Measuring Statistical Dependence with Hilbert-Schmidt Norms

Arthur Gretton[1], Olivier Bousquet[2], Alex Smola[3], and Bernhard Schölkopf[1]

[1] MPI for Biological Cybernetics, Spemannstr. 38, 72076 Tübingen, Germany
{arthur, bernhard.schoelkopf}@tuebingen.mpg.de
[2] Pertinence, 32, Rue des Jeûneurs, 75002 Paris, France
olivier.bousquet@pertinence.com
[3] National ICT Australia, North Road, Canberra 0200 ACT, Australia
alex.smola@nicta.com.au

Abstract. We propose an independence criterion based on the eigenspectrum of covariance operators in reproducing kernel Hilbert spaces (RKHSs), consisting of an empirical estimate of the Hilbert-Schmidt norm of the cross-covariance operator (we term this a Hilbert-Schmidt Independence Criterion, or HSIC). This approach has several advantages, compared with previous kernel-based independence criteria. First, the empirical estimate is simpler than any other kernel dependence test, and requires no user-defined regularisation. Second, there is a clearly defined population quantity which the empirical estimate approaches in the large sample limit, with exponential convergence guaranteed between the two: this ensures that independence tests based on HSIC do not suffer from slow learning rates. Finally, we show in the context of independent component analysis (ICA) that the performance of HSIC is competitive with that of previously published kernel-based criteria, and of other recently published ICA methods.

1 Introduction

Methods for detecting dependence using kernel-based approaches have recently found application in a wide variety of areas. Examples include independent component analysis [3, 10], gene selection [20], descriptions of gait in terms of hip and knee trajectories [15], feature selection [9], and dependence detection in fMRI signals [11]. The principle underlying these algorithms is that we may define covariance and cross-covariance operators in RKHSs, and derive statistics from these operators suited to measuring the dependence between functions in these spaces.

In the method of Bach and Jordan [3], a regularised correlation operator was derived from the covariance and cross-covariance operators, and its largest singular value (the kernel canonical correlation, or KCC) was used as a statistic to test independence. The approach of Gretton *et al.* [11] was to use the largest singular value of the cross-covariance operator, which behaves identically to the

S. Jain, H.U. Simon, and E. Tomita (Eds.): ALT 2005, LNAI 3734, pp. 63–77, 2005.

correlation operator at independence, but is easier to define and requires no regularisation — the resulting test is called the constrained covariance (COCO). Both these quantities fall within the framework set out by Rényi [17], namely that for sufficiently rich function classes, the functional correlation (or, alternatively, the cross-covariance) serves as an independence test, being zero only when the random variables tested are independent. Various empirical kernel quantities (derived from bounds on the mutual information that hold near independence)[1] were also proposed based on the correlation and cross-covariance operators in [3, 10], however their connection to the population covariance operators remains to be established (indeed, the population quantities to which these approximations converge are not yet known). Gretton *et al.* [11] showed that these various quantities are guaranteed to be zero for independent random variables only when the associated RKHSs are universal [19].

The present study extends the concept of COCO by using the *entire* spectrum of the cross-covariance operator to determine when all its singular values are zero, rather than looking only at the largest singular value; the idea being to obtain a more robust indication of independence. To this end, we use the sum of the squared singular values of the cross-covariance operator (i.e., its squared *Hilbert-Schmidt norm*) to measure dependence — we call the resulting quantity the Hilbert-Schmidt Independence Criterion (HSIC).[2] It turns out that the empirical estimate of HSIC is identical to the *quadratic dependence measure* of Achard *et al.* [1], although we shall see that their derivation approaches this criterion in a completely different way. Thus, the present work resolves the open question in [1] regarding the link between the quadratic dependence measure and kernel dependence measures based on RKHSs, and generalises this measure to metric spaces (as opposed to subsets of the reals). More importantly, however, we believe our proof assures that HSIC is indeed a dependence criterion under all circumstances (i.e., HSIC is zero if and only if the random variables are independent), which is not necessarily guaranteed in [1]. We give a more detailed analysis of Achard's proof in Appendix B.

