In Proc. of NIPS 2006, Vancouver, CA

A Kernel Method for the Two-Sample-Problem

Arthur Gretton
MPI for Biological Cybernetics
Tübingen, Germany
arthur@tuebingen.mpg.de

Karsten M. Borgwardt
Ludwig-Maximilians-Univ.
Munich, Germany
kb@dbs.ifi.lmu.de

Malte Rasch
Graz Univ. of Technology,
Graz, Austria
malte.rasch@igi.tu-graz.ac.at

Bernhard Schölkopf
MPI for Biological Cybernetics
Tübingen, Germany
bs@tuebingen.mpg.de

Alexander J. Smola
NICTA, ANU
Canberra, Australia
Alex.Smola@anu.edu.au

Abstract

We propose two statistical tests to determine if two samples are from different distributions. Our test statistic is in both cases the distance between the means of the two samples mapped into a reproducing kernel Hilbert space (RKHS). The first test is based on a large deviation bound for the test statistic, while the second is based on the asymptotic distribution of this statistic. The test statistic can be computed in $O(m^2)$ time. We apply our approach to a variety of problems, including attribute matching for databases using the Hungarian marriage method, where our test performs strongly. We also demonstrate excellent performance when comparing distributions over graphs, for which no alternative tests currently exist.

1 Introduction

We address the problem of comparing samples from two probability distributions, by proposing a statistical test of the hypothesis that these distributions are different (this is called the two-sample or homogeneity problem). This test has application in a variety of areas. In bioinformatics, it is of interest to compare microarray data from different tissue types, either to determine whether two subtypes of cancer may be treated as statistically indistinguishable from a diagnosis perspective, or to detect differences in healthy and cancerous tissue. In database attribute matching, it is desirable to merge databases containing multiple fields, where it is not known in advance which fields correspond: the fields are matched by maximising the similarity in the distributions of their entries.

In this study, we propose to test whether distributions p and q are different on the basis of samples drawn from each of them, by finding a smooth function which is large on the points drawn from p, and small (as negative as possible) on the points from q. We use as our test statistic the difference between the mean function values on the two samples; when this is large, the samples are likely from different distributions. We call this statistic the Maximum Mean Discrepancy (MMD).

Clearly the quality of MMD as a statistic depends heavily on the class $\mathcal{F}$ of smooth functions that define it. On one hand, $\mathcal{F}$ must be "rich enough" so that the population MMD vanishes if and only if $p = q$. On the other hand, for the test to be consistent, $\mathcal{F}$ needs to be "restrictive" enough for the empirical estimate of MMD to converge quickly to its expectation as the sample size increases. We shall use the unit balls in universal reproducing kernel Hilbert spaces [22] as our function class, since these will be shown to satisfy both of the foregoing properties. On a more practical note, MMD is cheap to compute: given m points sampled from p and n from q, the cost is $O(m + n)^2$ time.

We define two non-parametric statistical tests based on MMD. The first, which uses distribution-independent uniform convergence bounds, provides finite sample guarantees of test performance, at the expense of being conservative in detecting differences between p and q. The second test is

based on the asymptotic distribution of MMD, and is in practice more sensitive to differences in distribution at small sample sizes. These results build on our earlier work in [6] on MMD for the two sample problem, which addresses only the second kind of test. In addition, the present approach employs a more accurate approximation to the asymptotic distribution of the test statistic.

We begin our presentation in Section 2 with a formal definition of the MMD, and a proof that the population MMD is zero if and only if $p = q$ when $\mathcal{F}$ is the unit ball of a universal RKHS. We also give an overview of hypothesis testing as it applies to the two-sample problem, and review previous approaches. In Section 3, we provide a bound on the deviation between the population and empirical MMD, as a function of the Rademacher averages of $\mathcal{F}$ with respect to p and q. This leads to a first hypothesis test. We take a different approach in Section 4, where we use the asymptotic distribution of an unbiased estimate of the squared MMD as the basis for a second test. Finally, in Section 5, we demonstrate the performance of our method on problems from neuroscience, bioinformatics, and attribute matching using the Hungarian marriage approach. Our approach performs well on high dimensional data with low sample size; in addition, we are able to successfully apply our test to graph data, for which no alternative tests exist. Proofs and further details are provided in [13].

