The Annals of Applied Statistics
2009, Vol. 3, No. 4, 1285–1294
DOI: 10.1214/09-AOAS312E
Main article DOI: 10.1214/09-AOAS312

DISCUSSION OF: BROWNIAN DISTANCE COVARIANCE

BY ARTHUR GRETTON[1], KENJI FUKUMIZU AND BHARATH K. SRIPERUMBUDUR[2]

Carnegie Mellon University, MPI for Biological Cybernetics, The Institute of Statistical Mathematics, Department of Electrical and MPI for Biological Cybernetics and Computer Engineering, UCSD

1. Introduction. A dependence statistic, the Brownian Distance Covariance, has been proposed for use in dependence measurement and independence testing: we refer to this contribution henceforth as SR [we also note the earlier work on this topic of Székely, Rizzo and Bakirov (2007)]. Some advantages of the authors' approach are that the random variables X and Y being tested may have arbitrary dimension $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively; and the test is consistent against all alternatives subject to the conditions $\mathbf{E}\|X\|_p < \infty$ and $\mathbf{E}\|X\|_q < \infty$.

In our discussion we review and compare against a number of related dependence measures that have appeared in the statistics and machine learning literature. We begin with distances of the form of SR, equation (2.2), most notably the work of Feuerverger (1993); Kankainen (1995); Kankainen and Ushakov (1998); Ushakov (1999), which we describe in Section 2: these measures have been formulated only for the case $p = q = 1$, however. In Section 3 we turn to more recent dependence measures which are computed between mappings of the probability distributions $\mathbf{P}_x$, $\mathbf{P}_y$, and $\mathbf{P}_{xy}$ of X, Y, and (X, Y), respectively, to high dimensional feature spaces: specifically, reproducing kernel Hilbert spaces (RKHSs). The RKHS dependence statistics may be based on the distance [Smola et al. (2007), Section 2.3], covariance [Gretton et al. (2005a, 2005b, 2008)], or correlation [Dauxois and Nkiet (1998); Bach and Jordan (2002); Fukumizu, Bach and Gretton (2007); Fukumizu et al. (2008)] between the feature mappings, and make smoothness assumptions which can improve the power of the tests over approaches relying on distances between the unmapped variables. When the RKHSs are characteristic [Fukumizu et al. (2008); Sriperumbudur et al. (2008)], meaning that the feature mapping from the space of probability measures to the RKHS is injective, the kernel-based tests are consistent for all probability measures generating (X, Y).

[1]Supported by Grants DARPA IPTO FA8750-09-1-0141, ONR MURI N000140710747, and ARO MURI W911NF0810242.

[2]Supported by Max Planck Institute (MPI) for Biological Cybernetics, NSF Grant DMS-MSPA 0625409, the Fair Isaac Corporation, and the University of California MICRO program.

Key words and phrases. Independence testing, Brownian distance covariance, covariance operator, kernel methods, reproducing kernel Hilbert space.

RKHS-based tests apply on spaces $\mathbb{R}^p \times \mathbb{R}^q$ for arbitrary p and q. In fact, kernel independence tests are applicable on a still broader range of (possibly non-Euclidean) domains, which can include strings [Leslie et al. (2002)], graphs [Gärtner, Flach and Wrobel (2003)], and groups [Fukumizu et al. (2009)], making the kernel approach very general. In Section 4 we provide an empirical comparison between the approach of SR and the kernel statistic of Gretton et al. (2005b, 2008) on an independence testing benchmark.

2. Characteristic function-based dependence measures. We begin with a brief review of characteristic function-based independence measures related to the statistic $\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})$ in SR, equation (2.8); see also Ushakov (1999), Section 3.7.