Compared with previous kernel independence measures, HSIC has several advantages:

- The empirical estimate is much simpler — just the trace of a product of Gram matrices — and, unlike the canonical correlation or kernel generalised variance [3], HSIC does not require extra regularisation terms for good finite sample behaviour.
- The empirical estimate converges to the population estimate at rate $1/\sqrt{m}$, where m is the sample size, and thus independence tests based on HSIC do not suffer from slow learning rates [8]. In particular, as the sample size increases, we are guaranteed to detect any existing dependence with high

[1] Respectively the Kernel Generalised Variance (KGV) and the Kernel Mutual Information (KMI).

[2] The possibility of using a Hilbert-Schmidt norm was suggested by Fukumizu *et al.* [9], although the idea was not pursued further in that work.

probability. Of the alternative kernel dependence tests, this result is proved only for the constrained covariance [11].
- The finite sample bias of the estimate is $O(m^{-1})$, and is therefore negligible compared to the finite sample fluctuations (which underly the convergence rate in the previous point). This is currently proved for *no* other kernel dependence test, including COCO.
- Experimental results on an ICA problem show that the new independence test is superior to the previous ones, and competitive with the best existing specialised ICA methods. In particular, kernel methods are substantially more resistant to outliers than other specialised ICA algorithms.

We begin our discussion in Section 2, in which we define the cross-covariance operator between RKHSs, and give its Hilbert-Schmidt (HS) norm (this being the population HSIC). In Section 3, we given an empirical estimate of the HS norm, and establish the link between the population and empirical HSIC by determining the bias of the finite sample estimate. In Section 4, we demonstrate exponential convergence between the population HSIC and empirical HSIC. As a consequence of this fast convergence, we show in Section 5 that dependence tests formulated using HSIC do not suffer from slow learning rates. Also in this section, we describe an efficient approximation to the empirical HSIC based on the incomplete Cholesky decomposition. Finally, in Section 6, we apply HSIC to the problem of independent component analysis (ICA).

2 Cross-Covariance Operators

In this section, we provide the functional analytic background necessary in describing cross-covariance operators between RKHSs, and introduce the Hilbert-Schmidt norm of these operators. Our presentation follows [21, 12], the main difference being that we deal with cross-covariance operators rather than the covariance operators.[3] We also draw on [9], which uses covariance and cross-covariance operators as a means of defining conditional covariance operators, but does not investigate the Hilbert-Schmidt norm; and on [4], which characterises the covariance and cross-covariance operators for general Hilbert spaces.

2.1 RKHS Theory

Consider a Hilbert space $\mathcal{F}$ of functions from $\mathcal{X}$ to $\mathbb{R}$. Then $\mathcal{F}$ is a reproducing kernel Hilbert space if for each $x \in \mathcal{X}$, the Dirac evaluation operator $\delta_x : \mathcal{F} \to \mathbb{R}$, which maps $f \in \mathcal{F}$ to $f(x) \in \mathbb{R}$, is a bounded linear functional. To each point $x \in \mathcal{X}$, there corresponds an element $\phi(x) \in \mathcal{F}$ such that $\langle \phi(x), \phi(x') \rangle_{\mathcal{F}} = k(x, x')$, where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a unique positive definite kernel. We will require in particular that $\mathcal{F}$ be separable (it must have a complete orthonormal system).

[3] Briefly, a cross-covariance operator maps from one space to another, whereas a covariance operator maps from a space to itself. In the linear algebraic case, the covariance is $C_{xx} := \mathbf{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^\top] - \mathbf{E}_{\mathbf{x}}[\mathbf{x}]\mathbf{E}_{\mathbf{x}}[\mathbf{x}^\top]$, while the cross-covariance is $C_{xy} := \mathbf{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^\top] - \mathbf{E}_{\mathbf{x}}[\mathbf{x}]\mathbf{E}_{\mathbf{y}}[\mathbf{y}^\top]$.

As pointed out in [12–Theorem 7], any continuous kernel on a separable $\mathcal{X}$ (e.g. $\mathbb{R}^n$) induces a separable RKHS.[4] We likewise define a second separable RKHS, $\mathcal{G}$, with kernel $l(\cdot,\cdot)$ and feature map ψ, on the separable space $\mathcal{Y}$.

