A Hilbert Space Embedding for Distributions

Alex Smola[1], Arthur Gretton[2], Le Song[1], and Bernhard Schölkopf[2]

[1] National ICT Australia, North Road, Canberra 0200 ACT, Australia,
alex.smola@nicta.com.au,lesong@it.usyd.edu.au
[2] MPI for Biological Cybernetics, Spemannstr. 38, 72076 Tübingen, Germany,
{arthur,bernhard.schoelkopf}@tuebingen.mpg.de

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. Applications of this technique can be found in two-sample tests, which are used for determining whether two sets of observations arise from the same distribution, covariate shift correction, local learning, measures of independence, and density estimation.

Kernel methods are widely used in supervised learning [1, 2, 3, 4], however they are much less established in the areas of testing, estimation, and analysis of probability distributions, where information theoretic approaches [5, 6] have long been dominant. Recent examples include [7] in the context of construction of graphical models, [8] in the context of feature extraction, and [9] in the context of independent component analysis. These methods have by and large a common issue: to compute quantities such as the mutual information, entropy, or Kullback-Leibler divergence, we require sophisticated space partitioning and/or bias correction strategies [10, 9].

In this paper we give an overview of methods which are able to compute distances between distributions *without* the need for intermediate density estimation. Moreover, these techniques allow algorithm designers to specify which properties of a distribution are most relevant to their problems. We are optimistic that our embedding approach to distribution representation and analysis will lead to the development of algorithms which are simpler and more effective than entropy-based methods in a broad range of applications.

We begin our presentation in Section 1 with an overview of reproducing kernel Hilbert spaces (RKHSs), and a description of how probability distributions can be represented as elements in an RKHS. In Section 2, we show how these representations may be used to address a variety of problems, including homogeneity testing (Section 2.1), covariate shift correction (Section 2.2), independence measurement (Section 2.3), feature extraction (Section 2.4), and density estimation (Section 2.5).

1 Hilbert Space Embedding

1.1 Preliminaries

In the following we denote by $\mathcal{X}$ the domain of observations, and let $\mathbf{P}_x$ be a probability measure on $\mathcal{X}$. Whenever needed, $\mathcal{Y}$ will denote a second domain, with its own probability measure $\mathbf{P}_y$. A joint probability measure on $\mathcal{X} \times \mathcal{Y}$ will be denoted by $\mathbf{P}_{x,y}$. We will assume all measures are Borel measures, and the domains are compact.

We next introduce a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ of functions on $\mathcal{X}$ with kernel k (the analogous definitions hold for a corresponding RKHS $\mathcal{G}$ with kernel l on $\mathcal{Y}$). This is defined as follows: $\mathcal{H}$ is a Hilbert space of functions $\mathcal{X} \to \mathbb{R}$ with dot product $\langle \cdot, \cdot \rangle$, satisfying the reproducing property:

$$\langle f(\cdot), k(x, \cdot) \rangle = f(x) \tag{1a}$$

$$\text{and consequently } \langle k(x, \cdot), k(x', \cdot) \rangle = k(x, x'). \tag{1b}$$

This means we can view the linear map from a function f on $\mathcal{X}$ to its value at x as an inner product. The evaluation functional is then given by $k(x, \cdot)$, i.e. the kernel function. Popular kernel functions on $\mathbb{R}^n$ include the polynomial kernel $k(x, x') = \langle x, x' \rangle^d$, the Gaussian RBF kernel $k(x, x') = \exp\left(-\lambda \left\|x - x'\right\|^2\right)$, and the Laplace kernel $k(x, x') = \exp\left(-\lambda \left\|x - x'\right\|\right)$. Good kernel functions have been defined on texts, graphs, time series, dynamical systems, images, and structured objects. For recent reviews see [11, 12, 13].

An alternative view, which will come in handy when designing algorithms is that of a *feature map*. That is, we will consider maps $x \to \phi(x)$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle$ and likewise $f(x) = \langle w, \phi(x) \rangle$, where w is a suitably chosen "weight vector" (w can have infinite dimension, e.g. in the case of a Gaussian kernel).

Many kernels are universal in the sense of [14]. That is, their Hilbert spaces $\mathcal{H}$ are dense in the space of continuous bounded functions $C_0(\mathcal{X})$ on the compact domain $\mathcal{X}$. For instance, the Gaussian and Laplacian RBF kernels share this property. This is important since many results regarding distributions are stated with respect to $C_0(\mathcal{X})$ and we would like to translate them into results on Hilbert spaces.

1.2 Embedding

At the heart of our approach are the following two mappings:

$$\mu[\mathbf{P}_x] := \mathbf{E}_x \left[k(x, \cdot)\right] \tag{2a}$$

$$\mu[X] := \frac{1}{m} \sum_{i=1}^{m} k(x_i, \cdot). \tag{2b}$$

Here $X = \{x_1, \ldots, x_m\}$ is assumed to be drawn independently and identically distributed from $\mathbf{P}_x$. If the (sufficient) condition $\mathbf{E}_x \left[k(x, x)\right] < \infty$ is satisfied, then $\mu[\mathbf{P}_x]$ is an element of the Hilbert space (as is, in any case, $\mu[X]$). By virtue of the reproducing property of $\mathcal{H}$,

$$\langle \mu[\mathbf{P}_x], f \rangle = \mathbf{E}_x \left[f(x) \right] \text{ and } \langle \mu[X], f \rangle = \frac{1}{m} \sum_{i=1}^{m} f(x_i).$$

That is, we can compute expectations and empirical means with respect to $\mathbf{P}_x$ and X, respectively, by taking inner products with the means in the RKHS, $\mu[\mathbf{P}_x]$ and $\mu[X]$. The representations $\mu[\mathbf{P}_x]$ and $\mu[X]$ are attractive for the following reasons [15, 16]:

Theorem 1. *If the kernel k is universal, then the mean map $\mu : \mathbf{P}_x \to \mu[\mathbf{P}_x]$ is injective.*

Moreover, we have fast convergence of $\mu[X]$ to $\mu[\mathbf{P}_x]$ as shown in [17, Theorem 15]. Denote by $R_m(\mathcal{H}, \mathbf{P}_x)$ the Rademacher average [18] associated with $\mathbf{P}_x$ and $\mathcal{H}$ via

$$R_m(\mathcal{H}, \mathbf{P}_x) = \frac{1}{m} \mathbf{E}_{x_1,\ldots,x_m} \mathbf{E}_{\sigma_1,\ldots,\sigma_m} \left[\sup_{\|f\|_{\mathcal{H}} \leq 1} \left| \sum_{i=1}^{m} \sigma_i f(x_i) \right| \right]. \tag{3}$$

$R_m(\mathcal{H}, \mathbf{P}_x)$ can be used to measure the deviation between empirical means and expectations [17].

Theorem 2. *Assume that $\|f\|_\infty \leq R$ for all $f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} \leq 1$. Then with probability at least $1 - \delta$, $\|\mu[\mathbf{P}_x] - \mu[X]\| \leq 2R_m(\mathcal{H}, \mathbf{P}_x) + R\sqrt{-m^{-1} \log(\delta)}$*

This ensures that $\mu[X]$ is a good proxy for $\mu[\mathbf{P}_x]$, provided the Rademacher average is well behaved.