Feuerverger (1993) proposes two statistics for independence testing, in the case where X and Y are univariate. The first, described by Feuerverger [(1993), Section 4], is

$$T_n := \int\int \frac{|\Gamma_n'(s,t)|^2}{(1-e^{-s^2})(1-e^{-t^2})} W(s,t)\,ds\,dt,$$

where $W(s,t)$ is a weight function,

$$\Gamma_n'(s,t) := f_{\widetilde{X}\widetilde{Y}}^n(s,t) - f_{\widetilde{X}}^n(s) f_{\widetilde{Y}}^n(t),$$

and $f_{\widetilde{X}\widetilde{Y}}^n$, $f_{\widetilde{X}}^n$, and $f_{\widetilde{Y}}^n$ denote the empirical characteristic functions (in accordance with the notation of SR), however, these take as their argument the approximate normal scores of the sample points,

$$\widetilde{X}_i := \Phi^{-1}\left(\frac{\operatorname{rank}(X_i) - 3/8}{n + 1/4}\right). \tag{2.1}$$

With an appropriate choice of weight function, Feuerverger obtains the statistic

$$T_n^{(1)} = \frac{\pi^2}{n^2}\sum_{j,k} |\widetilde{X}_j - \widetilde{X}_k||\widetilde{Y}_j - \widetilde{Y}_k| - \frac{2\pi^2}{n^3}\sum_{j,q,r} |\widetilde{X}_j - \widetilde{X}_q||\widetilde{Y}_j - \widetilde{Y}_r| + \frac{\pi^2}{n^4}\sum_{j,k,q,r} |\widetilde{X}_j - \widetilde{X}_k||\widetilde{Y}_q - \widetilde{Y}_r|,$$

where the summation indices denote all r-tuples drawn with replacement from the set $\{1, \ldots, n\}$, and r is the number of indices of the sum. This statistic takes a form similar to the statistic $\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})$ in SR, equation (2.8), the main differences being the restriction to the univariate case, use of the 1-norm, and transformation (2.1). A second statistic, described by Feuerverger [(1993), Section 5], is written

$$T_n := \int\int |\Gamma_n(s,t)|^2 W(s,t)\,ds\,dt, \tag{2.2}$$

where the term $\Gamma_n(s,t)$ now simply denotes the difference between the joint characteristic function and the product of the marginals [in other words, the statistic

is identical to that in SR, equation (2.1)]. Feuerverger remarks that, for certain choices of $W(s,t)$, the resulting statistic resembles that of Rosenblatt (1975), being the ℓ_2 distance between the kernel density estimate of the joint distribution and that of the product of the marginals. As an illustration, Kankainen [(1995), page 54], makes this link explicit, employing a Gaussian weight function to obtain the statistic

$$T_n^{(2)} = \frac{1}{n^2}\sum_{j,k} k_{jk}l_{jk} - \frac{2}{n^3}\sum_{j,q,r} k_{jq}l_{jr} + \frac{1}{n^4}\sum_{j,k,qr} k_{jk}l_{qr}, \tag{2.3}$$

where

$$k_{jk} := \exp\left(\frac{-\|X_j - X_k\|^2}{2\sigma_x^2}\right) \quad \text{and} \quad l_{qr} := \exp\left(\frac{-\|Y_q - Y_r\|^2}{2\sigma_x^2}\right). \tag{2.4}$$

One can readily see that this involves transforming the distances of $\mathcal{V}_n^2(\mathbf{X},\mathbf{Y})$ in SR, equation (2.8), by passing them through a Gaussian distortion: this replaces the finite expected norm condition required by SR with a weaker requirement.

A further difference of Kankainen (1995) with respect to Feuerverger (1993) is that Kankainen generalizes to the problem of testing mutual independence, although the variables themselves remain univariate. Kankainen further enforces scale and location invariance by studentizing each variable. Finally, despite their superficial resemblance, a number of important differences nonetheless exist between the statistic in (2.2) and that of Rosenblatt (1975). Most crucially, the kernel bandwidth is kept fixed for the characteristic function-based test, rather than decreasing as n rises (a decreasing bandwidth is needed to ensure consistency of the kernel density estimates), resulting in very different forms for the null distribution; and there are more restrictive conditions on the Rosenblatt–Parzen test statistic [Rosenblatt (1975), conditions a.1–a.4]. These issues are discussed further by Feuerverger [(1993), Section 5], and Kankainen [(1995), Section 5.4]. An empirical comparison of the null distributions resulting from fixed vs decreasing bandwidth is provided by Gretton and Györfi (2008).