2 The Two-Sample-Problem

Our goal is to formulate a statistical test that answers the following question:

Problem 1 *Let p and q be distributions defined on a domain $\mathcal{X}$. Given observations $X := \{x_1, \ldots, x_m\}$ and $Y := \{y_1, \ldots, y_n\}$, drawn independently and identically distributed (i.i.d.) from p and q respectively, is $p \neq q$?*

To start with, we wish to determine a criterion that, in the population setting, takes on a unique and distinctive value only when $p = q$. It will be defined based on [10, Lemma 9.3.2].

Lemma 1 *Let $(\mathcal{X}, d)$ be a separable metric space, and let p, q be two Borel probability measures defined on $\mathcal{X}$. Then $p = q$ if and only if $\mathbf{E}_p(f(x)) = \mathbf{E}_q(f(x))$ for all $f \in C(\mathcal{X})$, where $C(\mathcal{X})$ is the space of continuous bounded functions on $\mathcal{X}$.*

Although $C(\mathcal{X})$ in principle allows us to identify $p = q$ uniquely, it is not practical to work with such a rich function class in the finite sample setting. We thus define a more general class of statistic, for as yet unspecified function classes $\mathcal{F}$, to measure the discrepancy between p and q, as proposed in [11].

Definition 2 *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$ and let p, q, X, Y be defined as above. Then we define the maximum mean discrepancy (MMD) and its empirical estimate as*

$$\mathrm{MMD}\left[\mathcal{F}, p, q\right] := \sup_{f \in \mathcal{F}} \left(\mathbf{E}_{x\sim p}[f(x)] - \mathbf{E}_{y\sim q}[f(y)]\right), \tag{1}$$

$$\mathrm{MMD}\left[\mathcal{F}, X, Y\right] := \sup_{f \in \mathcal{F}} \left(\frac{1}{m}\sum_{i=1}^{m} f(x_i) - \frac{1}{n}\sum_{i=1}^{n} f(y_i)\right). \tag{2}$$

We must now identify a function class that is rich enough to uniquely establish whether $p = q$, yet restrictive enough to provide useful finite sample estimates (the latter property will be established in subsequent sections). To this end, we select $\mathcal{F}$ to be the unit ball in a universal RKHS $\mathcal{H}$ [22]; we will henceforth use $\mathcal{F}$ only to denote this function class. With the additional restriction that $\mathcal{X}$ be compact, a universal RKHS is dense in $C(\mathcal{X})$ with respect to the L_∞ norm. It is shown in [22] that Gaussian and Laplace kernels are universal.

Theorem 3 *Let $\mathcal{F}$ be a unit ball in a universal RKHS $\mathcal{H}$, defined on the compact metric space $\mathcal{X}$, with associated kernel $k(\cdot, \cdot)$. Then $\mathrm{MMD}\left[\mathcal{F}, p, q\right] = 0$ if and only if $p = q$.*

This theorem is proved in [13]. We next express the MMD in a more easily computable form. This is simplified by the fact that in an RKHS, function evaluations can be written $f(x) = \langle \phi(x), f \rangle$,

where $\phi(x) = k(x, .)$. Denote by $\mu[p] := \mathbf{E}_{x\sim p(x)}[\phi(x)]$ the expectation of $\phi(x)$ (assuming that it exists).[1] Since $\mathbf{E}_p[f(x)] = \langle \mu[p], f\rangle$, we may rewrite

$$\mathrm{MMD}[\mathcal{F}, p, q] = \sup_{\|f\|_{\mathcal{H}}\leq 1} \langle \mu[p] - \mu[q], f\rangle = \|\mu[p] - \mu[q]\|_{\mathcal{H}}\,. \tag{3}$$

Using $\mu[X] := \frac{1}{m}\sum_{i=1}^{m}\phi(x_i)$ and $k(x, x') = \langle \phi(x), \phi(x')\rangle$, an empirical estimate of MMD is

$$\mathrm{MMD}\left[\mathcal{F}, X, Y\right] = \left[\frac{1}{m^2}\sum_{i,j=1}^{m} k(x_i, x_j) - \frac{2}{mn}\sum_{i,j=1}^{m,n} k(x_i, y_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i, y_j)\right]^{\frac{1}{2}}. \tag{4}$$

Eq. (4) provides us with a test statistic for $p \neq q$. We shall see in Section 3 that this estimate is biased, although it is straightforward to upper bound the bias (we give an unbiased estimate, and an associated test, in Section 4). Intuitively we expect $\mathrm{MMD}[\mathcal{F}, X, Y]$ to be small if $p = q$, and the quantity to be large if the distributions are far apart. Note that it costs $O((m+n)^2)$ time to compute the statistic.