Theorem 1 tells us that $\mu[\mathbf{P}_x]$ can be used to define distances between distributions $\mathbf{P}_x$ and $\mathbf{P}_y$, simply by letting $D(\mathbf{P}_x, \mathbf{P}_y) := \|\mu[\mathbf{P}_x] - \mu[\mathbf{P}_y]\|$. Theorem 2 tells us that we do not need to have access to actual distributions in order to compute $D(\mathbf{P}_x, \mathbf{P}_y)$ approximately — as long as $R_m(\mathcal{H}, \mathbf{P}_x) = O(m^{-\frac{1}{2}})$, a finite sample from the distributions will yield error of $O(m^{-\frac{1}{2}})$. See [18] for an analysis of the behavior of $R_m(\mathcal{H}, \mathbf{P}_x)$ when $\mathcal{H}$ is an RKHS.

This allows us to use $D(\mathbf{P}_x, \mathbf{P}_y)$ as a drop-in replacement wherever information theoretic quantities would have been used instead, e.g. for the purpose of determining whether two sets of observations have been drawn from the same distribution. Note that there is a strong connection between Theorem 2 and uniform convergence results commonly used in Statistical Learning Theory [19, 16]. This is captured in the theorem below:

Theorem 3. *Let $\mathcal{F}$ be the unit ball in the reproducing kernel Hilbert space $\mathcal{H}$. Then the deviation between empirical means and expectations for any $f \in \mathcal{F}$ is bounded:*

$$\sup_{f \in \mathcal{F}} \left| \mathbf{E}_x \left[f(x) \right] - \frac{1}{m} \sum_{i=1}^{m} f(x_i) \right| = \left\| \mu[\mathbf{P}_x] - \mu[X] \right\|.$$

Bounding the probability that this deviation exceeds some threshold ϵ is one of the key problems of statistical learning theory. See [16] for details. This means that we have at our disposition a large range of tools typically used to assess the quality of estimators. The key difference is that while those bounds are typically used to bound the deviation between empirical and expected means under the assumption that the data *are* drawn from the same distribution, we will use the bounds in Section 2.1 to test whether this assumption is actually true, and in Sections 2.2 and 2.5 to motivate strategies for approximating particular distributions.

This is analogous to what is commonly done in the univariate case: the Glivenko-Cantelli lemma allows one to bound deviations between empirical and expected means for functions of bounded variation, as generalized by the work of Vapnik and Chervonenkis [20, 21]. However, the Glivenko-Cantelli lemma also leads to the Kolmogorov-Smirnov statistic comparing distributions by comparing their cumulative distribution functions. Moreover, corresponding q-q plots can be used as a diagnostic tool to identify where differences occur.

1.3 A View from the Marginal Polytope

The space of all probability distributions $\mathcal{P}$ is a convex set. Hence, the image $\mathcal{M} := \mu[\mathcal{P}]$ of $\mathcal{P}$ under the linear map μ also needs to be convex. This set is commonly referred to as the marginal polytope. Such mappings have become a standard tool in deriving efficient algorithms for approximate inference in graphical models and exponential families [22, 23].

We are interested in the properties of $\mu[\mathbf{P}]$ in the case where $\mathbf{P}$ satisfies the conditional independence relations specified by an undirected graphical model. In [24], it is shown for this case that the sufficient statistics decompose along the maximal cliques of the conditional independence graph.

More formally, denote by $\mathcal{C}$ set of maximal cliques of the graph G and let x_c be the restriction of $x \in \mathcal{X}$ to the variables on clique $c \in \mathcal{C}$. Moreover, let k_c be universal kernels in the sense of [14] acting on the restrictions of $\mathcal{X}$ on clique $c \in \mathcal{C}$. In this case [24] show that

$$k(x, x') = \sum_{c \in \mathcal{C}} k_c(x_c, x'_c) \tag{4}$$

can be used to describe all probability distributions with the above mentioned conditional independence relations using an exponential family model with k as its kernel. Since for exponential families expectations of the sufficient statistics yield injections, we have the following result:

Corollary 1. *On the class of probability distributions satisfying conditional independence properties according to a graph G with maximal clique set $\mathcal{C}$ and with full support on their domain, the operator*

$$\mu[\mathbf{P}] = \sum_{c \in \mathcal{C}} \mu_c[\mathbf{P}_c] = \sum_{c \in \mathcal{C}} \mathbf{E}_{x_c}\left[k_c(x_c, \cdot)\right] \tag{5}$$

is injective if the kernels k_c are all universal. The same decomposition holds for the empirical counterpart $\mu[X]$.

The condition of full support arises from the conditions of the Hammersley-Clifford Theorem [25, 26]: without it, not all conditionally independent random variables can be represented as the product of potential functions. Corollary 1 implies that we will be able to perform all subsequent operations on structured domains simply by dealing with mean operators on the corresponding maximal cliques.

1.4 Choosing the Hilbert Space

Identifying probability distributions with elements of Hilbert spaces is not new: see e.g. [27]. However, this leaves the obvious question of which Hilbert space to employ. We could informally choose a space with a kernel equalling the Delta distribution $k(x, x') = \delta(x, x')$, in which case the operator μ would simply be the identity map (which restricts us to probability distributions with square integrable densities).

The latter is in fact what is commonly done on finite domains (hence the L_2 integrability condition is trivially satisfied). For instance, [22] effectively use the Kronecker Delta $\delta(x_c, x'_c)$ as their feature map. The use of kernels has additional advantages: we need not deal with the issue of representation of the sufficient statistics or whether such a representation is minimal (i.e. whether the sufficient statistics actually span the space).

Whenever we have knowledge about the class of functions $\mathcal{F}$ we would like to analyze, we should be able to trade off simplicity in $\mathcal{F}$ with better approximation behavior in $\mathcal{P}$. For instance, assume that $\mathcal{F}$ contains only linear functions. In this case, μ only needs to map $\mathcal{P}$ into the space of all expectations of x. Consequently, one may expect very good constants in the convergence of $\mu[X]$ to $\mu[\mathbf{P}_x]$.

2 Applications

While the previous description may be of interest on its own, it is in application to areas of statistical estimation and artificial intelligence that its relevance becomes apparent.

2.1 Two-Sample Test

Since we know that $\mu[X] \to \mu[\mathbf{P}_x]$ with a fast rate (given appropriate behavior of $R_m(\mathcal{H}, \mathbf{P}_x)$), we may compare data drawn from two distributions $\mathbf{P}_x$ and $\mathbf{P}_y$, with associated samples X and Y, to test whether both distributions are identical; that is, whether $\mathbf{P}_x = \mathbf{P}_y$. For this purpose, recall that we defined $D(\mathbf{P}_x, \mathbf{P}_y) = \|\mu[\mathbf{P}_x] - \mu[\mathbf{P}_y]\|$. Using the reproducing property of an RKHS we may show [16] that

$$D^2(\mathbf{P}_x, \mathbf{P}_y) = \mathbf{E}_{x,x'}\left[k(x, x')\right] - 2\mathbf{E}_{x,y}\left[k(x, y)\right] + \mathbf{E}_{y,y'}\left[k(y, y')\right],$$

where x' is an independent copy of x, and y' an independent copy of y. An unbiased empirical estimator of $D^2(\mathbf{P}_x, \mathbf{P}_y)$ is a U-statistic [28],

$$\hat{D}^2(X,Y) := \tfrac{1}{m(m-1)} \sum_{i \neq j} h((x_i, y_i), (x_j, y_j)), \tag{6}$$

where

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

An equivalent interpretation, also in [16], is that we find a function that maximizes the difference in expectations between probability distributions. The resulting problem may be written

$$D(\mathbf{P}_x, \mathbf{P}_y) := \sup_{f \in \mathcal{F}} \left| \mathbf{E}_x[f(x)] - \mathbf{E}_y[f(y)] \right|. \tag{7}$$

To illustrate this latter setting, we plot the witness function f in Figure 1, when $\mathbf{P}_x$ is Gaussian and $\mathbf{P}_y$ is Laplace, for a Gaussian RKHS kernel. This function is straightforward to obtain, since the solution to Eq. (7) can be written $f(x) = \langle \mu[\mathbf{P}_x] - \mu[\mathbf{P}_y], \phi(x) \rangle$.

