MMD_l is faster is that its null distribution is straightforward to compute, since its variance can be calculated at the same cost as the test statistic. A reasonable next step would be to find a compromise between these two extremes: to construct a statistic with a lower variance than MMD_l, while retaining an asymptotically Gaussian null distribution (and hence faster tests than MMD_u). In this paper, we propose a family of such test statistics, where we split the data into blocks of size B, compute the quadratic-time MMD_u on each block, and then average the resulting statistics. We call the resulting tests B-tests. As long as we choose the size B of blocks such that $n/B \to \infty$, we are still guaranteed asymptotic normality by the central limit theorem, and the null distribution can be computed with the same cost as the test statistic. For a given sample size n, however, the power of the test can increase dramatically over the MMD_l test, even for moderate block sizes B, making much better use of the available data with only a small increase in computation. Finally, the kernel learning approach of [8] applies straightforwardly, allowing us to maximize test power over a given kernel family.

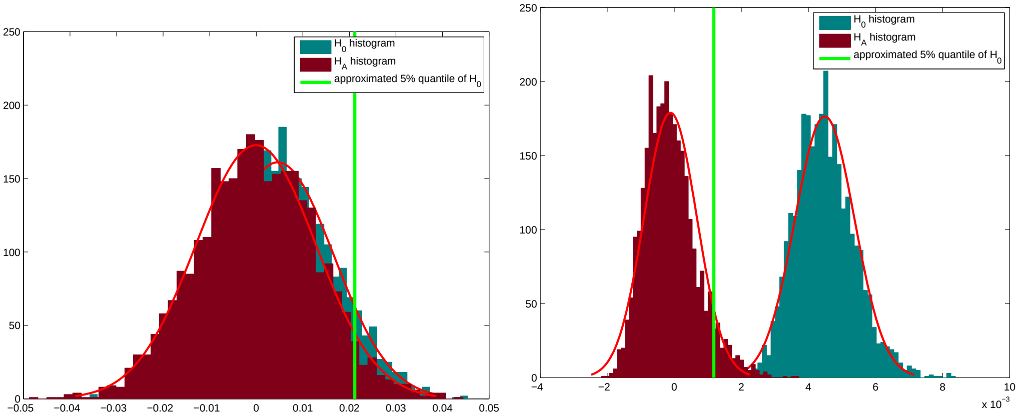
We begin our presentation in Section 2 with a brief overview of the MMD and its empirical estimates. We then propose our block test statistic, provide the relevant asymptotic analysis, and discuss the scaling of B with n. In Section 3, we provide experiments on a number of benchmark datasets, demonstrating the advantages in test power and computational efficiency that the B-test provides.

2 Theory

In this section we describe the mathematical foundations of the B-test. We begin with a brief review of kernel methods, and of the maximum mean discrepancy. We then present our block-based average MMD statistic, and derive its distribution under the $\mathcal{H}_0$ ($P = Q$) and $\mathcal{H}_A$ ($P \neq Q$) hypotheses. The central idea employed in the construction of the B-test is to generate low variance i.i.d. samples by combining multiple correlated but low variance kernel statistics computed over blocks of samples. We show simple sufficient conditions on the block size for consistency of the estimator. Furthermore, we analyze the properties of the finite sample estimate, and propose a consistent strategy for setting the block size as a function of the number of samples.

2.1 Definition and asymptotics of the block-MMD

Let $\mathcal{F}_k$ be an RKHS defined on a topological space $\mathcal{X}$ with reproducing kernel k, and P a Borel probability measure on $\mathcal{X}$. The *mean embedding* of P in $\mathcal{F}_k$, written $\mu_k(p) \in \mathcal{F}_k$ is defined such



(a) $B = 2$. This setting corresponds to the MMD_l statistic [9].