3. RKHS-based dependence measures. We now present a class of dependence measures (henceforth *kernel dependence measures*) based on mappings of the random variables to reproducing kernel Hilbert spaces, which encode features of interest for these variables. We first use Bochner's theorem to demonstrate that a subclass of kernel dependence measures is equivalent to SR, equation (2.2), under appropriate conditions on the weight function. Next, we give an interpretation in terms of covariances between feature space mappings, from which we may generalize to broader classes of kernel dependence measures, including correlations and estimates of the mean square contingency.

3.1. *Kernel dependence measures via Bochner's theorem.* We describe a dependence measure introduced by Gretton et al. (2005b, 2008) which constitutes the kernel statistic most closely resembling the characteristic function-based statistic of SR, equation (2.2). The present derivation follows Smola et al. (2007), Section 2.3. We begin with some necessary terminology and definitions. Let $z := (x, y) \in \mathbb{R}^{(p+q)}$, and $\mathcal{H}$ be an RKHS with the continuous feature mapping $\theta(z) \in \mathcal{H}$ for each $z \in \mathbb{R}^{(p+q)}$, such that the inner product between the features is given by the positive definite kernel function $h(z, z') := \langle \theta(z), \theta(z') \rangle_{\mathcal{H}}$. We remark that we never need deal with the feature representations $\theta(z)$ explicitly (indeed, these may be infinite dimensional): rather, we express our statistic entirely in terms of the kernel function, which is the inner product between two such mappings. If we restrict ourselves to kernels that can be written in terms of the difference of their arguments, $h(z, z') = \lambda(z - z')$, the following theorem applies [Wendland (2005), Theorem 6.6].

THEOREM 3.1 (Bochner). *A continuous function $\lambda : \mathbb{R}^{(p+q)} \to \mathbb{R}$ is positive definite if and only if it is the Fourier transform of a finite nonnegative Borel measure $W(u)\,du$ on $\mathbb{R}^{(p+q)}$, that is,*

$$\lambda(z) = \int_{\mathbb{R}^{(p+q)}} e^{-iz^T u} W(u)\,du, \qquad z \in \mathbb{R}^{(p+q)}. \tag{3.1}$$

Let us consider the following distance between the joint distribution $\mathbf{P} := \mathbf{P}_{xy}$ and the product of the marginals, $\mathbf{Q} := \mathbf{P}_x \mathbf{P}_y$:

$$H = \int |f_P(u) - f_Q(u)|^2 W(u)\,du,$$

where f_P and f_Q are the characteristic functions for $\mathbf{P}$ and $\mathbf{Q}$, respectively. Assuming further that we can decompose $\lambda(z - z') = k(x - x')l(y - y')$ (on which more below), we can rewrite H as

$$\begin{aligned}
H &= \int \left\{ \int e^{iz^T u}\,d\mathbf{P}(z) - \int e^{iz^T u}\,d\mathbf{Q}(z) \right\} \\
&\quad \times \left\{ \int e^{-iz'^T u}\,d\mathbf{P}(z') - \int e^{-iz'^T u}\,d\mathbf{Q}(z') \right\} W(u)\,du \\
&= \int \left\{ \int\int e^{i(z-z')^T u}\,d\mathbf{P}(z)\,d\mathbf{P}(z') - \int\int e^{i(z-z')^T u}\,d\mathbf{P}(z)\,d\mathbf{Q}(z') \right. \\
&\qquad \left. - \int\int e^{i(z-z')^T u}\,d\mathbf{Q}(z)\,d\mathbf{P}(z') + \int\int e^{i(z-z')^T u}\,d\mathbf{Q}(z)\,d\mathbf{Q}(z') \right\} W(u)\,du \\
&= \int\int \lambda(z - z')\,d\mathbf{P}(z)\,d\mathbf{P}(z') - \int\int \lambda(z - z')\,d\mathbf{P}(z)\,d\mathbf{Q}(z') \\
&\quad - \int\int \lambda(z - z')\,d\mathbf{Q}(z)\,d\mathbf{P}(z') + \int\int \lambda(z - z')\,d\mathbf{Q}(z)\,d\mathbf{Q}(z')
\end{aligned}$$