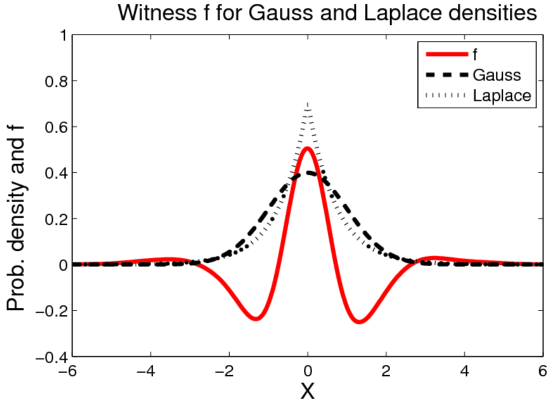

Fig. 1. Illustration of the function maximizing the mean discrepancy in the case where a Gaussian is being compared with a Laplace distribution. Both distributions have zero mean and unit variance. The function f that witnesses the difference in feature means has been scaled for plotting purposes, and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

The following two theorems give uniform convergence and asymptotic results, respectively. The first theorem is a straightforward application of [29, p. 25].

Theorem 4. *Assume that the kernel k is nonnegative and bounded by 1. Then with probability at least $1 - \delta$ the deviation $|D^2(\mathbf{P}_x, \mathbf{P}_y) - \hat{D}^2(X,Y)|$ is bounded by $4\sqrt{\log(2/\delta)/m}$.*

Note that an alternative uniform convergence bound is provided in [30], based on McDiarmid's inequality [31]. The second theorem appeared as [30, Theorem 8], and describes the asymptotic distribution of $\hat{D}^2(X,Y)$. When $\mathbf{P}_x \neq \mathbf{P}_y$, this distribution is given by [28, Section 5.5.1]; when $\mathbf{P}_x = \mathbf{P}_y$, it follows from [28, Section 5.5.2] and [32, Appendix].

Theorem 5. *We assume $\mathbf{E}\left(h^2\right) < \infty$. When $\mathbf{P}_x \neq \mathbf{P}_y$, $\hat{D}^2(X,Y)$ converges in distribution [33, Section 7.2] to a Gaussian according to*

$$m^{\frac{1}{2}}\left(\hat{D}^2(X,Y) - D^2(\mathbf{P}_x,\mathbf{P}_y)\right) \xrightarrow{D} \mathcal{N}\left(0,\sigma_u^2\right),$$

where $\sigma_u^2 = 4\left(\mathbf{E}_z\left[(\mathbf{E}_{z'}h(z,z'))^2\right] - \left[\mathbf{E}_{z,z'}(h(z,z'))\right]^2\right)$ and $z := (x,y)$, uniformly at rate $1/\sqrt{m}$ [28, Theorem B, p. 193]. When $\mathbf{P}_x = \mathbf{P}_y$, the U-statistic is degenerate, meaning $\mathbf{E}_{z'}h(z,z') = 0$. In this case, $\hat{D}^2(X,Y)$ converges in distribution according to

$$m\hat{D}^2(X,Y) \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l \left[g_l^2 - 2\right], \tag{8}$$

where $g_l \sim \mathcal{N}(0,2)$ 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 centered RKHS kernel.

We illustrate the MMD density by approximating it empirically for both $\mathbf{P}_x = \mathbf{P}_y$ (also called the null hypothesis, or H_0) and $\mathbf{P}_x \neq \mathbf{P}_y$ (the alternative hypothesis, or H_1). Results are plotted in Figure 2. We may use this theorem directly to test whether two distributions are identical, given an appropriate finite sample approximation to the $(1-\alpha)$th quantile of (8). In [16], this was achieved via two strategies: by using the bootstrap [34], and by fitting Pearson curves using the first four moments [35, Section 18.8].

While uniform convergence bounds have the theoretical appeal of making no assumptions on the distributions, they produce very weak tests. We find the test arising from Theorem 5 performs considerably better in practice. In addition, [36] demonstrate that this test performs very well in circumstances of high dimension and low sample size (i.e. when comparing microarray data), as well as being the only test currently applicable for structured data such as distributions on graphs. Moreover, the test can be used to determine whether records in databases may be matched based on their statistical properties. Finally, one may also apply it to extract features with the aim of *maximizing* discrepancy between sets of observations (see Section 2.4).

2.2 Covariate Shift Correction and Local Learning

A second application of the mean operator arises in situations of supervised learning where the training and test sets are drawn from different distributions,

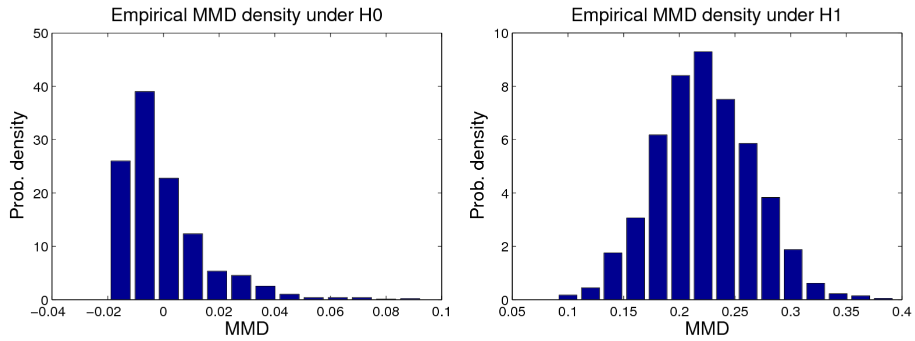

Fig. 2. Left: Empirical distribution of the MMD under H_0, with $\mathbf{P}_x$ and $\mathbf{P}_y$ both Gaussians with unit standard deviation, using 50 samples from each. **Right:** Empirical distribution of the MMD under H_1, with $\mathbf{P}_x$ a Laplace distribution with unit standard deviation, and $\mathbf{P}_y$ 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.

i.e. $X = \{x_1, \ldots, x_m\}$ is drawn from $\mathbf{P}_x$ and $X' = \{x'_1, \ldots, x'_{m'}\}$ is drawn from $\mathbf{P}_{x'}$. We assume, however, that the labels y are drawn from the same *conditional* distribution $\mathbf{P}_{y|x}$ on both the training and test sets.