Hilbert-Schmidt Norm. Denote by $C : \mathcal{G} \to \mathcal{F}$ a linear operator. Then provided the sum converges, the Hilbert-Schmidt (HS) norm of C is defined as

$$\|C\|_{\mathrm{HS}}^2 := \sum_{i,j} \langle Cv_i, u_j \rangle_{\mathcal{F}}^2, \tag{1}$$

where u_i and v_j are orthonormal bases of $\mathcal{F}$ and $\mathcal{G}$ respectively. It is easy to see that this generalises the Frobenius norm on matrices.

Hilbert-Schmidt Operator. A linear operator $C : \mathcal{G} \to \mathcal{F}$ is called a Hilbert-Schmidt operator if its HS norm exists. The set of Hilbert-Schmidt operators $\mathrm{HS}(\mathcal{G}, \mathcal{F}) : \mathcal{G} \to \mathcal{F}$ is a separable Hilbert space with inner product

$$\langle C, D \rangle_{\mathrm{HS}} := \sum_{i,j} \langle Cv_i, u_j \rangle_{\mathcal{F}} \langle Dv_i, u_j \rangle_{\mathcal{F}}.$$

Tensor Product. Let $f \in \mathcal{F}$ and $g \in \mathcal{G}$. Then the tensor product operator $f \otimes g : \mathcal{G} \to \mathcal{F}$ is defined as

$$(f \otimes g)h := f\langle g, h \rangle_{\mathcal{G}} \text{ for all } h \in \mathcal{G}. \tag{2}$$

Moreover, by the definition of the HS norm, we can compute the HS norm of $f \otimes g$ via

$$\begin{aligned} \|f \otimes g\|_{\mathrm{HS}}^2 = \langle f \otimes g, f \otimes g \rangle_{\mathrm{HS}} &= \langle f, (f \otimes g)g \rangle_{\mathcal{F}} \\ = \langle f, f \rangle_{\mathcal{F}} \langle g, g \rangle_{\mathcal{G}} &= \|f\|_{\mathcal{F}}^2 \|g\|_{\mathcal{G}}^2 \end{aligned} \tag{3}$$

2.2 The Cross-Covariance Operator

Mean. We assume that $(\mathcal{X}, \Gamma)$ and $(\mathcal{Y}, \Lambda)$ are furnished with probability measures p_x, p_y respectively (Γ being the Borel sets on $\mathcal{X}$, and Λ the Borel sets on $\mathcal{Y}$). We may now define the mean elements with respect to these measures as those members of $\mathcal{F}$ and $\mathcal{G}$ respectively for which

$$\begin{aligned} \langle \mu_x, f \rangle_{\mathcal{F}} &:= \mathbf{E}_x \left[\langle \phi(x), f \rangle_{\mathcal{F}} \right] = \mathbf{E}_x[f(x)], \\ \langle \mu_y, g \rangle_{\mathcal{G}} &:= \mathbf{E}_y \left[\langle \psi(y), g \rangle_{\mathcal{G}} \right] = \mathbf{E}_y[g(y)], \end{aligned} \tag{4}$$

where ϕ is the feature map from $\mathcal{X}$ to the RKHS $\mathcal{F}$, and ψ maps from $\mathcal{Y}$ to $\mathcal{G}$. Finally, $\|\mu_x\|_{\mathcal{F}}^2$ can be computed by applying the expectation twice via

$$\|\mu_x\|_{\mathcal{F}}^2 = \mathbf{E}_{x,x'} \left[\langle \phi(x), \phi(x') \rangle_{\mathcal{F}} \right] = \mathbf{E}_{x,x'}[k(x, x')]. \tag{5}$$

[4] For more detail on separable RKHSs and their properties, see [12] and references therein.

Here the expectation is taken over independent copies x, x' taken from p_x. The means μ_x, μ_y exist as long as their respective norms in $\mathcal{F}$ and $\mathcal{G}$ are bounded, which is true when the kernels k and l are bounded (since then $\mathbf{E}_{x,x'}[k(x,x')] < \infty$ and $\mathbf{E}_{y,y'}[l(y,y')] < \infty$). We are now in a position to define the cross-covariance operator.