$$= \mathbf{E}\{k(X - X')l(Y - Y')\} + \mathbf{E}\{k(X - X')\}\mathbf{E}\{l(Y - Y')\}$$
$$- 2\mathbf{E}\big\{\mathbf{E}\{k(X - X')|X\}\mathbf{E}\{k(Y - Y')|Y\}\big\}.$$

We call H the Hilbert–Schmidt independence criterion (HSIC). The test statistic in (2.3) is then interpreted as a biased empirical estimate of H [an unbiased estimate would replace the V-statistics with U-statistics; see Gretton et al. (2008)]. We remark at this point that the weight function $1/(|t|_p^{1+p}|s|_q^{1+q})$ is not integrable, hence, Bochner's theorem does not apply for this choice of $W(u)$. Thus, interpreting the statistic in SR, equation (2.6), as a kernel statistic is not straightforward.

3.2. *Kernel dependence measures via covariance operators.* We now obtain HSIC via a different argument, based on the covariance between feature mappings of the variables: we then generalize this to correlation-based dependence measures, with reference to the statistic $\mathcal{R}^2(X, Y)$ of SR. Our brief review draws heavily on the overview of Gretton and Györfi (2009), Section 4. Let $\mathcal{F}$ be an RKHS on $\mathbb{R}^p$ with feature map $\phi(X)$ and kernel $k(X, X') := \langle \phi(X), \phi(X') \rangle_{\mathcal{F}}$, and $\mathcal{G}$ be a second RKHS on $\mathbb{R}^q$ with kernel $l(\cdot, \cdot)$ and feature map $\psi(y)$. Following Baker (1973); Fukumizu, Bach and Jordan (2004); Gretton et al. (2005a); Fukumizu, Bach and Jordan (2009), the cross-covariance operator $C_{xy} : \mathcal{G} \to \mathcal{F}$ for the measure $\mathbf{P}_{xy}$ is defined such that, for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$,

$$\langle f, C_{xy} g \rangle_{\mathcal{F}} = \mathbf{E}\big([f(X) - \mathbf{E}(f(X))][g(Y) - \mathbf{E}(g(Y))]\big).$$

The cross-covariance operator can be thought of as a generalization of a cross-covariance matrix between the (potentially infinite dimensional) feature mappings $\phi(x)$ and $\psi(y)$.

To see how this operator may be used to test independence, we recall the following characterization of independence [see, e.g., Jacod and Protter (2000), Theorem 10.1e]:

THEOREM 3.2. *The random variables X and Y are independent if and only if* $\operatorname{cov}(f(X), g(Y)) = 0$ *for any pair* (f, g) *of bounded, continuous functions.*

While the bounded continuous functions are too rich a class to permit the construction of a covariance-based test statistic on a sample, Fukumizu et al. (2008); Sriperumbudur et al. (2008) show that when $\widetilde{\mathcal{F}}$ is the unit ball in a *characteristic*[3] RKHS $\mathcal{F}$, and $\widetilde{\mathcal{G}}$ the unit ball in a characteristic RKHS $\mathcal{G}$, then

$$\sup_{f \in \widetilde{\mathcal{F}}, g \in \widetilde{\mathcal{G}}} \mathbf{E}\big([f(X) - \mathbf{E}(f(X))][g(Y) - \mathbf{E}(g(Y))]\big) = 0 \quad \Longleftrightarrow \quad \mathbf{P}_{xy} = \mathbf{P}_x \mathbf{P}_y.$$

[3]The reader is referred to [Fukumizu et al. (2008); Sriperumbudur et al. (2008)] for conditions under which an RKHS is characteristic. We note here that the Gaussian kernel on $\mathbb{R}^p$ has this property, and provide further discussion below.