Overview of Statistical Hypothesis Testing, and of Previous Approaches Having defined our test statistic, we briefly describe the framework of statistical hypothesis testing as it applies in the present context, following [9, Chapter 8]. Given i.i.d. samples $X \sim p$ of size m and $Y \sim q$ of size n, the statistical test, $\mathcal{T}(X, Y) : \mathcal{X}^m \times \mathcal{X}^n \mapsto \{0, 1\}$ is used to distinguish between the null hypothesis $\mathcal{H}_0 : p = q$ and the alternative hypothesis $\mathcal{H}_1 : p \neq q$. This is achieved by comparing the test statistic $\mathrm{MMD}[\mathcal{F}, X, Y]$ with a particular threshold: if the threshold is exceeded, then the test rejects the null hypothesis (bearing in mind that a zero population MMD indicates $p = q$). The acceptance region of the test is thus defined as any real number below the threshold. Since the test is based on finite samples, it is possible that an incorrect answer will be returned: we define the Type I error as the probability of rejecting $p = q$ based on the observed sample, despite the null hypothesis being true. Conversely, the Type II error is the probability of accepting $p = q$ despite the underlying distributions being different. The level α of a test is an upper bound on the Type I error: this is a design parameter of the test, and is used to set the threshold to which we compare the test statistic (finding the test threshold for a given α is the topic of Sections 3 and 4). A consistent test achieves a level α, and a Type II error of zero, in the large sample limit. We will see that both of the tests proposed in this paper are consistent.

We next give a brief overview of previous approaches to the two sample problem for multivariate data. Since our later experimental comparison is with respect to certain of these methods, we give abbreviated algorithm names in italics where appropriate: these should be used as a key to the tables in Section 5. We provide further details in [13]. A generalisation of the Wald-Wolfowitz runs test to the multivariate domain was proposed and analysed in [12, 17] (*Wolf*), which involves counting the number of edges in the minimum spanning tree over the aggregated data that connect points in X to points in Y. The computational cost of this method using Kruskal's algorithm is $O((m + n)^2 \log(m + n))$, although more modern methods improve on the $\log(m + n)$ term. Two possible generalisations of the Kolmogorov-Smirnov test to the multivariate case were studied in [4, 12]. The approach of Friedman and Rafsky (*Smir*) in this case again requires a minimal spanning tree, and has a similar cost to their multivariate runs test. A more recent multivariate test was proposed in [20], which is based on the minimum distance non-bipartite matching over the aggregate data, at cost $O((m + n)^3)$. Another recent test was proposed in [15] (*Hall*): for each point from p, it requires computing the closest points in the aggregated data, and counting how many of these are from q (the procedure is repeated for each point from q with respect to points from p). The test statistic is costly to compute; [15] consider only tens of points in their experiments.

Yet another approach is to use some distance (e.g. L_1 or L_2) between Parzen window estimates of the densities as a test statistic [1, 3], based on the asymptotic distribution of this distance given $p = q$. When the L_2 norm is used, the test statistic is related to those we present here, although it is arrived at from a different perspective (see [13]: the Parzen window interpretation is problematic, since the resulting two-sample test is not consistent when the kernel size is decreased fast enough to ensure a consistent Parzen window estimate). The L_1 approach of [3] (*Biau*) requires the space to be partitioned into a grid of bins, which becomes difficult or impossible for high dimensional problems. Hence we use this test only for low-dimensional problems in our experiments.

[1] A sufficient condition for this is $\|\mu[p]\|^2_{\mathcal{H}} < \infty$, which is rearranged as $\mathbf{E}_p[k(x, x')] < \infty$, where x and x' are independent random variables drawn according to p.

3 A Test based on Uniform Convergence Bounds

In this section, we establish two properties of the MMD. First, we show that regardless of whether or not $p = q$, the empirical MMD converges in probability at rate $1/\sqrt{m+n}$ to its population value. This establishes the consistency of statistical tests based on MMD. Second, we give probabilistic bounds for large deviations of the empirical MMD in the case $p = q$. These bounds lead directly to a threshold for our first hypothesis test.

