A Kernel Approach to Comparing Distributions

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 describe a technique for comparing distributions without the need for density estimation as an intermediate step. Our approach relies on mapping the distributions into a Reproducing Kernel Hilbert Space. We apply this technique to construct a two-sample test, which is used for determining whether two sets of observations arise from the same distribution. We use this test in attribute matching for databases using the Hungarian marriage method, where it performs strongly. We also demonstrate excellent performance when comparing distributions over graphs, for which no alternative tests currently exist.

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 (Steinwart 2002) 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 develop a non-parametric statistical test for the two-sample problem, based on the asymptotic distribution of an unbiased empirical estimate of the MMD. This result builds on our earlier work in (Borgwardt *et al.* 2006), although the present approach employs a more accurate approximation to the asymptotic distribution of the test statistic; the test described here was originally presented in (Gretton *et al.* 2007). We demonstrate the good performance of our test on problems from bioinformatics and attribute matching using the Hungarian marriage approach. In addition, we are able to successfully apply our test to graph data, for which no alternative tests exist. Matlab software for the test may be downloaded from `http : //www.kyb.mpg.de/bs/people/arthur/mmd.htm`

The Two-Sample-Problem

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, we wish to test whether $p \neq q$.

To start with, we must 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 (Dudley 2002, 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

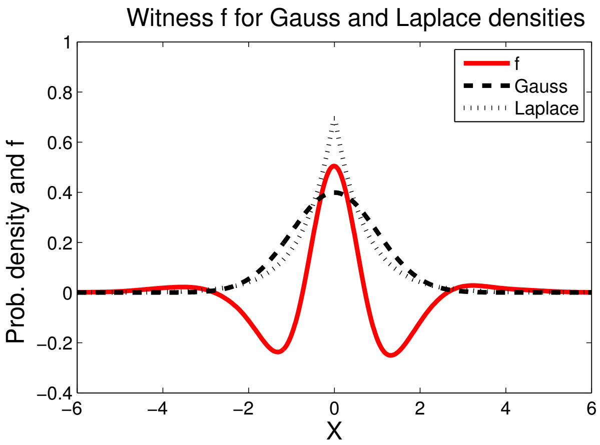

Figure 1: Illustration of the function maximizing the mean discrepancy in the case where a Gaussian is being compared with a Laplace distribution. Both distributions have zero mean and unit variance. The function f that witnesses the MMD has been scaled for plotting purposes, and was computed empirically on the basis of 2×10^4 samples, using a Gaussian kernel with $\sigma = 0.5$.

$\mathcal{F}$, to measure the discrepancy between p and q, as proposed in (Fortet & Mourier 1953).

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

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

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 $\mathcal{H}$ (Steinwart 2002); 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 (Steinwart 2002) 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 $p = q$.*

See (Gretton *et al.* 2007) for more detail. We plot the witness function f from Definition 2 in Figure 1, when p is Gaussian and q is Laplace, for a Gaussian RKHS kernel.

We next express the MMD in a more easily computable form.

Lemma 4 *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*

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

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

$$\mathrm{MMD}_u^2\left[\mathcal{F}, X, Y\right] = \frac{1}{(m)(m-1)} \sum_{i \neq j}^{m} h(z_i, z_j), \quad (3)$$

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

Proof [Eq. (2) in Lemma 4] In an RKHS, function evaluations can be written $f(x) = \langle \phi(x), f \rangle$, where $\phi(x) = k(x, .)$. Denote by $\mu_p := \mathbf{E}_{x\sim p(x)}\left[\phi(x)\right]$ the expectation of $\phi(x)$ (assuming that it exists),[1] and note that $\mathbf{E}_p[f(x)] = \langle \mu_p, f \rangle$. Then