(b) $B = 250$

Figure 1: Empirical distributions under $\mathcal{H}_0$ and $\mathcal{H}_A$ for different regimes of B for the music experiment (Section 3.2). In both plots, the number of samples is fixed at 500. As we vary B, we trade off the approximation of a Gaussian in the finite sample case, as in Theorem 2.3, with the variances of the $\mathcal{H}_0$ and $\mathcal{H}_A$ distributions, as outlined in Section 2.1. In (b) the distribution under $\mathcal{H}_0$ resembles a Gaussian, but due to the skewness of the weighted sum of χ^2 distributions, the empirical distribution results in a conservative threshold (vertical green line). The empirical distribution under $\mathcal{H}_0$ on the right does not pass a Komogorov-Smirnov (KS) normality test [13, 17]. The remaining empirical distributions all pass a KS normality test.

that $E_{x\sim p} f(x) = \langle f, \mu_k(p)\rangle_{\mathcal{F}_k}$ for all $f \in \mathcal{F}_k$, and exists for all Borel probability measures when k is bounded and continuous [3, 9]. The maximum mean discrepancy (MMD) between a Borel probability measure P and a second Borel probability measure Q is the squared RKHS distance between their respective mean embeddings,

$$\eta_k(P,Q) = \|\mu_k(P) - \mu_k(Q)\|^2_{\mathcal{F}_k} = E_{xx'} k(x,x') + E_{yy'} k(y,y') - 2E_{xy} k(x,y), \tag{1}$$

where x' denotes an independent copy of x [10]. By introducing the notation $z = (x,y)$, we may write

$$h(z,z') = k(x,x') + k(y,y') - k(x,y') - k(x',y),$$

and

$$\eta_k(P,Q) = E_{zz'} h_k(z,z'). \tag{2}$$

When the kernel k is characteristic, then $\eta_k\,(P,Q) = 0$ iff $P = Q$ [18].

By analogy with MMD_u, we make use of averages of $h(x,y,x',y')$ to construct our two-sample test. We denote by $\hat{\eta}_k(i)$ the ith empirical estimate MMD_u based on a subsample of size B, where $1 \le i \le \frac{n}{B}$.[1] More precisely,

$$\hat{\eta}_k(i) = \frac{1}{B(B-1)} \sum_{a=(i-1)B+1}^{iB} \sum_{b=(i-1)B+1, b\neq a}^{iB} h(z_a, z_b). \tag{3}$$

The B-test simply estimates the MMD by averaging $\hat{\eta}_k(i)$. Each $\hat{\eta}_k(i)$ under $\mathcal{H}_0$ converges to an infinite sum of weighted χ^2 [7]. Although setting $B = n$ would lead to the lowest variance estimate of the MMD, computing sound thresholds for a given p-value is expensive, involving repeated bootstrap sampling [5, 12], or computing the eigenvalues of a Gram matrix [7].

In contrast, we note that $\hat{\eta}_k(i)_{i=1,\ldots,\frac{n}{B}}$ are i.i.d. variables, and averaging them allows us to apply the central limit theorem in order to estimate p-values from a normal distribution. We denote the average of $\hat{\eta}_k(i)$ by $\hat{\eta}_k$:

$$\hat{\eta}_k = \frac{B}{n} \sum_{i=1}^{\frac{n}{B}} \hat{\eta}_k(i) \tag{4}$$

[1] For notational purposes, we will index samples as though they are presented in a random fixed order.

We would like to apply the central limit theorem to variables $\hat{\eta}_k(i)_{i=1,\dots,\frac{n}{B}}$. It remains for us to derive the distribution of $\hat{\eta}_k$ under $\mathcal{H}_0$ and under $\mathcal{H}_A$.

We rely on the result from [10, Theorem 8] for $\mathcal{H}_A$. According to our notation, for every i, it holds that:

Theorem 2.1 *Assume $0 < \mathbb{E}(h^2) < \infty$, then under $\mathcal{H}_A$, $\hat{\eta}_k$ converges in distribution to a Gaussian according to*

$$B^{\frac{1}{2}}(\hat{\eta}_k(i) - \mathrm{MMD}^2) \xrightarrow{D} \mathcal{N}(0, \sigma_u^2) \tag{5}$$

where $\sigma_u^2 = 4\left(\mathbb{E}_z[(\mathbb{E}_{z'}h(z,z'))^2 - \mathbb{E}_{z,z'}(h(z,z'))]^2\right)$.