We begin our discussion of the convergence of $\mathrm{MMD}[\mathcal{F}, X, Y]$ to $\mathrm{MMD}[\mathcal{F}, p, q]$.

Theorem 4 *Let p, q, X, Y be defined as in Problem 1, and assume $|k(x, y)| \leq K$. Then*

$$\Pr\left\{|\mathrm{MMD}[\mathcal{F}, X, Y] - \mathrm{MMD}[\mathcal{F}, p, q]| > 2\left((K/m)^{1/2} + (K/n)^{1/2}\right) + \epsilon\right\} \leq 2\exp\left(-\frac{\epsilon^2 mn}{2K(m+n)}\right).$$

Our next goal is to refine this result in a way that allows us to define a test threshold under the null hypothesis $p = q$. Under this circumstance, the constants in the exponent are slightly improved.

Theorem 5 *Under the conditions of Theorem 4 where additionally $p = q$ and $m = n$,*

$$\mathrm{MMD}[\mathcal{F}, X, Y] > \underbrace{m^{-\frac{1}{2}}\sqrt{2\mathbf{E}_p\left[k(x,x) - k(x,x')\right]}}_{B_1(\mathcal{F},p)} + \epsilon > \underbrace{2(K/m)^{1/2}}_{B_2(\mathcal{F},p)} + \epsilon,$$

both with probability less than $\exp\left(-\frac{\epsilon^2 m}{4K}\right)$ (see [13] for the proof).

In this theorem, we illustrate two possible bounds $B_1(\mathcal{F}, p)$ and $B_2(\mathcal{F}, p)$ on the bias in the empirical estimate (4). The first inequality is interesting inasmuch as it provides a link between the bias bound $B_1(\mathcal{F}, p)$ and kernel size (for instance, if we were to use a Gaussian kernel with large σ, then $k(x, x)$ and $k(x, x')$ would likely be close, and the bias small). In the context of testing, however, we would need to provide an additional bound to show convergence of an empirical estimate of $B_1(\mathcal{F}, p)$ to its population equivalent. Thus, in the following test for $p = q$ based on Theorem 5, we use $B_2(\mathcal{F}, p)$ to bound the bias.

Lemma 6 *A hypothesis test of level α for the null hypothesis $p = q$ (equivalently $\mathrm{MMD}[\mathcal{F}, p, q] = 0$) has the acceptance region $\mathrm{MMD}[\mathcal{F}, X, Y] < 2\sqrt{K/m}\left(1 + \sqrt{\log \alpha^{-1}}\right)$.*

We emphasise that Theorem 4 guarantees the consistency of the test, and that the Type II error probability decreases to zero at rate $1/\sqrt{m}$ (assuming $m = n$). To put this convergence rate in perspective, consider a test of whether two normal distributions have equal means, given they have unknown but equal variance [9, Exercise 8.41]. In this case, the test statistic has a Student-t distribution with $n + m - 2$ degrees of freedom, and its error probability converges at the same rate as our test.

4 An Unbiased Test Based on the Asymptotic Distribution of the U-Statistic

We now propose a second test, which is based on the asymptotic distribution of an unbiased estimate of MMD^2. We begin by defining this test statistic.

Lemma 7 *Given x and x' independent random variables with distribution p, and y and y' independent random variables with distribution q, the population MMD^2 is*

$$\mathrm{MMD}^2\left[\mathcal{F}, p, q\right] = \mathbf{E}_{x,x'\sim p}\left[k(x, x')\right] - 2\mathbf{E}_{x\sim p, y\sim q}\left[k(x, y)\right] + \mathbf{E}_{y,y'\sim q}\left[k(y, y')\right] \tag{5}$$

(see [13] for details). Let $Z := (z_1, \ldots, z_m)$ be m i.i.d. random variables, where $z_i := (x_i, y_i)$ (i.e. we assume $m = n$). An unbiased empirical estimate of MMD^2 is

$$\mathrm{MMD}_u^2\left[\mathcal{F}, X, Y\right] = \frac{1}{(m)(m-1)}\sum_{i\neq j}^{m} h(z_i, z_j), \tag{6}$$

which is a one-sample U-statistic with $h(z_i, z_j) := k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i)$.