$$\begin{aligned}
&\mathrm{MMD}^2[\mathcal{F}, p, q] \\
&= \left(\sup_{\|f\|_{\mathcal{H}} \le 1} \mathbf{E}_p\left[f(x)\right] - \mathbf{E}_q\left[f(y)\right] \right)^2 \\
&= \left(\sup_{\|f\|_{\mathcal{H}} \le 1} \mathbf{E}_p\left[\langle \phi(x), f\rangle_{\mathcal{H}}\right] - \mathbf{E}_q\left[\langle \phi(y), f\rangle_{\mathcal{H}}\right] \right)^2 \\
&= \left(\sup_{\|f\|_{\mathcal{H}} \le 1} \langle \mu_p - \mu_q, f\rangle_{\mathcal{H}} \right)^2 \\
&= \|\mu_p - \mu_q\|_{\mathcal{H}}^2 \\
&= \langle \mu_p, \mu_p\rangle_{\mathcal{H}} + \langle \mu_q, \mu_q\rangle_{\mathcal{H}} - 2\langle \mu_p, \mu_q\rangle_{\mathcal{H}} \\
&= \mathbf{E}_p\langle \phi(x), \phi(x')\rangle_{\mathcal{H}} + \mathbf{E}_q\langle \phi(y), \phi(y')\rangle_{\mathcal{H}} \\
&\quad - 2\mathbf{E}_{p,q}\langle \phi(x), \phi(y)\rangle_{\mathcal{H}},
\end{aligned}$$

where x' is a random variable independent of x with distribution p, and y' is a random variable independent of y with distribution q. The proof is completed by applying $\langle \phi(x), \phi(x')\rangle_{\mathcal{H}} = k(x, x')$. ■

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 (Serfling 1980, Section 5.1.4)). Intuitively we expect $\mathrm{MMD}_u^2[\mathcal{F}, X, Y]$ 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. We remark that these quantities can easily be linked with a simple kernel between probability measures: (2) is a special case of the Hilbertian metric (Hein, Lal, & Bousquet 2004, Eq. (4)) with the associated kernel $\mathfrak{K}(p,q) = \mathbf{E}_{p,q}k(x,y)$ (Hein, Lal, & Bousquet 2004, Theorem 4).

Having defined our test statistic, we briefly describe the framework of statistical hypothesis testing as it applies in the

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

present context, following (Casella & Berger 2002, 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. A consistent test achieves a level α, and a Type II error of zero, in the large sample limit. We will see that the test proposed in this paper is consistent.

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

We now propose a statistical test of whether $p \neq q$, which is based on the asymptotic distribution of MMD_u^2. This distribution under $\mathcal{H}_1$ is given by (Serfling 1980, Section 5.5.1), and the distribution under $\mathcal{H}_0$ is computed based on (Serfling 1980, Section 5.5.2) and (Anderson, Hall, & Titterington 1994, Appendix); see (Gretton *et al.* 2007) for details.

Theorem 5 *We assume* $\mathbf{E}\left(h^2\right) < \infty$. *Under* $\mathcal{H}_1$, MMD_u^2 *converges in distribution 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 $\sigma_u^2 = 4\left(\mathbf{E}_z\left[(\mathbf{E}_{z'}h(z,z'))^2\right] - \left[\mathbf{E}_{z,z'}(h(z,z'))\right]^2\right)$, *uniformly at rate* $1/\sqrt{m}$ *(Serfling 1980, 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*

$$m\mathrm{MMD}_u^2 \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l \left[z_l^2 - 2\right], \tag{4}$$

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.*

We illustrate the MMD density under both the null and alternative hypotheses by approximating it empirically for both $p = q$ and $p \neq q$. Results are plotted in Figure 2.