This in turn implies that:

$$\hat{\eta}_k(i) \xrightarrow{D} \mathcal{N}(\mathrm{MMD}^2, \frac{\sigma_u^2}{B}) \tag{6}$$

For an average of $\{\hat{\eta}_k(i)\}_{i=1,\dots,\frac{n}{B}}$ based on the central limit theorem we have that under $\mathcal{H}_A$:

$$\hat{\eta}_k \xrightarrow{D} \mathcal{N}\left(\mathrm{MMD}^2, \frac{\sigma_u^2}{B\frac{n}{B}}\right) = \mathcal{N}\left(\mathrm{MMD}^2, \frac{\sigma_u^2}{n}\right). \tag{7}$$

This result shows that asymptotically the distribution of $\mathcal{H}_A$ is independent from the choice of block size, B.

Turning to the null hypothesis, [10, Theorem 8] additionally implies that under $\mathcal{H}_0$ for every i:

Theorem 2.2

$$B\hat{\eta}_k(i) \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l[z_l^2 - 2] \tag{8}$$

where $z_l \sim \mathcal{N}(0,2)^2$ i.i.d, λ_l are the solutions to the eigenvalue equation

$$\int_{\mathcal{X}} \bar{k}(x,x')\psi_l(x)dp(x) = \lambda_l \psi_l(x') \tag{9}$$

and $\bar{k}(x_i,x_j) := k(x_i,x_j) - \mathbb{E}_x k(x_i,x) - \mathbb{E}_x k(x,x_j) + \mathbb{E}_{x,x'}k(x,x')$ is the centered RKHS kernel.

As a consequence, under $\mathcal{H}_0$, $\hat{\eta}_k(i)$ has expected variance $2\frac{\sum_{l=1}^{\infty}\lambda^2}{B^2}$. We will denote this variance by $\frac{C}{B^2}$. The central limit theorem implies under $\mathcal{H}_0$:

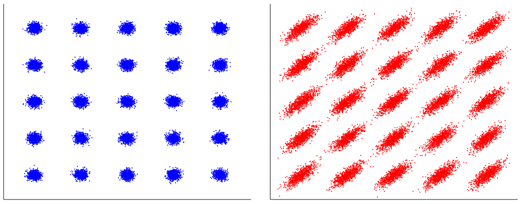
$$\hat{\eta}_k \xrightarrow{D} \mathcal{N}\left(0, \frac{C}{B^2\frac{n}{B}}\right) = \mathcal{N}\left(0, \frac{C}{nB}\right) \tag{10}$$

The asymptotic distributions for $\hat{\eta}_k$ under $\mathcal{H}_0$ and $\mathcal{H}_A$ are Gaussian, and consequently it is easy to calculate hypothesis thresholds. Asymptotically, it is always beneficial to increase B, as the distributions for η under $\mathcal{H}_0$ and $\mathcal{H}_A$ will be better separated. For consistency, it is sufficient to ensure that $\frac{n}{B} \to \infty$.

2.2 Convergence of Moments

In this section, we show that the B-test in the finite sample case gives conservative bounds on the Type I error. We do so by analyzing the convergence of the moments of the distributions.

The central limit theorem implies that the empirical mean of $\{\hat{\eta}_k(i)\}_{i=1,\dots,\frac{n}{B}}$ converges to $\mathbb{E}(\hat{\eta}_k(i))$. Moreover it states that the variance $\{\hat{\eta}_k(i)\}_{i=1,\dots,\frac{n}{B}}$ converges to $\mathbb{E}(\hat{\eta}_k(i))^2 - \mathbb{E}(\hat{\eta}_k(i)^2)$. Finally, all remaining moments tend to zero, where the rate of convergence for the jth moment is of the order



(a) Distribution P

(b) Distribution Q

Figure 2: Synthetic data distributions P and Q. Samples belonging to these classes are difficult to distinguish.

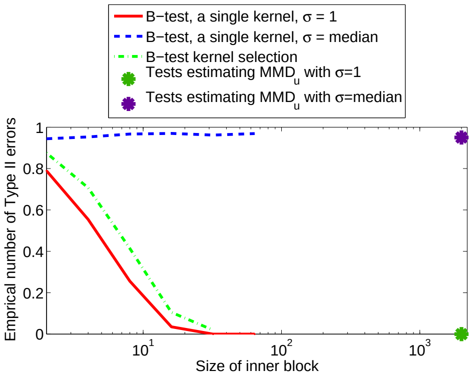


Figure 3: Synthetic experiment in which the empirical number of Type II errors is plotted for a fixed threshold on Type I errors, $\alpha = 5\%$. As B grows, the Type II error drops quickly when the kernel is appropriately chosen. The kernel selection method is described in [8], and closely approximates the best-case performance of a kernel selected a priori to match the data generating process ($\sigma = 1$).