The empirical statistic is an unbiased estimate of MMD^2, although it does not have minimum variance (the minimum variance estimate is almost identical: see [21, Section 5.1.4]). We remark that these quantities can easily be linked with a simple kernel between probability measures: (5) is a special case of the Hilbertian metric [16, Eq. (4)] with the associated kernel $\mathfrak{K}(p,q) = \mathbf{E}_{p,q} k(x,y)$ [16, Theorem 4]. The asymptotic distribution of this test statistic under $\mathcal{H}_1$ is given by [21, Section 5.5.1], and the distribution under $\mathcal{H}_0$ is computed based on [21, Section 5.5.2] and [1, Appendix]; see [13] for details.

Theorem 8 *We assume* $\mathbf{E}\left(h^2\right) < \infty$. *Under* $\mathcal{H}_1$, MMD_u^2 *converges in distribution (defined e.g. in [14, Section 7.2]) to a Gaussian according to*

$$m^{\frac{1}{2}}\left(\mathrm{MMD}_u^2 - \mathrm{MMD}^2\left[\mathcal{F}, p, q\right]\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma_u^2\right),$$

where $\sigma_u^2 = 4\left(\mathbf{E}_z\left[\left(\mathbf{E}_{z'} h(z,z')\right)^2\right] - \left[\mathbf{E}_{z,z'}\left(h(z,z')\right)\right]^2\right)$, *uniformly at rate* $1/\sqrt{m}$ *[21, Theorem B, p. 193]. Under* $\mathcal{H}_0$, *the U-statistic is degenerate, meaning* $\mathbf{E}_{z'} h(z,z') = 0$. *In this case,* MMD_u^2 *converges in distribution according to*

$$m\mathrm{MMD}_u^2 \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l \left[z_l^2 - 2\right], \tag{7}$$

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

$$\int_{\mathcal{X}} \tilde{k}(x,x')\psi_i(x) dp(x) = \lambda_i \psi_i(x'),$$

and $\tilde{k}(x_i,x_j) := k(x_i,x_j) - \mathbf{E}_x k(x_i,x) - \mathbf{E}_x k(x,x_j) + \mathbf{E}_{x,x'} k(x,x')$ *is the centred RKHS kernel.*

Our goal is to determine whether the empirical test statistic MMD_u^2 is so large as to be outside the $1-\alpha$ quantile of the null distribution in (7) (consistency of the resulting test is guaranteed by the form of the distribution under $\mathcal{H}_1$). One way to estimate this quantile is using the bootstrap [2] on the aggregated data. Alternatively, we may approximate the null distribution by fitting Pearson curves to its first four moments [18, Section 18.8]. Taking advantage of the degeneracy of the U-statistic, we obtain (see [13])

$$\mathbf{E}\left(\left[\mathrm{MMD}_u^2\right]^2\right) = \frac{2}{m(m-1)} \mathbf{E}_{z,z'}\left[h^2(z,z')\right]$$

and

$$\mathbf{E}\left(\left[\mathrm{MMD}_u^2\right]^3\right) = \frac{8(m-2)}{m^2(m-1)^2} \mathbf{E}_{z,z'}\left[h(z,z')\mathbf{E}_{z''}\left(h(z,z'')h(z',z'')\right)\right] + O(m^{-4}). \tag{8}$$

The fourth moment $\mathbf{E}\left(\left[\mathrm{MMD}_u^2\right]^4\right)$ is not computed, since it is both very small ($O(m^{-4})$) and expensive to calculate ($O(m^4)$). Instead, we replace the kurtosis with its lower bound $\mathrm{kurt}\left(\mathrm{MMD}_u^2\right) \geq \left(\mathrm{skew}\left(\mathrm{MMD}_u^2\right)\right)^2 + 1$.

5 Experiments

We conducted distribution comparisons using our MMD-based tests on datasets from three real-world domains: database applications, bioinformatics, and neurobiology. We investigated the uniform convergence approach (MMD), the asymptotic approach with bootstrap (MMD_u^2 B), and the asymptotic approach with moment matching to Pearson curves (MMD_u^2 M). We also compared against several alternatives from the literature (where applicable): the multivariate t-test, the Friedman-Rafsky Kolmogorov-Smirnov generalisation (*Smir*), the Friedman-Rafsky Wald-Wolfowitz generalisation (*Wolf*), the Biau-Györfi test (*Biau*), and the Hall-Tajvidi test (*Hall*). Note that we do not apply the Biau-Györfi test to high-dimensional problems (see end of Section 2), and that MMD is the only method applicable to structured data such as graphs.