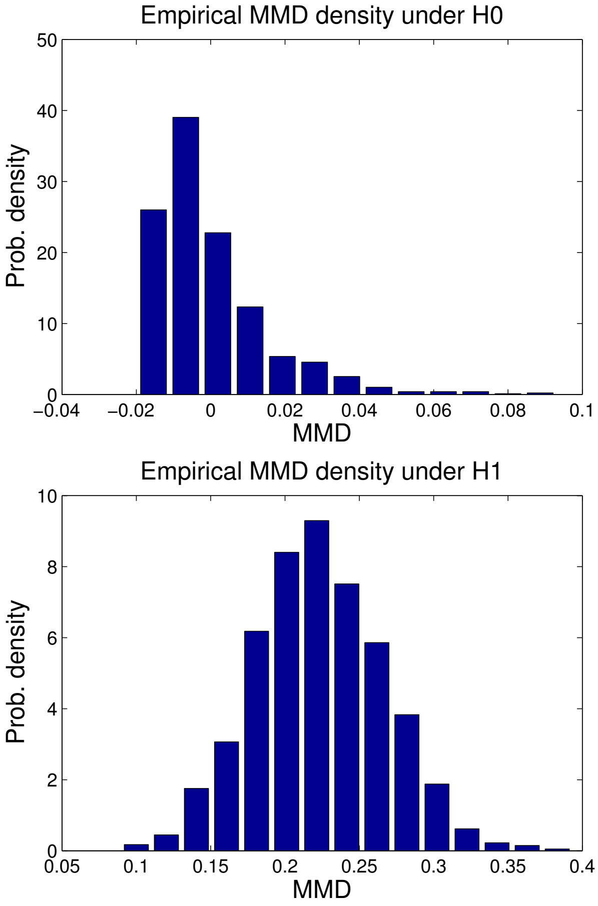

Figure 2: **Top:** Empirical distribution of the MMD under $\mathcal{H}_0$, with p and q both Gaussians with unit standard deviation, using 50 samples from each. **Bottom:** Empirical distribution of the MMD under $\mathcal{H}_1$, with p a Laplace distribution with unit standard deviation, and q a Laplace distribution with standard deviation $3\sqrt{2}$, using 100 samples from each. In both cases, the histograms were obtained by computing 2000 independent instances of the MMD.

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 (4) (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 (Arcones & Giné 1992) on the aggregated data. Alternatively, we may approximate the null distribution by fitting Pearson curves to its first four moments (Johnson, Kotz, & Balakrishnan 1994, Section 18.8).

Experiments

We conducted distribution comparisons using our MMD-based tests on datasets from bioinformatics and database applications. We applied tests based on both moment matching to Pearson curves (MMD_u^2 M) and the bootstrap (MMD_u^2 B). For our kernel, we used a Gaussian with σ set to the me-

Dataset	Attr.	MMD_u^2 B	MMD_u^2 M	t-test	Wolf	Smir	Hall	Biau
BIO	Same	93.8	94.8	95.2	90.3	95.8	95.3	99.3
	Different	**17.2**	17.6	36.2	**17.2**	18.6	17.9	42.1
FOREST	Same	96.4	96.0	97.4	94.6	99.8	95.5	100.0
	Different	**0.0**	**0.0**	0.2	3.8	**0.0**	50.1	**0.0**
CNUM	Same	94.5	93.8	94.0	98.4	97.5	91.2	98.5
	Different	2.7	**2.5**	19.17	22.5	11.6	79.1	50.5
FOR10D	Same	94.0	94.0	100.0	93.5	96.5	97.0	100.0
	Different	**0.0**	**0.0**	**0.0**	**0.0**	1.0	72.0	100.0

Table 1: Attribute matching on univariate (BIO, FOREST, CNUM) and multivariate data (FOR10D). 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); FOR10D 1000 (10; 2; 100).

dian distance between points in the aggregate sample, besides on the graph data, where we used the graph kernel for proteins from (Borgwardt *et al.* 2005). We also compared against several alternatives from the literature (see (Gretton *et al.* 2007) for descriptions): 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 the Biau-Györfi test does not apply to very high-dimensional problems (since it requires partitioning of the space into a grid), and that MMD is the only method applicable to structured data such as graphs.