$\left(\frac{n}{B}\right)^{\frac{j+1}{2}}$ [1]. This indicates that the skewness dominates the difference of the distribution from a Gaussian.

Under $\mathcal{H}_A$, which has a Gaussian distribution, thresholds computed from normal distribution tables are unbiased.

Under $\mathcal{H}_0$, from Equation (8) we have that the summands, $\hat{\eta}_k(i)$, converge in distribution to infinite weighted sums of χ^2 distributions. Every unweighted term of this infinite sum is distributed with $\mathcal{N}(0, 2)^2$, which has finite skewness equal to 8. The skewness for the entire sum is finite and positive:

$$C = \sum_{l=1}^{\infty} 8\lambda_l^3, \tag{11}$$

as $\lambda_l \geq 0$ for all l due to the positive definiteness of the kernel k. The skew for the mean of summands, $\hat{\eta}_k(i)$, converges to 0 and is positively biased. We anticipate that in the finite sample case, the bounds from the standard normal table will tend to be conservative for quantiles over 50%, which is the regime of interest (cf. Figure 1). This is consistently demonstrated in our experiments, where Type I error is slightly overestimated at the computed thresholds (Figures 4 and 5).

2.3 Finite Sample Case

In the finite sample case, we apply the Berry-Esséen theorem, which gives conservative bounds on the ℓ_∞ convergence of a series of finite sample random variables to a Gaussian distribution [4].

Theorem 2.3 *Let $X_1, X_2, \ldots, X_n$ be i.i.d. variables. $\mathbb{E}(X_1) = 0$, $\mathbb{E}(X_1^2) = \sigma^2 > 0$, and $\mathbb{E}(|X_1|^3) = \rho < \infty$. Let F_n be a cumulative distribution of $\frac{\sum_{i=1}^n X_i}{\sqrt{n}\sigma}$, and let Φ denote the standard normal distribution. Then for every x*

$$|F_n(x) - \Phi(x)| \leq \frac{C\rho}{\sigma^3\sqrt{n}} \tag{12}$$

where $C < 1$.

This result allows us to ensure fast point-wise convergence of the B-test. We have that $\rho(\hat{\eta}_k) = O(1)$, i.e. it is dependent only on the underlying distributions of the samples and not on the sample size. The number of i.i.d. samples is $\frac{n}{B}$. Based on Theorem 2.3, the point-wise error can be

Method	Kernel parameters	Additional parameters	Minimum number of samples	Computation time (s)	Consistent
B-test	$\sigma = 1$	$B = 2$	26400	0.0012	✓
		$B = 8$	3850	0.0039	✓
		$B = \sqrt{n}$	886	0.0572	✓
	$\sigma =$ median	any B	> 60000		✓
	multiple kernels	$B = 2$	37000	0.0700	✓
		$B = 8$	5400	0.1295	✓
		$B = \sqrt{\frac{n}{2}}$	1700	0.8332	✓
Pearson curves	$\sigma = 1$	$B = n$	186	387.4649	×
Gamma approximation			183	0.2667	×
Gram matrix spectrum			186	407.3447	✓
Bootstrap			190	129.4094	✓
Pearson curves	$\sigma =$ median		> 60000, or 2h per iteration timeout		×
Gamma approximation					×
Gram matrix spectrum					✓
Bootstrap					✓

Table 1: Sample complexity for the distributions described in Figure 2. The fourth column indicates the minimum number of samples necessary to achieve Type I and Type II errors of 5%. The fifth column is the computation time required for 2000 samples, and is not presented for settings that have unsatisfactory sample complexity.

upper bounded by $\frac{O(1)}{O(B^{-1})^{\frac{3}{2}}\sqrt{\frac{n}{B}}} = O(\frac{B^2}{\sqrt{n}})$ under $\mathcal{H}_A$. Under $\mathcal{H}_0$, the error can be bounded by $\frac{O(1)}{O(B^{-2})^{\frac{3}{2}}\sqrt{\frac{n}{B}}} = O(\frac{B^{3.5}}{\sqrt{n}})$.