An important issue in the practical application of the MMD-based tests is the selection of the kernel parameters. We illustrate this with a Gaussian RBF kernel, where we must choose the kernel width

σ (we use this kernel for univariate and multivariate data, but not for graphs). The empirical MMD is zero both for kernel size $\sigma = 0$ (where the aggregate Gram matrix over X and Y is a unit matrix), and also approaches zero as $\sigma \to \infty$ (where the aggregate Gram matrix becomes uniformly constant). We set σ to be the median distance between points in the aggregate sample, as a compromise between these two extremes: this remains a heuristic, however, and the optimum choice of kernel size is an ongoing area of research.

Data integration As a first application of MMD, we performed distribution testing for data integration: the objective is to aggregate two datasets into a single sample, with the understanding that both original samples are generated from the same distribution. Clearly, it is important to check this last condition before proceeding, or an analysis could detect patterns in the new dataset that are caused by combining the two different source distributions, and not by real-world phenomena. We chose several real-world settings to perform this task: we compared microarray data from normal and tumor tissues (Health status), microarray data from different subtypes of cancer (Subtype), and local field potential (LFP) electrode recordings from the Macaque primary visual cortex (V1) with and without spike events (Neural Data I and II). In all cases, the two data sets have different statistical properties, but the detection of these differences is made difficult by the high data dimensionality.

We applied our tests to these datasets in the following fashion. Given two datasets A and B, we either chose one sample from A and the other from B (*attributes = different*); or both samples from either A or B (*attributes = same*). We then repeated this process up to 1200 times. Results are reported in Table 1. Our asymptotic tests perform better than all competitors besides *Wolf*: in the latter case, we have greater Type II error for one neural dataset, lower Type II error on the Health Status data (which has very high dimension and low sample size), and identical (error-free) performance on the remaining examples. We note that the Type I error of the bootstrap test on the Subtype dataset is far from its design value of 0.05, indicating that the Pearson curves provide a better threshold estimate for these low sample sizes. For the remaining datasets, the Type I errors of the Pearson and Bootstrap approximations are close. Thus, for larger datasets, the bootstrap is to be preferred, since it costs $O(m^2)$, compared with a cost of $O(m^3)$ for Pearson (due to the cost of computing (8)). Finally, the uniform convergence-based test is too conservative, finding differences in distribution only for the data with largest sample size.

Dataset	Attr.	MMD	MMD_u^2 B	MMD_u^2 M	t-test	Wolf	Smir	Hall
Neural Data I	Same	100.0	96.5	96.5	100.0	97.0	95.0	96.0
	Different	50.0	**0.0**	**0.0**	42.0	**0.0**	10.0	49.0
Neural Data II	Same	100.0	94.6	95.2	100.0	95.0	94.5	96.0
	Different	100.0	3.3	3.4	100.0	**0.8**	31.8	5.9
Health status	Same	100.0	95.5	94.4	100.0	94.7	96.1	95.6
	Different	100.0	1.0	**0.8**	100.0	2.8	44.0	35.7
Subtype	Same	100.0	99.1	96.4	100.0	94.6	97.3	96.5
	Different	100.0	**0.0**	**0.0**	100.0	**0.0**	28.4	0.2

Table 1: Distribution testing for data integration on multivariate data. Numbers indicate the percentage of repetitions for which the null hypothesis (p=q) was accepted, given $\alpha = 0.05$. Sample size (dimension; repetitions of experiment): Neural I 4000 (63; 100) ; Neural II 1000 (100; 1200); Health Status 25 (12,600; 1000); Subtype 25 (2,118; 1000).

Attribute matching Our second series of experiments addresses automatic attribute matching. Given two databases, we want to detect corresponding attributes in the schemas of these databases, based on their data-content (as a simple example, two databases might have respective fields Wage and Salary, which are assumed to be observed via a subsampling of a particular population, and we wish to automatically determine that both Wage and Salary denote to the same underlying attribute). We use a two-sample test on pairs of attributes from two databases to find corresponding pairs.[2] This procedure is also called *table matching* for tables from different databases. We performed attribute matching as follows: first, the dataset D was split into two halves A and B. Each of the n attributes

[2]Note that corresponding attributes may have different distributions in real-world databases. Hence, schema matching cannot solely rely on distribution testing. Advanced approaches to schema matching using MMD as one key statistical test are a topic of current research.

in A (and B, resp.) was then represented by its instances in A (resp. B). We then tested all pairs of attributes from A and from B against each other, to find the optimal assignment of attributes $A_1, \ldots, A_n$ from A to attributes $B_1, \ldots, B_n$ from B. We assumed that A and B contain the same number of attributes.