The goal in this case is to find a weighting of the training set such that minimizing a reweighted empirical error on the training set will come close to minimizing the expected loss on the test set. That is, we would like to find weights $\{\beta_1, \ldots, \beta_m\}$ for X with $\sum_i \beta_i = 1$.

Obviously, if $\mathbf{P}_{y|x}$ is a rapidly changing function of x, or if the loss measuring the discrepancy between y and its estimate is highly non-smooth, this problem is difficult to solve. However, under regularity conditions spelled out in [37], one may show that by minimizing

$$\varDelta := \left\| \sum_{i=1}^{m} \beta_i k(x_i, \cdot) - \mu[X'] \right\|$$

subject to $\beta_i \geq 0$ and $\sum_i \beta_i = 1$, we will obtain weights which achieve this task. The idea here is that the expected loss with the expectation taken over $y|x$ should not change too quickly as a function of x. In this case we can use points x_i "nearby" to estimate the loss at location x'_j on the test set. Hence we are re-weighting the empirical distribution on the training set X such that the distribution behaves more like the empirical distribution on X'.

Note that by re-weighting X we will assign some observations a higher weight than $\frac{1}{m}$. This means that the statistical guarantees can no longer be stated in terms of the sample size m. One may show [37], however, that $\|\beta\|_2^{-2}$ now behaves like the effective sample size. Instead of minimizing $\varDelta$, it pays to minimize $\varDelta^2 + \lambda \|\beta\|_2^2$ subject to the above constraints. It is easy to show using the reproducing property of $\mathcal{H}$ that this corresponds to the following quadratic program:

$$\underset{\beta}{\text{minimize}} \ \frac{1}{2}\beta^\top (K + \lambda \mathbf{1})\beta - \beta^\top l \tag{9a}$$

$$\text{subject to } \beta_i \geq 0 \text{ and } \sum_i \beta_i = 1. \tag{9b}$$

Here $K_{ij} := k(x_i, x_j)$ denotes the kernel matrix and $l_i := \frac{1}{m}\sum_{j=1}^{m'} k(x_i, x'_j)$ is the expected value of $k(x_i, \cdot)$ on the test set X', i.e. $l_i = \langle k(x_i, \cdot), \mu[X'] \rangle$.

Experiments show that solving (9) leads to sample weights which perform very well in covariate shift. Remarkably, the approach can even outperform "importance sampler" weights, i.e. weights β_i obtained by computing the ratio $\beta_i = \mathbf{P}_{x'}(x_i)/\mathbf{P}_x(x_i)$. This is surprising, since the latter provide unbiased estimates of the expected error on X'. A point to bear in mind is that the kernels employed in the classification/regression learning algorithms of [37] are somewhat large, suggesting that the feature mean matching procedure is helpful when the learning algorithm returns relatively smooth classification/regression functions (we observe the same situation in the example of [38, Figure 1], where the model is "simpler" than the true function generating the data).

In the case where X' contains only a single observation, i.e. $X' = \{x'\}$, the above procedure leads to estimates which try to find a subset of observations in X and a weighting scheme such that the error at x' is approximated well. In practice, this leads to a local sample weighting scheme, and consequently an algorithm for local learning [39]. Our key advantage, however, is that we do not need to define the shape of the neighborhood in which we approximate the error at x'. Instead, this is automatically taken care of via the choice of the Hilbert space $\mathcal{H}$ and the location of x' relative to X.

2.3 Independence Measures

A third application of our mean mapping arises in measures of whether two random variables x and y are independent. Assume that pairs of random variables (x_i, y_i) are jointly drawn from some distribution $\mathbf{P}_{x,y}$. We wish to determine whether this distribution factorizes.

Having a measure of (in)dependence between random variables is a very useful tool in data analysis. One application is in independent component analysis [40], where the goal is to find a linear mapping of the observations x_i to obtain mutually independent outputs. One of the first algorithms to gain popularity was InfoMax, which relies on information theoretic quantities [41]. Recent developments using cross-covariance or correlation operators between Hilbert space representations have since improved on these results significantly [42, 43, 44]; in particular, a faster and more accurate quasi-Newton optimization procedure for kernel ICA is given in [45]. In the following we re-derive one of the above kernel independence measures using mean operators instead.

We begin by defining

$$\mu[\mathbf{P}_{xy}] := \mathbf{E}_{x,y}\left[v((x,y),\cdot)\right]$$
$$\text{and } \mu[\mathbf{P}_x \times \mathbf{P}_y] := \mathbf{E}_x\mathbf{E}_y\left[v((x,y),\cdot)\right].$$

Here we assumed that $\mathcal{V}$ is an RKHS over $\mathcal{X} \times \mathcal{Y}$ with kernel $v((x,y),(x',y'))$. If x and y *are* dependent, the equality $\mu[\mathbf{P}_{xy}] = \mu[\mathbf{P}_x \times \mathbf{P}_y]$ will not hold. Hence we may use $\Delta := \|\mu[\mathbf{P}_{xy}] - \mu[\mathbf{P}_x \times \mathbf{P}_y]\|$ as a measure of dependence.

Now assume that $v((x,y),(x',y')) = k(x,x')l(y,y')$, i.e. that the RKHS $\mathcal{V}$ is a direct product $\mathcal{H} \otimes \mathcal{G}$ of the RKHSs on $\mathcal{X}$ and $\mathcal{Y}$. In this case it is easy to see that

$$\begin{aligned}\Delta^2 &= \left\|\mathbf{E}_{xy}\left[k(x,\cdot)l(y,\cdot)\right] - \mathbf{E}_x\left[k(x,\cdot)\right]\mathbf{E}_y\left[l(y,\cdot)\right]\right\|^2 \\ &= \mathbf{E}_{xy}\mathbf{E}_{x'y'}\left[k(x,x')l(y,y')\right] - 2\mathbf{E}_x\mathbf{E}_y\mathbf{E}_{x'y'}\left[k(x,x')l(y,y')\right] \\ &\quad + \mathbf{E}_x\mathbf{E}_y\mathbf{E}_{x'}\mathbf{E}_{y'}\left[k(x,x')l(y,y')\right]\end{aligned}$$

The latter, however, is exactly what [43] show to be the Hilbert-Schmidt norm of the covariance operator between RKHSs: this is zero if and only if x and y are independent, for universal kernels. We have the following theorem:

Theorem 6. *Denote by C_{xy} the covariance operator between random variables x and y, drawn jointly from $\mathbf{P}_{xy}$, where the functions on $\mathcal{X}$ and $\mathcal{Y}$ are the reproducing kernel Hilbert spaces $\mathcal{F}$ and $\mathcal{G}$ respectively. Then the Hilbert-Schmidt norm $\|C_{xy}\|_{\mathrm{HS}}$ equals Δ.*

Empirical estimates of this quantity are as follows:

Theorem 7. *Denote by K and L the kernel matrices on X and Y respectively. Moreover, denote by $H = I - \mathbf{1}/m$ the projection matrix onto the subspace orthogonal to the vector with all entries set to 1. Then $m^{-2}\operatorname{tr} HKHL$ is an estimate of Δ^2 with bias $O(m^{-1})$. With high probability the deviation from Δ^2 is $O(m^{-\frac{1}{2}})$.*

See [43] for explicit constants. In certain circumstances, including in the case of RKHSs with Gaussian kernels, the empirical Δ^2 may also be interpreted in terms of a smoothed difference between the joint empirical characteristic function (ECF) and the product of the marginal ECFs [46, 47]. This interpretation does not hold in all cases, however, e.g. for kernels on strings, graphs, and other structured spaces. An illustration of the witness function of the equivalent optimization problem in Eq. 7 is provided in Figure 3. We observe that this is a smooth function which has large magnitude where the joint density is most different from the product of the marginals.