While the asymptotic results indicate that convergence to an optimal predictor is fastest for larger B, the finite sample results support decreasing the size of B in order to have a sufficient number of samples for application of the central limit theorem. As long as $B \to \infty$ and $\frac{n}{B} \to \infty$, the assumptions of the B-test are fulfilled.

By varying B, we make a fundamental tradeoff in the construction of our two sample test. When B is small, we have many samples, hence the null distribution is close to the asymptotic limit provided by the central limit theorem, and the Type I error is estimated accurately. The disadvantage of a small B is a lower test power for a given sample size. Conversely, if we increase B, we will have a lower variance empirical distribution for $\mathcal{H}_0$, hence higher test power, but we will tend to overestimate the number of Type I errors (Figure 1). A sensible family of heuristics therefore is to set

$$B = n^{\gamma} \tag{13}$$

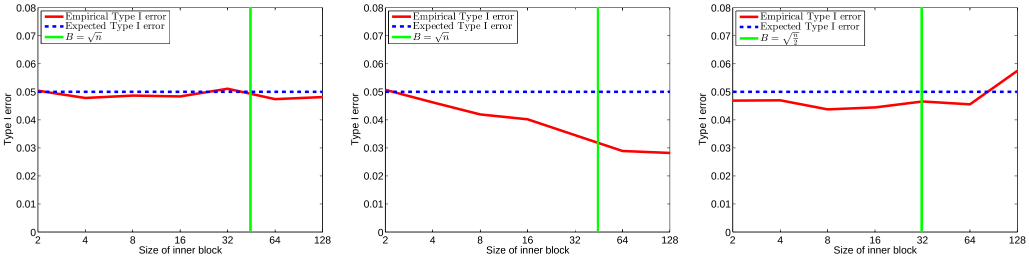
for some $0 < \gamma < 1$. In this setting the number of samples available for application of the central limit theorem will be $n^{(1-\gamma)}$. We note that any value of $\gamma \in (0, 1)$ yields a consistent estimator. We have chosen $\gamma = \frac{1}{2}$ in the experimental results section.

3 Experiments

We have conducted experiments on challenging synthetic and real datasets in order to empirically measure (i) sample complexity, (ii) computation time, and (iii) Type I / Type II errors. We evaluate B-test performance in comparison to MMD_l estimators as well as MMD_u estimators, including Pearson curves, gamma approximation, Gram matrix spectrum, and bootstrap. We note that Pearson curves and gamma approximation are not statistically consistent.

3.1 Synthetic data

Following previous work on kernel hypothesis testing [8], our synthetic setting is a 5×5 grid of 2D Gaussians. We specify two distributions, P and Q. For distribution P each Gaussian has identity covariance matrix, while for distribution Q the covariance is non-spherical. Samples drawn from P and Q are presented in Figure 2. These distributions have proven to be very challenging for existing non-parametric two-sample tests [8].



(a) (b) (c)

Figure 4: Empirical number of Type I errors on the distributions shown in Figure 2 for $\alpha = 5\%$: (a) a single kernel test with $\sigma = 1$, (b) a single kernel test with σ set to the median pairwise distance, and (c) for a multiple kernel test. The experiment was repeated 30000 times. Error bars are not visible at this scale.

We have used several different kernels to compute the hypothesis test. We have used a Gaussian kernel with $\sigma = 1$, which approximately matches the scale of the variance of each Gaussian in mixture P, and can be considered a best-case baseline. Next, we empirically set σ equal to the median pairwise distance over the training data, which is a standard way of choosing the Gaussian kernel bandwidth [14]. Finally, we applied a kernel learning strategy in which the kernel is optimized to reveal if $P \neq Q$ [8]. We have learned a combination of kernels based on half of the samples, and we test on the remaining half. The base kernels were set to Gaussian kernels with bandwidth in the set $\sigma \in \{-15, \ldots, 10\}$.

Finally, for comparison we evaluated four methods estimating one-sample U-statistics (i) Pearson curves, (ii) gamma approximation, (iii) Gram matrix spectrum, and (iv) bootstrap. These methods rely on hyperparameters. For methods using Pearson curves and the Gram matrix spectrum, we set the number of samples to 500 to estimate the null distribution. For bootstrap, we fixed the number of shuffles to 1000. We considered only the setting with $\sigma = 1$ and σ set to the median pairwise distance, as kernel selection is not applicable for any test employing MMD_u [8].