In other words, the spectral norm of the covariance operator C_{xy} between characteristic RKHSs is zero only at independence, and is an independence statistic [Gretton et al. (2005a)]. Rather than the spectral norm, Gretton et al. (2005b) propose to use the squared Hilbert–Schmidt norm (the sum of the squared singular values), which has a population expression identical to HSIC, defined earlier. The RKHS norm implies a smoothness penalty on the functions f and g Schölkopf and Smola [(2002), Chapter 4], resulting in $O_p(n^{-1/2})$ convergence of the finite sample estimate: interestingly, this rate does *not* depend on the dimensions p and q of X and Y, respectively. Following Serfling [(1980), Chapter 5], the asymptotic distribution of the statistic under the alternative hypothesis $\mathcal{H}_1$ of dependence is Gaussian, and the distribution under the null hypothesis $\mathcal{H}_0$ of independence is an infinite weighted sum of independent χ^2 random variables; see [Gretton et al. (2008)] for details.

As long as k and l are characteristic kernels, then $\mathrm{H}(\mathbf{P}_{xy}; \mathcal{F}, \mathcal{G}) = 0$ iff X and Y are independent. The Gaussian and Laplace kernels are characteristic on $\mathbb{R}^p$ [Fukumizu et al. (2008)], and universal kernels [as defined by Steinwart, (2001)] are characteristic on compact domains [Gretton et al. (2005b), Theorem 6]. Sriperumbudur et al. (2008) provide a simple necessary and sufficient condition for a translation invariant kernel to be characteristic on $\mathbb{R}^p$: the Fourier spectrum of the kernel must be supported on the entire domain. Note that characteristic kernels need not be functions of the distance between points: an example is the kernel

$$k(x, x') = \exp(x^T x'/\sigma)$$

from Steinwart [(2001), Section 3, Example 1], which is characteristic on compact subsets of $\mathbb{R}^p$ since it is universal. Finally, an appropriate choice of kernels allows testing of dependence in non-Euclidean settings, such as distributions on groups, graphs, and strings [see, for instance, Gretton et al. (2008), who described independence testing between text fragments in English and French, where the null hypothesis was rejected when the French extracts were translations from the English].

Interestingly, the first RKHS-based independence measures were based on the canonical correlation, rather than the covariance: in this respect, they more strongly resemble the statistic $\mathcal{R}_n^2$ of SR. Dauxois and Nkiet (1998) propose the canonical correlation between variables in a spline-based RKHS as a dependence measure, using projection on a finite basis to regularize: this dependence measure follows the suggestion of Rényi (1959), but with a more restrictive pair of function classes used to compute the correlation (rather than the set of all square integrable functions). The variables are assumed in this case to be univariate. Likewise, Bach and Jordan (2002) use the canonical correlation between RKHS feature mappings as a measure of dependence between pairs of random variables. Bach and Jordan employ a different regularization strategy, however, which is a roughness penalty on the canonical correlates. For an appropriate rate of decay of this regularization with

increasing sample size, the empirical estimate of the canonical correlation converges in probability [Leurgans, Moyeed and Silverman (1993); Fukumizu, Bach and Gretton (2007)]. Finally, Fukumizu et al. (2008) provide a consistent RKHS-based estimate of the mean-square contingency, which is also based on the feature space correlation. This final independence measure is asymptotically independent of the kernel choice. When used as a statistic in an independence test, this last statistic was found empirically to have power superior to the HSIC-based test.