Note that if $v((x,y),\cdot)$ does *not* factorize we obtain a more general measure of dependence. In particular, we might not care about all types of interaction between x and y to an equal extent, and use an ANOVA kernel. Computationally efficient recursions are due to [48], as reported in [49]. More importantly, this representation will allow us to deal with *structured* random variables which are *not* drawn independently and identically distributed, such as time series.

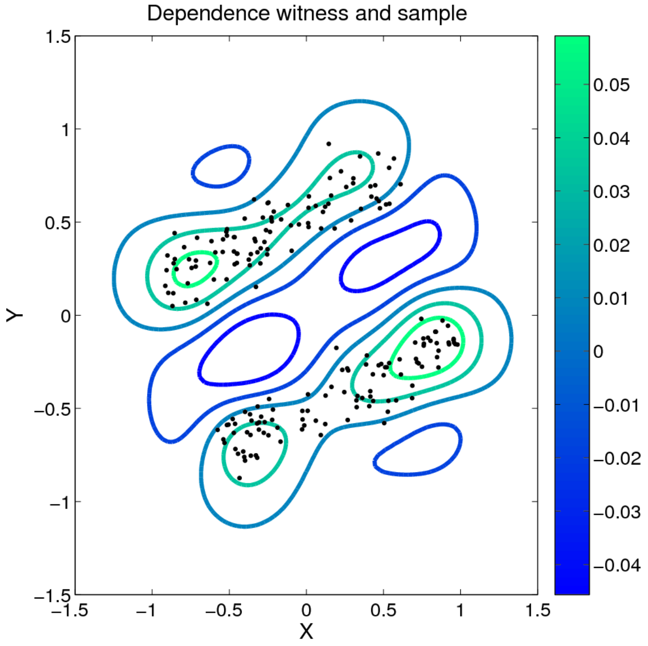

Fig. 3. Illustration of the function maximizing the mean discrepancy when MMD is used as a measure of independence. A sample from dependent random variables x and y is shown in black, and the associated function f that witnesses the MMD is plotted as a contour. The latter was computed empirically on the basis of 200 samples, using a Gaussian kernel with $\sigma = 0.2$.

For instance, in the case of EEG (electroencephalogram) data, we have both spatial and temporal structure in the signal. That said, few algorithms take full advantage of this when performing independent component analysis [50]. The pyramidal kernel of [51] is one possible choice for dependent random variables.

2.4 Feature Extraction

Kernel measures of statistical dependence need not be applied *only* to the analysis of independent components. To the contrary, we may also use them to extract highly dependent random variables, i.e. features. This procedure leads to variable selection algorithms with very robust properties [52].

The idea works as follows: given a set of patterns X and a set of labels Y, find a subset of features from X which maximizes $m^{-2} \operatorname{tr} HKHL$. Here L is the kernel matrix on the labels. In the most general case, the matrix K will arise from an arbitrary kernel k, for which no efficient decompositions exist. In this situation [52] suggests the use of a greedy feature removal procedure, i.e. to remove subsets of features iteratively such that $m^{-2} \operatorname{tr} HKHL$ is maximized for the remaining features.

In general, for particular choices of k and l, it is possible to recover well known feature selection methods, such as Pearson's correlation, shrunken centroid, or signal-to-noise ratio selection. Below we give some examples, mainly when a linear kernel $k(x, x') = \langle x, x' \rangle$. For more details see [53].

Pearson's Correlation is commonly used in microarray analysis [54, 55]. It is defined as

$$R_j := \frac{1}{m} \sum_{i=1}^{m} \left(\frac{x_{ij} - x_j}{s_{x_j}} \right) \left(\frac{y_i - y}{s_y} \right) \text{ where} \tag{10}$$

$$x_j = \frac{1}{m} \sum_{i=1}^{m} x_{ij} \text{ and } y = \frac{1}{m} \sum_{i=1}^{m} y_i$$

$$s_{x_j}^2 = \frac{1}{m} \sum_{i=1}^{m} (x_{ij} - x_j)^2 \text{ and } s_y^2 = \frac{1}{m} \sum_{i=1}^{m} (y_i - y)^2. \tag{11}$$

This means that all features are individually centered by x_j and scaled by their coordinate-wise variance s_{x_j} as a preprocessing step. Performing those operations before applying a linear kernel yields the formulation:

$$\operatorname{tr} KHLH = \operatorname{tr} \left(XX^\top Hyy^\top H \right) = \left\| HX^\top Hy \right\|^2 \tag{12}$$

$$= \sum_{j=1}^{d} \left(\sum_{i=1}^{m} \left(\frac{x_{ij} - x_j}{s_{x_j}} \right) \left(\frac{y_i - y}{s_y} \right) \right)^2 = \sum_{j=1}^{d} R_j^2. \tag{13}$$

Hence $\operatorname{tr} KHLH$ computes the sum of the squares of the Pearson Correlation (pc) coefficients. Since the terms are additive, feature selection is straightforward by picking the list of best performing features.

Centroid The difference between the means of the positive and negative classes at the jth feature, $(x_{j+} - x_{j-})$, is useful for scoring individual features. With different normalization of the data and the labels, many variants can be derived.

To obtain the centroid criterion [56] use $v_j := \lambda x_{j+} - (1-\lambda) x_{j-}$ for $\lambda \in (0,1)$ as the score[3] for feature j. Features are subsequently selected according to the absolute value $|v_j|$. In experiments the authors typically choose $\lambda = \frac{1}{2}$. For $\lambda = \frac{1}{2}$ we can achieve the same goal by choosing $L_{ii'} = \frac{y_i y_{i'}}{m_{y_i} m_{y_{i'}}}$ $(y_i, y_{i'} \in \{\pm 1\})$, in which case $HLH = L$, since the label kernel matrix is already centered. Hence we have

$$\operatorname{tr} KHLH = \sum_{i,i'=1}^{m} \frac{y_i y_{i'}}{m_{y_i} m_{y_{i'}}} x_i^\top x_{i'} \tag{14}$$

$$= \sum_{j=1}^{d} \left(\sum_{i,i'=1}^{m} \frac{y_i y_{i'} x_{ij} x_{i'j}}{m_{y_i} m_{y_{i'}}} \right) = \sum_{j=1}^{d} (x_{j+} - x_{j-})^2. \tag{15}$$

This proves that the centroid feature selector can be viewed as a special case of BAHSIC in the case of $\lambda = \frac{1}{2}$. From our analysis we see that other values of λ amount to effectively rescaling the patterns x_i *differently* for different classes, which may lead to undesirable features being selected.