In the first experiment we find a threshold such that Type I error is 5% and we measure empirically the Type II error. We have conducted these experiments on 2000 samples with 1000 repetitions. We consider settings with varying block size, B. Figure 3 presents results for both methods as the block size, B, increases.

In the second experiment, we measure the empirical sample complexity of the various methods. We have fixed for all methods the same Type I error and Type II error of 5%, and we vary only the number of samples. Column four of Table 1 shows the number of samples required in each setting to achieve these error rates. We additionally compare the computational efficiency of the various methods. All experiments were run on a single 2.4 GHz core. The computation time for each method with a fixed sample size of 2000 is presented in column five of Table 1.

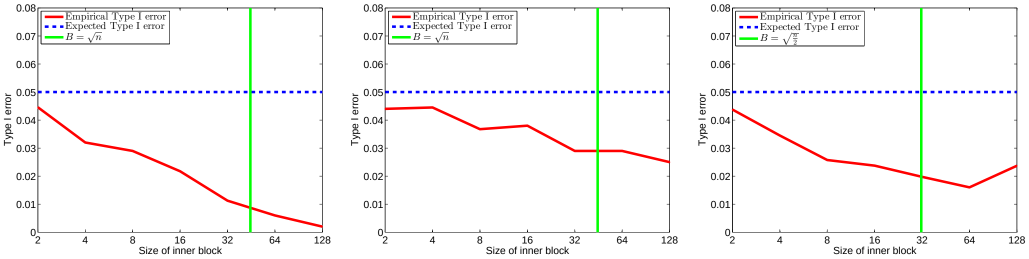
Finally, we evaluate empirical Type I error for $\alpha = 5\%$ and increasing B. Figure 4 displays the empirical Type I error. When the kernel is chosen to be discriminative for the generating distribution ($\sigma = 1$, Figure 4(a)), the number of observed errors closely matches the threshold. However, when the kernel is incorrectly specified (σ set to the median pairwise distance, Figure 4(b)), the expected number of Type I errors does not match the empirical observation. Kernel selection alleviates this problem (Figure 4(c)). We also illustrate the threshold selected when $\gamma = \frac{1}{2}$ in Equation (13). This setting coincides with a block size substantially larger than 2 (MMD_l), and therefore achieves lower Type II errors while retaining good Type I error estimates.

3.2 Musical experiments

In this set of experiments, two amplitude modulated Rammstein songs were compared (*Sehnsucht* vs. *Engel*, from the album *Sehnsucht*). Following the experimental setting in [8, Section 5], samples from P and Q are extracts from AM signals of time duration 8.3×10^{-3} seconds in the original audio. Feature extraction is identical to [8] except that the amplitude scaling parameter was set to 0.3 instead of 0.5. As the feature vector has size 1000 we set the block size $B = \lceil \sqrt{1000} \rceil = 32$. Table 2 summarizes the empirical Type I and Type II errors over 1000 repetitions as well as the

Method	Kernel parameters	Additional parameters	Type I error	Type II error	Computational time (s)
B-test	$\sigma = 1$	$B = 2$	0.038	0.927	0.039
		$B = \sqrt{n}$	0.006	0.597	1.276
	$\sigma =$ median	$B = 2$	0.043	0.786	0.047
		$B = \sqrt{n}$	0.026	0	1.259
	multiple kernels	$B = 2$	0.0481	0.867	0.607
		$B = \sqrt{\frac{n}{2}}$	0.025	0.012	18.285
Gram matrix spectrum	$\sigma = 1$	$B = 2000$	0	0	160.1356
Bootstrap			0.01	0	121.2570
Gram matrix spectrum	$\sigma =$ median		0	0	286.8649
Bootstrap			0.01	0	122.8297

Table 2: A comparison of consistent methods on the music experiment described in Section 3.2. Here computation time is reported for the test achieving the stated error rates.



(a) (b) (c)

Figure 5: Empirical Type I error rate for $\alpha = 5\%$ on the music data (Section 3.2). (a) A single kernel test with $\sigma = 1$, (b) A single kernel test with $\sigma =$ median, and (c) for multiple kernels. Error bars are not visible at this scale. The results broadly follow the trend visible from the synthetic experiments.

average computational times. Figure 5 shows the average empirical number of Type I errors as a function of B, while Figure 1 shows the empirical $\mathcal{H}_0$ and $\mathcal{H}_A$ distributions for different settings of B.