Cross-Covariance. Following [4, 9],[5] the *cross-covariance operator* associated with the joint measure $p_{x,y}$ on $(\mathcal{X} \times \mathcal{Y}, \Gamma \times \Lambda)$ is a linear operator $C_{xy} : \mathcal{G} \to \mathcal{F}$ defined as

$$C_{xy} := \mathbf{E}_{x,y}\left[(\phi(x) - \mu_x) \otimes (\psi(y) - \mu_y)\right] = \underbrace{\mathbf{E}_{x,y}\left[\phi(x) \otimes \psi(y)\right]}_{:=\tilde{C}_{xy}} - \underbrace{\mu_x \otimes \mu_y}_{:=M_{xy}}. \quad (6)$$

Here (6) follows from the linearity of the expectation. We will use $\tilde{C}_{xy}$ and M_{xy} as the basis of our measure of dependence. Our next goal is to derive the Hilbert-Schmidt norm of the above quantity; the conditions under which C_{xy} is a HS operator will then follow from the existence of the norm.

2.3 Hilbert-Schmidt Independence Criterion

Definition 1 (HSIC). *Given separable RKHSs $\mathcal{F}, \mathcal{G}$ and a joint measure p_{xy} over $(\mathcal{X} \times \mathcal{Y}, \Gamma \times \Lambda)$, we define the Hilbert-Schmidt Independence Criterion (HSIC) as the squared HS-norm of the associated cross-covariance operator C_{xy}:*

$$\mathrm{HSIC}(p_{xy}, \mathcal{F}, \mathcal{G}) := \|C_{xy}\|^2_{\mathrm{HS}}. \quad (7)$$

To compute it we need to express HSIC in terms of kernel functions. This is achieved by the following lemma:

Lemma 1 (HSIC in terms of kernels).

$$\begin{aligned}\mathrm{HSIC}(p_{xy}, \mathcal{F}, \mathcal{G}) = \mathbf{E}_{x,x',y,y'}[k(x,x')l(y,y')] + \mathbf{E}_{x,x'}[k(x,x')]\mathbf{E}_{y,y'}[l(y,y')] \quad (8)\\ -2\mathbf{E}_{x,y}\left[\mathbf{E}_{x'}[k(x,x')]\mathbf{E}_{y'}[l(y,y')]\right]\end{aligned}$$

Here $\mathbf{E}_{x,x',y,y'}$ denotes the expectation over independent pairs (x, y) and (x', y') drawn from p_{xy}. This lemma is proved in Appendix A. It follows from Lemma 8 that the HS norm of C_{xy} exists when the various expectations over the kernels are bounded, which is true as long as the kernels k and l are bounded.

3 Empirical Criterion

In order to show that HSIC is a practical criterion for testing independence, and to obtain a formal independence test on the basis of HSIC, we need to perform three more steps. First, we need to approximate $\mathrm{HSIC}(p_{xy}, \mathcal{F}, \mathcal{G})$ given a finite

[5] Our operator (and that in [9]) differs from Baker's in that Baker defines all measures directly on the function spaces.

number of observations. Second, we need to show that this approximation converges to HSIC sufficiently quickly. Third, we need to show that HSIC is, indeed, an indicator for the independence of random variables (subject to appropriate choice of $\mathcal{F}$ and $\mathcal{G}$). We address the first step in this section, and the remaining two steps in the two sections that follow.

3.1 Estimator of HSIC

Definition 2 (Empirical HSIC). *Let $Z := \{(x_1, y_1), \ldots, (x_m, y_m)\} \subseteq \mathcal{X} \times \mathcal{Y}$ be a series of m independent observations drawn from p_{xy}. An estimator of HSIC, written $\mathrm{HSIC}(Z, \mathcal{F}, \mathcal{G})$, is given by*

$$\mathrm{HSIC}(Z, \mathcal{F}, \mathcal{G}) := (m-1)^{-2}\mathbf{tr}KHLH \tag{9}$$

where $H, K, L \in \mathbb{R}^{m\times m}$, $K_{ij} := k(x_i, x_j), L_{ij} := l(y_i, y_j)$ and $H_{ij} := \delta_{ij} - m^{-1}$.