***t*-Statistic** The normalization for the jth feature is computed as

$$\bar{s}_j = \left[\frac{s_{j+}^2}{m_+} + \frac{s_{j-}^2}{m_-} \right]^{\frac{1}{2}} \tag{16}$$

In this case we define the t-statistic for the jth feature via $t_j = (x_{j+} - x_{j-})/\bar{s}_j$. Compared to the Pearson correlation, the key difference is that now we normalize each feature not by the overall sample standard deviation but rather by a value which takes each of the two classes separately into account.

Signal to noise ratio is yet another criterion to use in feature selection. The key idea is to normalize each feature by $\bar{s}_j = s_{j+} + s_{j-}$ instead. Subsequently the $(x_{j+} - x_{j-})/\bar{s}_j$ are used to score features.

Moderated *t*-score is similar to t-statistic and is used for microarray analysis [57]. Its normalization for the jth feature is derived via a Bayes approach as

$$\tilde{s}_j = \frac{m \bar{s}_j^2 + m_0 \bar{s}_0^2}{m + m_0} \tag{17}$$

where $\bar{s}_j$ is from (16), and $\bar{s}_0$ and m_0 are hyperparameters for the prior distribution on $\bar{s}_j$ (all $\bar{s}_j$ are assumed to be iid). $\bar{s}_0$ and m_0 are estimated using

[3] The parameterization in [56] is different but it can be shown to be equivalent.

information from all feature dimensions. This effectively borrows information from the ensemble of features to aid with the scoring of an individual feature. More specifically, $\bar{s}_0$ and m_0 can be computed as [57]

$$m_0 = 2\Gamma'^{-1}\left(\frac{1}{d}\sum_{j=1}^{d}(z_j - \bar{z})^2 - \Gamma'\left(\frac{m}{2}\right)\right), \tag{18}$$

$$\bar{s}_0^2 = \exp\left(\bar{z} - \Gamma\left(\frac{m}{2}\right) + \Gamma\left(\frac{m_0}{2}\right) - \ln\left(\frac{m_0}{m}\right)\right), \tag{19}$$

where $\Gamma(\cdot)$ is the gamma function, $'$ denotes derivative, $z_j = \ln(\bar{s}_j^2)$ and $\bar{z} = \frac{1}{d}\sum_{j=1}^{d} z_j$.

***B*-statistic** is the logarithm of the posterior odds (lods) that a feature is differentially expressed. [58, 57] show that, for large number of features, the B-statistic is given by

$$B_j = a + b\tilde{t}_j^2, \tag{20}$$

where both a and b are constant ($b > 0$), and $\tilde{t}_j$ is the moderated-t statistic for the jth feature. Here we see that B_j is monotonic increasing in $\tilde{t}_j$, and thus results in the same gene ranking as the moderated-t statistic.

2.5 Density Estimation

General setting Obviously, we may also use the connection between mean operators and empirical means for the purpose of estimating densities. In fact, [59, 17, 60] show that this may be achieved in the following fashion:

$$\underset{\mathbf{P}_x}{\text{maximize}}\, H(\mathbf{P}_x) \text{ subject to } \|\mu[X] - \mu[\mathbf{P}_x]\| \leq \epsilon. \tag{21}$$

Here H is an entropy-like quantity (e.g. Kullback Leibler divergence, Csiszar divergence, Bregmann divergence, Entropy, Amari divergence) that is to be maximized subject to the constraint that the expected mean should not stray too far from its empirical counterpart. In particular, one may show that this approximate maximum entropy formulation is the dual of a maximum-a-posteriori estimation problem.

In the case of conditional probability distributions, it is possible to recover a raft of popular estimation algorithms, such as Gaussian Process classification, regression, and conditional random fields. The key idea in this context is to identify the sufficient statistics in generalized exponential families with the map $x \to k(x, \cdot)$ into a reproducing kernel Hilbert space.

Mixture model In problem (21) we try to find the optimal $\mathbf{P}_x$ over the entire space of probability distributions on $\mathcal{X}$. This can be an exceedingly costly optimization problem, in particular in the nonparametric setting. For instance, computing the normalization of the density itself may be intractable, in particular for high-dimensional data. In this case we may content ourselves with finding a suitable mixture distribution such that $\|\mu[X] - \mu[\mathbf{P}_x]\|$ is minimized with respect to the mixture coefficients. The diagram below summarizes our approach:

$$\text{density } \mathbf{P}_x \longrightarrow \text{ sample } X \longrightarrow \text{ emp. mean } \mu[X] \longrightarrow \text{ estimate via } \mu[\widehat{\mathbf{P}}_x] \tag{22}$$

The connection between $\mu[\mathbf{P}_x]$ and $\mu[X]$ follows from Theorem 2. To obtain a density estimate from $\mu[X]$ assume that we have a set of candidate densities $\mathbf{P}_x^i$ on $\mathcal{X}$. We want to use these as basis functions to obtain $\widehat{\mathbf{P}}_x$ via

$$\widehat{\mathbf{P}}_x = \sum_{i=1}^{M} \beta_i \mathbf{P}_x^i \text{ where } \sum_{i=1}^{M} \beta_i = 1 \text{ and } \beta_i \geq 0. \tag{23}$$

In other words we wish to estimate $\mathbf{P}_x$ by means of a mixture model with mixture densities $\mathbf{P}_x^i$. The goal is to obtain good estimates for the coefficients β_i and to obtain performance guarantees which specify how well $\widehat{\mathbf{P}}_x$ is capable of estimating $\mathbf{P}_x$ in the first place. This is possible using a very simple optimization problem:

$$\underset{\beta}{\text{minimize}} \left\| \mu[X] - \mu[\widehat{\mathbf{P}}_x] \right\|_{\mathcal{H}}^2 \text{ subject to } \beta^\top \mathbf{1} = 1 \text{ and } \beta \geq 0. \tag{24}$$

To ensure good generalization performance we add a regularizer $\Omega[\beta]$ to the optimization problem, such as $\frac{1}{2}\|\beta\|^2$. It follows using the expansion of $\widehat{\mathbf{P}}_x$ in (23) that the resulting optimization problem can be reformulated as a quadratic program via

$$\underset{\beta}{\text{minimize}} \frac{1}{2}\beta^\top [Q + \lambda \mathbf{1}]\beta - l^\top \beta \text{ subject to } \beta^\top \mathbf{1} = 1 \text{ and } \beta \geq 0. \tag{25}$$

Here $\lambda > 0$ is a regularization constant, and the quadratic matrix $Q \in \mathbb{R}^{M \times M}$ and the vector $l \in \mathbb{R}^M$ are given by

$$Q_{ij} = \left\langle \mu[\mathbf{P}_x^i], \mu[\mathbf{P}_x^j] \right\rangle = \mathop{\mathbf{E}}_{x^i, x^j} \left[k(x^i, x^j) \right] \tag{26}$$

$$\text{and } l_j = \left\langle \mu[X], \mu[\mathbf{P}_x^j] \right\rangle = \frac{1}{m} \sum_{i=1}^{m} \mathop{\mathbf{E}}_{x^j} \left[k(x_i, x^j) \right]. \tag{27}$$

By construction $Q \succeq 0$ is positive semidefinite, hence the quadratic program (25) is convex. For a number of kernels and mixture terms $\mathbf{P}_x^i$ we are able to compute Q, l in closed form.