4 Discussion

We have presented experimental results on both a difficult pair of synthetic distributions, and on real-world data from audio recordings. The results show that the B-test has a much better sample complexity than any other method based on MMD_l. Moreover, it is an order of magnitude faster than any consistent method that approximates MMD_u, namely those based on the Gram matrix eigenspectrum and the bootstrap. Methods approximating MMD_u may be infeasible for data with a large number of samples due to their computational complexity. Additionally, the B-test remains statistically consistent, with the best convergence rates achieved for large B. The B-test therefore combines the best features of MMD_l and MMD_u based two-sample tests: consistency, high statistical efficiency, and high computational efficiency.

A number of further interesting experimental trends may be seen in these results. First, as predicted by the analysis of the convergence of moments in Section 2.2, we have experimentally observed that the empirical rate of Type I errors is typically less than the 5% estimated by the threshold based on a Gaussian assumption (Figures 4 and 5). This, however, has not been matched by an increase in Type II errors (Tables 1 and 2), as the increased statistical power of the B-tests has improved overall discrimination rates between hypotheses (cf. Figure 1).

Finally, while Equation (7) implies the size of B does not influence the asymptotic variance under $\mathcal{H}_A$, we nonetheless observe in Figure 1 that the empirical variance of $\mathcal{H}_A$ drops with larger B. Note, however, that for these P and Q and small B, the null and alternative distributions have considerable overlap. Hence for these sample sizes, the variance of the alternative distribution as a function of B is expected to behave more like that of $\mathcal{H}_0$ (cf. Equation (10)). In other words, this behavior is an artifact of the small sample regime given the P and Q being compared, and will vanish in the asymptotic limit.

Source code is available for download from `https://github.com/wojzaremba/btest/`.

References

[1] Bengt Von Bahr. On the convergence of moments in the central limit theorem. *The Annals of Mathematical Statistics*, 36(3):pp. 808–818, 1965.

[2] L. Baringhaus and C. Franz. On a new multivariate two-sample test. *J. Multivariate Anal.*, 88:190–206, 2004.

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

[4] Andrew C Berry. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1):122–136, 1941.

[5] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

[6] M. Fromont, B. Laurent, M. Lerasle, and P. Reynaud-Bouret. Kernels based tests with non-asymptotic bootstrap approaches for two-sample problems. In *COLT*, pages 23.1 – 23.23, 2012.

[7] A Gretton, K Fukumizu, Z Harchaoui, and BK Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems 22*, pages 673–681, Red Hook, NY, USA, 2009. Max-Planck-Gesellschaft, Curran.

[8] A Gretton, B Sriperumbudur, D Sejdinovic, H Strathmann, S Balakrishnan, M Pontil, and K Fukumizu. Optimal kernel choice for large-scale two-sample tests. In P Bartlett, FCN Pereira, CJC. Burges, L Bottou, and KQ Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1214–1222, 2012.

[9] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, March 2012.

[10] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel method for the two-sample-problem. In *NIPS*, pages 513–520, 2006.

[11] Z. Harchaoui, F. Bach, and E. Moulines. Testing for homogeneity with kernel Fisher discriminant analysis. In *NIPS*, pages 609–616. MIT Press, Cambridge, MA, 2008.

[12] Norman Lloyd Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. *Continuous univariate distributions*. Distributions in statistics. Wiley, New York, NY [u.a.], 2. ed. edition, 1994.

[13] Andrey N Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dellIstituto Italiano degli Attuari*, 4(1):83–91, 1933.

[14] B Schölkopf. *Support vector learning*. Oldenbourg, München, Germany, 1997. Zugl.: Berlin, Techn. Univ., Diss., 1997.

[15] D. Sejdinovic, A. Gretton, B. Sriperumbudur, and K. Fukumizu. Hypothesis testing using pairwise distances and associated kernels. In *ICML*, 2012.

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

[17] Nickolay Smirnov. Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, 19(2):279–281, 1948.

[18] B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

[19] G. Székely and M. Rizzo. Testing for equal distributions in high dimension. *InterStat*, (5), November 2004.

[20] M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural Computation*, 25(5):1324–1370, 2013.