An advantage of $\mathrm{HSIC}(Z, \mathcal{F}, \mathcal{G})$ is that it can be computed in $O(m^2)$ time, whereas other kernel methods cost at least $O(m^3)$ before approximations are made (although in practice, this advantage is somewhat academic, since good approximations to all kernel dependence criteria can be computed in similar time: see [3–Section 4] and Section 5.2). What we now need to show is that it is indeed related to $\mathrm{HSIC}(p_{xy}, \mathcal{F}, \mathcal{G})$:

Theorem 1 ($O(m^{-1})$ Bias of Estimator). *Let $\mathbf{E}_Z$ denote the expectation taken over m independent copies (x_i, y_i) drawn from p_{xy}. Then*

$$\mathrm{HSIC}(p_{xy}, \mathcal{F}, \mathcal{G}) = \mathbf{E}_Z\left[\mathrm{HSIC}(Z, \mathcal{F}, \mathcal{G})\right] + O(m^{-1}).$$

This means that if the variance of $\mathrm{HSIC}(Z, \mathcal{F}, \mathcal{G})$ is larger than $O(m^{-1})$ (and indeed, the uniform convergence bounds we derive with respect to $p_{x,y}$ will be $O(m^{-1/2})$), the bias arising from the definition of $\mathrm{HSIC}(Z, \mathcal{F}, \mathcal{G})$ is negligible in the overall process. The proof is in Appendix A.

4 Large Deviation Bounds

As a next step we need to show that the deviation between $\mathrm{HSIC}[Z, \mathcal{F}, \mathcal{G}]$ and its expectation is not too large. This section repeatedly uses a bound from [13–p.25], which applies to U-statistics of the form we encounter in the previous section.

Theorem 2 (Deviation bound for U-statistics). *A one-sample U-statistic is defined as the random variable*

$$u := \tfrac{1}{(m)_r}\sum_{\mathbf{i}_r^m} g(x_{i_1}, \ldots, x_{i_r}),$$

where g is called the kernel of the U-statistic.[6] *If $a \le g \le b$, then for all $t > 0$ the following bound holds:*

$$\mathbf{P}_u\left\{u - \mathbf{E}_u[u] \ge t\right\} \le \exp\left(-\frac{2t^2\lceil m/r\rceil}{(b-a)^2}\right).$$

We now state our main theorem.The proof is in Appendix A.

[6] We denote $(m)_n := \frac{m!}{(m-n)!}$.

Theorem 3 (Bound on Empirical HSIC). *Assume that k and l are bounded almost everywhere by 1, and are non-negative. Then for $m > 1$ and all $\delta > 0$, with probability at least $1 - \delta$, for all $p_{x,y}$,*

$$|\mathrm{HSIC}(p_{xy}, \mathcal{F}, \mathcal{G}) - \mathrm{HSIC}(Z, \mathcal{F}, \mathcal{G})| \leq \sqrt{\frac{\log(6/\delta)}{\alpha^2 m}} + \frac{C}{m},$$

where $\alpha^2 > 0.24$ and C are constants.

5 Independence Tests Using HSIC

In this section, we describe how HSIC can be used as an independence measure, and as the basis for an independence test. We also describe an approximation to HSIC which is more efficient to compute. We begin by demonstrating that the Hilbert-Schmidt norm can be used as a measure of independence, as long as the associated RKHSs are universal [19].

Theorem 4 (C_{xy} and Independence). *Denote by $\mathcal{F}, \mathcal{G}$ RKHSs with universal kernels k, l on the compact domains $\mathcal{X}$ and $\mathcal{Y}$ respectively. We assume without loss of generality that $\|f\|_\infty \leq 1$ and $\|g\|_\infty \leq 1$ for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$. Then $\|C_{xy}\|_{\mathrm{HS}} = 0$ if and only if x and y are independent.*

Proof. According to Gretton *et al.* [11], the largest singular value (i.e., the *spectral norm*) $\|C_{xy}\|_{\mathrm{S}}$ is zero if and only if x and y are independent, under the conditions specified in the theorem. Since $\|C_{xy}\|_{\mathrm{S}} = 0$ if and only if $\|C_{xy}\|_{\mathrm{HS}} = 0$, it follows that $\|C_{xy}\|_{\mathrm{HS}} = 0$ if and only if x and y are independent. ■

5.1 Independence Tests

We now describe how to use HSIC as the basis of an independence test. Consider a set $\mathcal{P}$ of probability distributions $p_{x,y}$. We may decompose $\mathcal{P}$ into two subsets: $\mathcal{P}_i$ contains distributions $p^{(i)}_{x,y}$ under which x and y are independent, and $\mathcal{P}_d$ contains distributions $p^{(d)}_{x,y}$ under which x and y are dependent.