4. Experiments. In comparing the independence tests $\mathcal{V}_n^2$ (henceforth denoted *Dist*) and HSIC, we used an artificial benchmark proposed by Gretton et al. (2008). We tested the independence in two, four, and eight dimensions (i.e., $p \in 1, 2, 4$ and $p = q =: d$). We reproduce here the data description of Gretton *et al.* for ease of reference. First, we generated n samples of two independent univariate random variables, each drawn at random from the ICA benchmark densities of Bach and Jordan [(2002), Figure 5]: these included super-Gaussian, sub-Gaussian, multimodal, and unimodal distributions, with the common property of zero mean and unit variance. Second, we mixed these random variables using a rotation matrix parametrized by an angle θ, varying from 0 to $\pi/4$ (a zero angle meant the data were independent, while dependence became easier to detect as the angle increased to $\pi/4$; see the two plots in Figure 1). Third, in the cases $d = 2$ and $d = 4$, independent Gaussian noise of zero mean and unit variance was used to fill the remaining dimensions, and the resulting vectors were multiplied by independent random two- or four-dimensional orthogonal matrices, to obtain random vectors X and Y dependent across all observed dimensions. The resulting random variables were dependent but uncorrelated. We investigated sample sizes $n = 128, 512, 1024$, and 2048. In estimating the the test threshold (i.e., the $1 - \alpha$ quantile of the HSIC and *Dist* null distributions), we randomly permuted the Y sample ordering 200 times, and used the appropriate quantile of the resulting histogram of values. The kernel bandwidths for HSIC were set to the median distance between samples of the respective variables.[4] Note that a more sophisticated but computationally costly approach to bandwidth selection is described by Fukumizu et al. (2008), which involves matching the closed-form expression for the variance of HSIC with an estimate obtained by data shuffling.

Results are plotted in Figure 1 (average over 500 independent generations of the data). The y-intercept on these plots corresponds to the acceptance rate of the null hypothesis $\mathcal{H}_0$ of independence, or $1 -$ (Type I error), and should be close to the design parameter of $1 - \alpha = 0.95$. Elsewhere, the plots indicate acceptance of $\mathcal{H}_0$ where the alternative hypothesis $\mathcal{H}_1$ of dependence holds, that is, the Type II error.

[4]A Matlab implementation of the HSIC test, including the kernel bandwidth selection step, may be downloaded from http://www.kyb.mpg.de/bs/people/arthur/indep.htm. The software also includes a faster Gamma approximation to the null distribution.