As a naive approach, one could assume that any possible pair of attributes might correspond, and thus that every attribute of A needs to be tested against all the attributes of B to find the optimal match. We report results for this naive approach, aggregated over all pairs of possible attribute matches, in Table 2. We used three datasets: the census income dataset from the UCI KDD archive (CNUM), the protein homology dataset from the 2004 KDD Cup (BIO) [8], and the forest dataset from the UCI ML archive [5]. For the final dataset, we performed univariate matching of attributes (FOREST) and multivariate matching of tables (FOREST10D) from two different databases, where each table represents one type of forest. Both our asymptotic MMD_u^2-based tests perform as well as or better than the alternatives, notably for CNUM, where the advantage of MMD_u^2 is large. Unlike in Table 1, the next best alternatives are not consistently the same across all data: e.g. in BIO they are *Wolf* or *Hall*, whereas in FOREST they are *Smir*, *Biau*, or the t-test. Thus, MMD_u^2 appears to perform more consistently across the multiple datasets. The Friedman-Rafsky tests do not always return a Type I error close to the design parameter: for instance, *Wolf* has a Type I error of 9.7% on the BIO dataset (on these data, MMD_u^2 has the joint best Type II error without compromising the designed Type I performance). Finally, our uniform convergence approach performs much better than in Table 1, although surprisingly it fails to detect differences in FOREST10D.

A more principled approach to attribute matching is also possible. Assume that $\phi(A) = (\phi_1(A_1), \phi_2(A_2), ..., \phi_n(A_n))$: in other words, the kernel decomposes into kernels on the individual attributes of A (and also decomposes this way on the attributes of B). In this case, MMD^2 can be written $\sum_{i=1}^{n} \|\mu_i(A_i) - \mu_i(B_i)\|^2$, where we sum over the MMD terms on each of the attributes. Our goal of optimally assigning attributes from B to attributes of A via MMD is equivalent to finding the optimal permutation π of attributes of B that minimizes $\sum_{i=1}^{n} \|\mu_i(A_i) - \mu_i(B_{\pi(i)})\|^2$. If we define $C_{ij} = \|\mu_i(A_i) - \mu_i(B_j)\|^2$, then this is the same as minimizing the sum over $C_{i,\pi(i)}$. This is the linear assignment problem, which costs $O(n^3)$ time using the Hungarian method [19].

Dataset	Attr.	MMD	MMD_u^2 B	MMD_u^2 M	t-test	Wolf	Smir	Hall	Biau
BIO	Same	100.0	93.8	94.8	95.2	90.3	95.8	95.3	99.3
	Different	20.0	**17.2**	17.6	36.2	**17.2**	18.6	17.9	42.1
FOREST	Same	100.0	96.4	96.0	97.4	94.6	99.8	95.5	100.0
	Different	4.9	**0.0**	**0.0**	0.2	3.8	**0.0**	50.1	**0.0**
CNUM	Same	100.0	94.5	93.8	94.0	98.4	97.5	91.2	98.5
	Different	15.2	2.7	**2.5**	19.17	22.5	11.6	79.1	50.5
FOREST10D	Same	100.0	94.0	94.0	100.0	93.5	96.5	97.0	100.0
	Different	100.0	**0.0**	**0.0**	**0.0**	**0.0**	1.0	72.0	100.0

Table 2: Naive attribute matching on univariate (BIO, FOREST, CNUM) and multivariate data (FOREST10D). Numbers indicate the percentage of accepted null hypothesis (p=q) pooled over attributes. $\alpha = 0.05$. Sample size (dimension; attributes; repetitions of experiment): BIO 377 (1; 6; 100); FOREST 538 (1; 10; 100); CNUM 386 (1; 13; 100); FOREST10D 1000 (10; 2; 100).