Since $\widehat{\mathbf{P}}_x$ is an empirical estimate it is quite unlikely that $\widehat{\mathbf{P}}_x = \mathbf{P}_x$. This raises the question of how well expectations with respect to $\mathbf{P}_x$ are approximated by those with respect to $\widehat{\mathbf{P}}_x$. This can be answered by an extension of the Koksma-Hlawka inequality [61].

Lemma 1. *Let $\epsilon > 0$ and let $\epsilon' := \left\| \mu[X] - \mu[\widehat{\mathbf{P}}_x] \right\|$. Under the assumptions of Theorem 2 we have that with probability at least $1 - \exp(-\epsilon^2 m R^{-2})$,*

$$\sup_{\|f\|_{\mathcal{H}} \leq 1} \left| \mathbf{E}_{x \sim \mathbf{P}_x}[f(x)] - \mathbf{E}_{x \sim \widehat{\mathbf{P}}_x}[f(x)] \right| \leq 2R_m(\mathcal{H}, \mathbf{P}_x) + \epsilon + \epsilon'. \tag{28}$$

Proof We use that in Hilbert spaces, $\mathbf{E}_{x \sim \mathbf{P}_x}[f(x)] = \langle f, \mu[\mathbf{P}_x] \rangle$ and $\mathbf{E}_{x \sim \widehat{\mathbf{P}}_x}[f(x)] = \left\langle f, \mu[\widehat{\mathbf{P}}_x] \right\rangle$ both hold. Hence the LHS of (28) equates to $\sup_{\|f\|_{\mathcal{H}} \leq 1} \left| \left\langle \mu[\mathbf{P}_x] - \mu[\widehat{\mathbf{P}}_x], f \right\rangle \right|$, which is given by the norm of $\left\| \mu[\mathbf{P}_x] - \mu[\widehat{\mathbf{P}}_x] \right\|$. The triangle inequality, our assumption on $\mu[\widehat{\mathbf{P}}_x]$, and Theorem 2 complete the proof. ■

This means that we have good control over the behavior of expectations of random variables, as long as they belong to "smooth" functions on $\mathcal{X}$ — the uncertainty increases with their RKHS norm.

The above technique is useful when it comes to representing distributions in message passing and data compression. Rather than minimizing an information theoretic quantity, we can choose a Hilbert space which accurately reflects the degree of smoothness required for any subsequent operations carried out by the estimate. For instance, if we are only interested in linear functions, an accurate match of the first order moments will suffice, without requiring a good match in higher order terms.

2.6 Kernels on Sets

Up to now we used the mapping $X \to \mu[X]$ to compute the distance between two distributions (or their samples). However, since $\mu[X]$ itself is an element of an RKHS we can define a kernel on sets (and distributions) directly via

$$k(X, X') := \langle \mu[X], \mu[X'] \rangle = \tfrac{1}{mm'} \sum_{i,j}^{m,m'} k(x_i, x'_j). \tag{29}$$

In other words, $k(X, X')$, and by analogy $k(\mathbf{P}_x, \mathbf{P}_{x'}) := \langle \mu[\mathbf{P}_x], \mu[\mathbf{P}_{x'}] \rangle$, define kernels on sets and distributions, and obviously also between sets and distributions. If we have multisets and sample weights for instances we may easily include this in the computation of $\mu[X]$. It turns out that (29) is exactly the set kernel proposed by [62], when dealing with multiple instance learning. This notion was subsequently extended to deal with intermediate density estimates by [63]. We have therefore that in situations where estimation problems are well described by distributions we inherit the consistency properties of the underlying RKHS simply by using a universal set kernel for which $\mu[X]$ converges to $\mu[\mathbf{P}_x]$. We have the following corollary:

Corollary 2. *If k is universal the kernel matrix defined by the set/distribution kernel (29) has full rank as long as the sets/distributions are not identical.*

Note, however, that the set kernel may not be ideal for all multi instance problems: in the latter one assumes that at least a *single instance* has a given property, whereas for the use of (29) one needs to assume that at least a certain *fraction of instances* have this property.

3 Summary

We have seen that Hilbert space embeddings of distributions are a powerful tool to deal with a broad range of estimation problems, including two-sample tests, feature extractors, independence tests, covariate shift, local learning, density estimation, and the measurement of similarity between sets. Given these successes, we are very optimistic that these embedding techniques can be used to address further problems, ranging from issues in high dimensional numerical integration (the connections to lattice and Sobol sequences are apparent) to more advanced nonparametric property testing.

Acknowledgments We thank Karsten Borgwardt, Kenji Fukumizu, Jiayuan Huang, Quoc Le, Malte Rasch, and Vladimir Vapnik for helpful discussions. 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

[1] Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)

[2] Schölkopf, B., Smola, A.: Learning with Kernels. MIT Press, Cambridge, MA (2002)

[3] Joachims, T.: Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms. Kluwer Academic Publishers, Boston (2002)

[4] Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA (2006)

[5] Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley and Sons, New York (1991)

[6] Amari, S., Nagaoka, H.: Methods of Information Geometry. Oxford University Press (1993)

[7] Krause, A., Guestrin, C.: Near-optimal nonmyopic value of information in graphical models. In: Uncertainty in Artificial Intelligence UAI'05. (2005)

[8] Slonim, N., Tishby, N.: Agglomerative information bottleneck. In Solla, S.A., Leen, T.K., Müller, K.R., eds.: Advances in Neural Information Processing Systems 12, Cambridge, MA, MIT Press (2000) 617–623

[9] Stögbauer, H., Kraskov, A., Astakhov, S., Grassberger, P.: Least dependent component analysis based on mutual information. Phys. Rev. E **70**(6) (2004) 066123

[10] Nemenman, I., Shafee, F., Bialek, W.: Entropy and inference, revisited. In: Neural Information Processing Systems. Volume 14., Cambridge, MA, MIT Press (2002)

[11] Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge, UK (2004)

[12] Schölkopf, B., Tsuda, K., Vert, J.P.: Kernel Methods in Computational Biology. MIT Press, Cambridge, MA (2004)

[13] Hofmann, T., Schölkopf, B., Smola, A.J.: A review of kernel methods in machine learning. Technical Report 156, Max-Planck-Institut für biologische Kybernetik (2006)

[14] Steinwart, I.: The influence of the kernel on the consistency of support vector machines. Journal of Machine Learning Research **2** (2002)

[15] Fukumizu, K., Bach, F.R., Jordan, M.I.: Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. J. Mach. Learn. Res. **5** (2004) 73–99

[16] Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., Smola, A.J.: A kernel method for the two-sample-problem. In Schölkopf, B., Platt, J., Hofmann, T., eds.: Advances in Neural Information Processing Systems. Volume 19., The MIT Press, Cambridge, MA (2007)

[17] Altun, Y., Smola, A.: Unifying divergence minimization and statistical inference via convex duality. In Simon, H., Lugosi, G., eds.: Proc. Annual Conf. Computational Learning Theory. LNCS, Springer (2006) 139–153

[18] Bartlett, P.L., Mendelson, S.: Rademacher and gaussian complexities: Risk bounds and structural results. J. Mach. Learn. Res. **3** (2002) 463–482

[19] Koltchinskii, V.: Rademacher penalties and structural risk minimization. IEEE Trans. Inform. Theory **47** (2001) 1902–1914