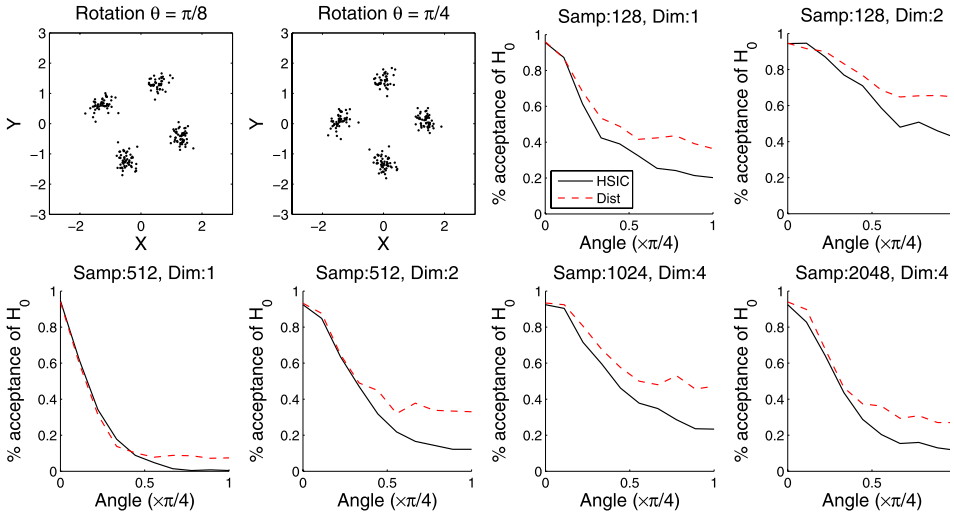

FIG. 1. Top left plots: *Example data set for* $p = q = 1$, $n = 200$, *and rotation angles* $\theta = \pi/8$ (*left*) *and* $\theta = \pi/4$ (*right*). *In this case, both sources are mixtures of two Gaussians* [*source* (g) *of Bach and Jordan* (2002), *Figure* 5]. *We remark that the random variables appear "more dependent" as the angle* θ *increases, although their correlation is always zero*. Remaining plots: *Rate of acceptance of* $\mathcal{H}_0$ *for the* Dist *and* HSIC *tests. "Samp" is the number* m *of samples, and "dim" is the dimension* d *of* x *and* y.

We observe dependence becomes easier to detect as θ increases from 0 to $\pi/4$, when n increases, and when d decreases. HSIC does as well as or better than *Dist* in all experiments, with a particular advantage at low sample sizes. In this respect, it appears that the additional smoothing employed by the RKHS approach has made the associated independence test more robust. Earlier experiments by Gretton et al. (2008) indicate that both HSIC and *Dist* outperform the power-divergence statistic of Read and Cressie (1988) on these data. This is unsurprising, since, for higher dimensions, a space partitioning approach results in too few samples per bin.

Acknowledgments. We would like to acknowledge Bernhard Schölkopf and Alexander Smola for their collaboration on several of the works referenced in this discussion.

REFERENCES

BACH, F. R. and JORDAN, M. I. (2002). Kernel independent component analysis. *J. Mach. Learn. Res.* **3** 1–48. MR1966051

BAKER, C. (1973). Joint measures and cross-covariance operators. *Trans. Amer. Math. Soc.* **186** 273–289. MR0336795

DAUXOIS, J. and NKIET, G. M. (1998). Nonlinear canonical analysis and independence tests. *Ann. Statist.* **26** 1254–1278. MR1647653

FEUERVERGER, A. (1993). A consistent test for bivariate dependence. *International Statistical Review* **61** 419–433.

FUKUMIZU, K., BACH, F. R. and JORDAN, M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *J. Mach. Learn. Res.* **5** 73–99. MR2247974

FUKUMIZU, K., BACH, F. and GRETTON, A. (2007). Statistical consistency of kernel canonical correlation analysis. *J. Mach. Learn. Res.* **8** 361–383. MR2320675

FUKUMIZU, K., BACH, F. R. and JORDANM M. I. (2009). Kernel dimension reduction in regression. *Ann. Statist.* **37** 1871–1905. MR2533474

FUKUMIZU, K., GRETTON, A., SUN, X. and SCHÖLKOPF, B. (2008). Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems* **20** 489–496. MIT Press, Cambridge, MA.

FUKUMIZU, K., SRIPERUMBUDUR, B., GRETTON, A. and SCHÖLKOPF, B. (2009). Characteristic kernels on groups and semigroups. In *Advances in Neural Information Processing Systems* **21** 473–480. Curran Associates Inc., Red Hook, NY.

GÄRTNER, T., FLACH, P. and WROBEL, S. (2003). On Graph Kernels: Hardness Results and Efficient Alternatives. In *Proc. Annual Conf. Computational Learning Theory* (B. Schölkopf and M. K. Warmuth, eds.) 129–143. Springer, Berlin.

GRETTON, A. and GYÖRFI, L. (2008). Nonparametric independence tests: Space partitioning and kernel approaches. In *Algorithmic Learning Theory: 19th International Conference* 183–198. Springer, Berlin.

GRETTON, A. and GYÖRFI, L. (2009). Consistent nonparametric tests of independence. Technical Report No. 172.