We next introduce a test $\Delta(Z)$, which takes a data set $Z \sim p_Z$, where p_Z is the distribution corresponding to m independent draws from $p_{x,y}$, and returns

$$\Delta(Z) = \begin{cases} 1 & \text{if } Z \sim p^{(d)}_Z \\ 0 & \text{if } Z \sim p^{(i)}_Z \end{cases}$$

Given that the test sees only a finite sample, it cannot determine with complete certainty from which class of distributions the data are drawn. We call Δ an α-test when

$$\sup_{p^{(i)}_{x,y} \in \mathcal{P}_i} \mathbf{E}_{Z \sim p^{(i)}_Z}[\Delta(Z) = 1] \leq \alpha.$$

In other words α upper bounds the probability of a Type I error. It follows from Theorem 3 that the empirical HSIC converges to the population HSIC at speed

$1/\sqrt{m}$. This means that if we define the independence test $\Delta(Z)$ as the indicator that HSIC is larger than a term of the form $C\sqrt{\log(1/\alpha)/m}$, with C a suitable constant, then $\Delta(Z)$ is an α-test with Type II error upper bounded by a term approaching zero as $1/\sqrt{m}$.

5.2 Efficient Computation

Computational cost is another factor in using HSIC as an independence criterion. As in [3], we use a low rank decomposition of the Gram matrices via an incomplete Cholesky decomposition, which permits an accurate approximation to HSIC as long as the kernel has a fast decaying spectrum. This results in the following cost saving, which we use in our experiments. The proof is in Appendix A.

Lemma 2 (Efficient approximation to HSIC). *Let $K \approx AA^\top$ and $L \approx BB^\top$, where $A \in \mathbb{R}^{m\times d_f}$ and $B \in \mathbb{R}^{m\times d_g}$. Then we may approximate $\mathbf{tr}HKHL$ in $O(m(d_f^2 + d_g^2))$ time.*

Finally, note that although the present measure of dependence pertains only to the two-variable case, a test of pairwise dependence for a greater number of variables may easily be defined by summing HSIC over every pair of variables — this quantity vanishes if and only if the random variables are pairwise independent. We use this generalisation in the experiments of Section 6.

6 Experimental Results

We apply our estimates of statistical dependence to the problem of linear instantaneous independent component analysis [14]. In this setting, we assume a random source vector $\mathbf{s}$ of dimension n, where $s_i \in \mathbb{R}$, such that the components are mutually independent; $p_{\mathbf{s}}(\mathbf{s}) = \prod_{i=1}^n p_{s_i}(s_i)$. We observe a vector $\mathbf{t}$ that corresponds to a linear mixing $\mathbf{t} = \mathbf{A}\mathbf{s}$ of the sources $\mathbf{s}$, where $\mathbf{A}$ is an $n \times n$ matrix with full rank.[7] We wish to recover an estimate $\mathbf{x}$ of the unmixed elements $\mathbf{s}$ given m i.i.d. samples from $p_{\mathbf{t}}(\mathbf{t})$, and using *only* the linear mixing model and the fact that the unmixed components are independent. This problem is indeterminate in certain respects: for instance, the ordering and scale of the sources cannot be recovered using independence alone.

It is clear that the various cross-covariance based kernel dependence tests, including HSIC, can each be used to determine when the inverse $\mathbf{V}$ of $\mathbf{A}$ is found,[8] by testing the pairwise independence of the components in $\mathbf{x} = \mathbf{V}\mathbf{t}$ (bearing in mind Theorem 4 and its implications for the various kernel dependence tests). This requires a gradient descent procedure in which the kernel contrasts are minimised as a function of $\mathbf{V}$; see [3, 10] for details. The Amari divergence [2],

[7] This assumes the number of sources is equal to the number of sensors, and the sources are spatially distinct.