[20] Vapnik, V., Chervonenkis, A.: On the uniform convergence of relative frequencies of events to their probabilities. Theory Probab. Appl. **16**(2) (1971) 264–281

[21] Vapnik, V., Chervonenkis, A.: The necessary and sufficient conditions for the uniform convergence of averages to their expected values. Teoriya Veroyatnostei i Ee Primeneniya **26**(3) (1981) 543–564

[22] Wainwright, M.J., Jordan, M.I.: Graphical models, exponential families, and variational inference. Technical Report 649, UC Berkeley, Department of Statistics (September 2003)

[23] Ravikumar, P., Lafferty, J.: Variational chernoff bounds for graphical models. In: Uncertainty in Artificial Intelligence UAI04. (2004)

[24] Altun, Y., Smola, A.J., Hofmann, T.: Exponential families for conditional random fields. In: Uncertainty in Artificial Intelligence (UAI), Arlington, Virginia, AUAI Press (2004) 2–9

[25] Hammersley, J.M., Clifford, P.E.: Markov fields on finite graphs and lattices. unpublished manuscript (1971)

[26] Besag, J.: Spatial interaction and the statistical analysis of lattice systems (with discussion). J. Roy. Stat. Soc. Ser. B Stat. Methodol. **36**(B) (1974) 192–326

[27] Hein, M., Bousquet, O.: Hilbertian metrics and positive definite kernels on probability measures. In Ghahramani, Z., Cowell, R., eds.: Proc. of AI & Statistics. Volume 10. (2005)

[28] Serfling, R.: Approximation Theorems of Mathematical Statistics. Wiley, New York (1980)

[29] Hoeffding, W.: Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association **58** (1963) 13–30

[30] Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., Smola, A.: A kernel method for the two-sample-problem. In: Advances in Neural Information Processing Systems 19, Cambridge, MA, MIT Press (2007)

[31] McDiarmid, C.: On the method of bounded differences. Surveys in Combinatorics (1969) 148–188 Cambridge University Press.

[32] Anderson, N., Hall, P., Titterington, D.: Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. Journal of Multivariate Analysis **50** (1994) 41–54

[33] Grimmet, G.R., Stirzaker, D.R.: Probability and Random Processes. Third edn. Oxford University Press, Oxford (2001)

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

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

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

[37] Huang, J., Smola, A., Gretton, A., Borgwardt, K., Schölkopf, B.: Correcting sample selection bias by unlabeled data. In Schölkopf, B., Platt, J., Hofmann, T., eds.: Advances in Neural Information Processing Systems. Volume 19., The MIT Press, Cambridge, MA (2007)

[38] Shimodaira, H.: Improving predictive inference under convariance shift by weighting the log-likelihood function. Journal of Statistical Planning and Inference **90** (2000)

[39] Bottou, L., Vapnik, V.N.: Local learning algorithms. Neural Computation **4**(6) (1992) 888–900

[40] Comon, P.: Independent component analysis, a new concept? Signal Processing **36** (1994) 287–314

[41] Lee, T.W., Girolami, M., Bell, A., Sejnowski, T.: A unifying framework for independent component analysis. Comput. Math. Appl. **39** (2000) 1–21

[42] Bach, F.R., Jordan, M.I.: Kernel independent component analysis. J. Mach. Learn. Res. **3** (2002) 1–48

[43] Gretton, A., Bousquet, O., Smola, A., Schölkopf, B.: Measuring statistical dependence with Hilbert-Schmidt norms. In Jain, S., Simon, H.U., Tomita, E., eds.: Proceedings Algorithmic Learning Theory, Berlin, Germany, Springer-Verlag (2005) 63–77

[44] Gretton, A., Herbrich, R., Smola, A., Bousquet, O., Schölkopf, B.: Kernel methods for measuring independence. J. Mach. Learn. Res. **6** (2005) 2075–2129

[45] Shen, H., Jegelka, S., Gretton, A.: Fast kernel ICA using an approximate newton method. In: AISTATS 11. (2007)

[46] Feuerverger, A.: A consistent test for bivariate dependence. International Statistical Review **61**(3) (1993) 419–433

[47] Kankainen, A.: Consistent Testing of Total Independence Based on the Empirical Characteristic Function. PhD thesis, University of Jyväskylä (1995)

[48] Burges, C.J.C., Vapnik, V.: A new method for constructing artificial neural networks. Interim technical report, ONR contract N00014-94-c-0186, AT&T Bell Laboratories (1995)

[49] Vapnik, V.: Statistical Learning Theory. John Wiley and Sons, New York (1998)

[50] Anemuller, J., Duann, J.R., Sejnowski, T.J., Makeig, S.: Spatio-temporal dynamics in fmri recordings revealed with complex independent component analysis. Neurocomputing **69** (2006) 1502–1512

[51] Schölkopf, B.: Support Vector Learning. R. Oldenbourg Verlag, Munich (1997) Download: http://www.kernel-machines.org.

[52] Song, L., Smola, A., Gretton, A., Borgwardt, K., Bedo, J.: Supervised feature selection via dependence estimation. In: Proc. Intl. Conf. Machine Learning. (2007)

[53] Song, L., Bedo, J., Borgwardt, K., Gretton, A., Smola, A.: Gene selection via the BAHSIC family of algorithms. In: Bioinformatics (ISMB). (2007) To appear.

[54] van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A.M., et al.: Gene expression profiling predicts clinical outcome of breast cancer. Nature **415** (2002) 530–536

[55] Ein-Dor, L., Zuk, O., Domany, E.: Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc. Natl. Acad. Sci. USA **103**(15) (Apr 2006) 5923–5928

[56] Bedo, J., Sanderson, C., Kowalczyk, A.: An efficient alternative to svm based recursive feature elimination with applications in natural language processing and bioinformatics. In: Artificial Intelligence. (2006)

[57] Smyth, G.: Linear models and empirical bayes methods for assessing differential expressionin microarray experiments. Statistical Applications in Genetics and Molecular Biology **3** (2004)

[58] Lönnstedt, I., Speed, T.: Replicated microarray data. Statistica Sinica **12** (2002) 31–46

[59] Dudík, M., Schapire, R., Phillips, S.: Correcting sample selection bias in maximum entropy density estimation. In: Advances in Neural Information Processing Systems 17. (2005)

[60] Dudík, M., Schapire, R.E.: Maximum entropy distribution estimation with generalized regularization. In Lugosi, G., Simon, H.U., eds.: Proc. Annual Conf. Computational Learning Theory, Springer Verlag (June 2006)

[61] Hlawka, E.: Funktionen von beschränkter variation in der theorie der gleichverteilung. Annali di Mathematica Pura ed Applicata **54** (1961)

[62] Gärtner, T., Flach, P.A., Kowalczyk, A., Smola, A.J.: Multi-instance kernels. In: Proc. Intl. Conf. Machine Learning. (2002)

[63] Jebara, T., Kondor, I.: Bhattacharyya and expected likelihood kernels. In Schölkopf, B., Warmuth, M., eds.: Proceedings of the Sixteenth Annual Conference on Computational Learning Theory. Number 2777 in Lecture Notes in Computer Science, Heidelberg, Germany, Springer-Verlag (2003) 57–71