GRETTON, A., HERBRICH, R., SMOLA, A., BOUSQUET, O. and SCHÖLKOPF, B. (2005a). Kernel methods for measuring independence. *J. Mach. Learn. Res.* **6** 2075–2129. MR2249882

GRETTON, A., BOUSQUET, O., SMOLA, A. and SCHÖLKOPF, B. (2005b). Measuring statistical dependence with Hilbert–Schmidt norms. In *Algorithmic Learning Theory: 16th International Conference* (S. Jain, H. U. Simon and E. Tomita, eds.) 63–77. Springer, Berlin.

GRETTON, A., FUKUMIZU, K., TEO, C.-H., SONG, L., SCHÖLKOPF, B. and SMOLA, A. (2008). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20* 585–592. MIT Press, Cambridge, MA.

JACOD, J. and PROTTER, P. (2000). *Probability Essentials*. Springer, New York. MR1736066

KANKAINEN, A. (1995). Consistent testing of total independence based on the empirical characteristic function. Ph.D. thesis, University of Jyväskylä.

KANKAINEN, A. and USHAKOV, N. (1998). A consistent modification of a test for independence based on the empirical characteristic function. *J. Math. Sci.* **89** 1486–1494. MR1632247

LESLIE, C., ESKIN, E., WESTON, J. and NOBLE, W. S. (2002). Mismatch string kernels for SVM protein classification. In *Advances in Neural Information Processing Systems* (S. Becker, S. Thrun and K. Obermayer, eds.) **15**. MIT Press, Cambridge, MA.

LEURGANS, S. E., MOYEED, R. A. and SILVERMAN, B. W. (1993). Canonical correlation analysis when the data are curves. *J. Roy. Statist. Soc. Ser. B* **55** 725–740. MR1223939

READ, T. and CRESSIE, N. (1988). *Goodness-Of-Fit Statistics for Discrete Multivariate Analysis*. Springer, New York. MR0955054

RÉNYI, A. (1959). On measures of dependence. *Acta Math. Acad. Sci. Hungar.* **10** 441–451. MR0115203

ROSENBLATT, M. (1975). A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *Ann. Statist.* **3** 1–14. MR0428579

SCHÖLKOPF, B. and SMOLA, A. (2002). *Learning With Kernels*. MIT Press, Cambridge, MA.

SERFLING, R. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York. MR0595165

SMOLA, A. J., GRETTON, A., SONG, L. and SCHÖLKOPF, B. (2007). A hilbert space embedding for distributions. In *Proc. Intl. Conf. Algorithmic Learning Theory*. *LNAI* **4754** 13–31. Springer, Berlin.

SRIPERUMBUDUR, B., GRETTON, A., FUKUMIZU, K., LANCKRIET, G. and SCHÖLKOPF, B. (2008). Injective Hilbert space embeddings of probability measures. In *Proceedings of the 21st Annual Conference on Learning Theory* 111–122. Omnipress, Madison, WI.

STEINWART, I. (2001). On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.* **2** 67–93. MR1883281

SZÉKELY, G., RIZZO, M. and BAKIROV, N. (2007). Measuring and testing independence by correlation of distances. *Ann. Statist.* **35** 2769–2794. MR2382665

USHAKOV, N. (1999). *Selected Topics in Characteristic Functions*. *Modern Probability and Statistics*. Walter de Gruyter, Berlin. MR1745554

WENDLAND, H. (2005). *Scattered Data Approximation*. Cambridge Univ. Press, Cambridge, UK. MR2131724

A. GRETTON
MACHINE LEARNING DEPARTMENT, CMU
5000 FORBES AVE
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: arthur.gretton@gmail.com

K. FUKUMIZU
THE INSTITUTE OF STATISTICAL MATHEMATICS
4-6-7 MINAMI-AZABU, MINATO-KU
TOKYO 106-8569
JAPAN
E-MAIL: fukumizu@ism.ac.jp

B. K. SRIPERUMBUDUR
DEPARTMENT OF ELECTRICAL
AND COMPUTER ENGINEERING
UNIVERSITY OF CALIFORNIA, SAN DIEGO
LA JOLLA, CALIFORNIA 92093-0407
USA
E-MAIL: bharathsv@ucsd.edu