Our experiments address 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 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 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 contained 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 1. We used three datasets: the census income dataset from the UCI KDD archive (CNUM), the protein homology dataset from the 2004 KDD Cup (BIO) (Caruana & Joachims 2004), and the forest dataset from the UCI ML archive (Blake & Merz 1998). For the final dataset, we performed univariate matching of attributes (FOREST) and multivariate matching of tables (FOR10D) 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. 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).

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 (Kuhn 1955).

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 (FOR10D). To study MMD_u^2 B on structured data, we obtained two datasets of protein graphs (PROT and ENZYM) and used the graph kernel for proteins from (Borgwardt *et al.* 2005) for table matching (the other tests were not applicable to this graph data). The challenge on these graph datasets is to match tables representing one functional class of proteins (or enzymes) from dataset A to the corresponding tables (func-

[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.

tional classes) in B. The graph kernel we apply is a special instance of Haussler's convolution kernels (Haussler 1999), and counts common substructures in two graphs, for instance walks, shortest paths, subtrees, or cyclic patterns. To keep graph kernel computation efficient, only substructures that can be computed in polynomial runtime are considered. Results for attribute matching are shown in Table 2. Besides on the BIO dataset, MMD_u^2 B made almost no errors.

Dataset	Type	# attr.	# samp.	Rep.	% corr.
BIO	univ.	6	377	100	90.0
CNUM	univ.	13	386	100	99.8
FOREST	univ.	10	538	100	100.0
FOR10D	multiv.	2	1000	100	100.0
ENZYM	struct.	6	50	50	100.0
PROT	struct.	2	200	50	100.0

Table 2: Hungarian Method for attribute matching via MMD_u^2 B on univariate (BIO, CNUM, FOREST), multivariate (FOR10D), and structured data (ENZYM, PROT) ($\alpha = 0.05$; '# attr.' is number of attributes, '# samp.' is the sample size, 'Rep.' is the number of repetitions, and '% corr.' is the percentage of correct attribute matches detected over all repetitions).

Summary and Discussion

We have established a simple statistical test for comparing two distributions p and q. The test statistic is 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 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. We remark that other applications can be built on the Hilbert space representation of distributions. For instance, it is possible to correct for covariate shift, where we wish to perform supervised learning when the test distribution differs from the training distribution (Huang *et al.* 2007). This is achieved by by matching feature space means of the test and training distributions through a training sample reweighting.

Acknowledgments NICTA is funded through the Australian Government's *Baking Australia's Ability* initiative, in part through the Australian Research Council. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

References

Anderson, N.; Hall, P.; and Titterington, D. 1994. 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.

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

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

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

Borgwardt, K. M.; Gretton, A.; Rasch, M. J.; Kriegel, H.-P.; Schölkopf, B.; and Smola, A. J. 2006. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* 22(14):e49–e57.

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

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

Dudley, R. M. 2002. *Real analysis and probability*. Cambridge, UK: Cambridge University Press.

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

Gretton, A.; Borgwardt, K.; Rasch, M.; Schölkopf, B.; and Smola, A. J. 2007. A kernel method for the two-sample-problem. In Schölkopf, B.; Platt, J.; and Hofmann, T., eds., *Advances in Neural Information Processing Systems*, volume 19. The MIT Press, Cambridge, MA.

Haussler, D. 1999. Convolutional kernels on discrete structures. Technical Report UCSC-CRL-99-10, Computer Science Department, UC Santa Cruz.

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

Huang, J.; Smola, A.; Gretton, A.; Borgwardt, K.; and Schölkopf, B. 2007. Correcting sample selection bias by unlabeled data. In Schölkopf, B.; Platt, J.; and Hofmann, T., eds., *Advances in Neural Information Processing Systems*, volume 19. The MIT Press, Cambridge, MA.

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

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

Serfling, R. 1980. *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Steinwart, I. 2002. On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.* 2:67–93.