We tested this 'Hungarian approach' to attribute matching via MMD_u^2 B on three univariate datasets (BIO, CNUM, FOREST) and for table matching on a fourth (FOREST10D). To study MMD_u^2 B on structured data, we obtained two datasets of protein graphs (PROTEINS and ENZYMES) and used the graph kernel for proteins from [7] for table matching via the Hungarian method (the other tests were not applicable to this graph data). The challenge here is to match tables representing one functional class of proteins (or enzymes) from dataset A to the corresponding tables (functional classes) in B. Results are shown in Table 3. Besides on the BIO dataset, MMD_u^2 B made no errors.

6 Summary and Discussion

We have established two simple multivariate tests for comparing two distributions p and q. The test statistics are based on the maximum deviation of the expectation of a function evaluated on each of the random variables, taken over a sufficiently rich function class. We do not require density

Dataset	Data type	No. attributes	Sample size	No. repetitions	% correct matches
BIO	univariate	6	377	100	90.0
CNUM	univariate	13	386	100	99.8
FOREST	univariate	10	538	100	100.0
FOREST10D	multivariate	2	1000	100	100.0
ENZYME	structured	6	50	50	100.0
PROTEINS	structured	2	200	50	100.0

Table 3: Hungarian Method for attribute matching via MMD_u^2 B on univariate (BIO, CNUM, FOREST), multivariate (FOREST10D), and structured data (ENZYMES, PROTEINS) ($\alpha = 0.05$; '% correct matches' is the percentage of the correct attribute matches detected over all repetitions).

estimates as an intermediate step. Our method either outperforms competing methods, or is close to the best performing alternative. Finally, our test was successfully used to compare distributions on graphs, for which it is currently the only option.

Acknowledgements: The authors thank Matthias Hein for helpful discussions, Patrick Warnat (DKFZ, Heidelberg) for providing the microarray datasets, and Nikos Logothetis for providing the neural datasets. NICTA is funded through the Australian Government's *Backing Australia's Ability* initiative, in part through the ARC. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

References

[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] M. Arcones and E. Giné. On the bootstrap of u and v statistics. *The Annals of Statistics*, 20(2):655–674, 1992.

[3] G. Biau and L. Gyorfi. On the asymptotic properties of a nonparametric l_1-test statistic of homogeneity. *IEEE Transactions on Information Theory*, 51(11):3965–3973, 2005.

[4] P. Bickel. A distribution free version of the Smirnov two sample test in the p-variate case. *The Annals of Mathematical Statistics*, 40(1):1–23, 1969.

[5] C. L. Blake and C. J. Merz. UCI repository of machine learning databases, 1998.

[6] K. M. Borgwardt, A. Gretton, M.J. Rasch, H.P. Kriegel, B. Schölkopf, and A.J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. In *ISMB*, 2006.

[7] K. M. Borgwardt, C. S. Ong, S. Schonauer, S. V. N. Vishwanathan, A. J. Smola, and H. P. Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21(Suppl 1):i47–i56, Jun 2005.

[8] R. Caruana and T. Joachims. Kdd cup. http://kodiak.cs.cornell.edu/kddcup/index.html, 2004.

[9] G. Casella and R. Berger. *Statistical Inference*. Duxbury, Pacific Grove, CA, 2nd edition, 2002.

[10] R. M. Dudley. *Real analysis and probability*. Cambridge University Press, Cambridge, UK, 2002.

[11] R. Fortet and E. Mourier. Convergence de la réparation empirique vers la réparation théorique. *Ann. Scient. École Norm. Sup.*, 70:266–285, 1953.

[12] J. Friedman and L. Rafsky. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, 7(4):697–717, 1979.

[13] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two sample problem. Technical Report 157, MPI for Biological Cybernetics, 2007.

[14] G. R. Grimmet and D. R. Stirzaker. *Probability and Random Processes*. Oxford University Press, Oxford, third edition, 2001.

[15] P. Hall and N. Tajvidi. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):359–374, 2002.

[16] M. Hein, T.N. Lal, and O. Bousquet. Hilbertian metrics on probability measures and their application in svm's. In *Proceedings of the 26th DAGM Symposium*, pages 270–277, Berlin, 2004. Springer.

[17] N. Henze and M. Penrose. On the multivariate runs test. *The Annals of Statistics*, 27(1):290–298, 1999.

[18] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions. Volume 1 (Second Edition)*. John Wiley and Sons, 1994.

[19] H.W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

[20] P. Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society B*, 67(4):515–530, 2005.

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

[22] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2002.