[8] Up to permutation and scaling, and assuming no more than one source is Gaussian [14].

which is invariant to permutation and scaling, is used to compare $\mathbf{V}$ and $\mathbf{A}^{-1}$. We acknowledge that the application of a general dependence function to linear ICA is not an optimal non-parametric approach to the problem of estimating the entries in $\mathbf{A}$, as discussed in [18]. Indeed, most specialised ICA algorithms exploit the linear mixing structure of the problem to avoid having to conduct a general test of independence, which makes the task of recovering $\mathbf{A}$ easier. That said, ICA is in general a good benchmark for dependence measures, in that it applies to a problem with a known "ground truth", and tests that the dependence measures approach zero gracefully as dependent random variables are made to approach independence (through optimisation of the unmixing matrix).

As well as the kernel algorithms, we also compare with three standard ICA methods (FastICA [14], Jade [6], and Infomax [5]); and two recent state of the art methods, neither of them based on kernels: RADICAL [16], which uses order statistics to obtain entropy estimates; and characteristic function based ICA (CFICA) [7].[9] It was recommended to run the CFICA algorithm with a good initialising guess; we used RADICAL for this purpose. All kernel algorithms were initialised using Jade (except for the 16 source case, where Fast ICA was used due to its more stable output). RADICAL is based on an exhaustive grid search over all the Jacobi rotations, and does not require an initial guess.

Table 1. Densities used, and their respective kurtoses. Densities have zero mean and unit variance.

Density	Kurtosis
Student, 3 DOF	∞
Double exponential	3.00
Uniform	-1.20
Student, 5 DOF	6.00
Exponential	6.00
2 double exponentials	-1.70
Symmetric. 2 Gaussians, multimodal	-1.85
As above, transmodal	-0.75
As above, unimodal	-0.50
Asymmetric. 2 Gaussians, multimodal	-0.57
As above, transmodal	-0.29
As above, unimodal	-0.20
Symmetric. 4 Gaussians, multimodal	-0.91
As above, transmodal	-0.34
As above, unimodal	-0.40
Asymmetric. 4 Gaussians, multimodal	-0.67
As above, transmodal	-0.59
As above, unimodal	-0.82

Our first experiment consisted in demixing data drawn independently from several distributions chosen at random with replacement from Table 1, and mixed with a random matrix having condition number between 1 and 2. In the case of the KCC and KGV, we use the parameters recommended in [3]: namely, $\kappa = 2 \times 10^{-2}$ and $\sigma = 1$ for $m \leq 1000$, $\kappa = 2 \times 10^{-3}$ and $\sigma = 0.5$ for $m > 1000$ (σ being the kernel size, and κ the coefficient used to scale the regularising terms). In the case of our dependence tests (COCO, KMI, HSIC), we used $\sigma = 1$ for the Gaussian kernel, and $\sigma = 3$ for the Laplace kernel. After convergence, the kernel size was halved for all methods, and the solution refined in a "polishing" step. Results are given in Table 2.

[9] We are aware that the same authors propose an alternative algorithm, "Efficient ICA". We did not include results from this algorithm in our experiments, since it is unsuited to mixtures of Gaussians (which have fast decaying tails) and discontinuous densities (such as the uniform density on a finite interval), which both occur in our benchmark set.

We note that HSIC with a Gaussian kernel performs on par with the best alternatives in the final four experiments, and that HSIC with a Laplace kernel gives joint best performance in six of the seven experiments. On the other hand, RADICAL and the KGV perform better than HSIC in the $m = 250$ case. While the Laplace kernel clearly gives superior performance, this comes at an increased computational cost, since the eigenvalues of the associated Gram matrices decay more slowly than for the Gaussian kernel, necessitating the use of a higher rank in the incomplete Cholesky decomposition. Interestingly, the Laplace kernel can improve on the Gaussian kernel even with sub-Gaussian sources, as seen for instance in [10–Table 6.3] for the KMI and COCO.[10] This is because the slow decay of the eigenspectrum of the Laplace kernel improves the detection of dependence encoded at higher frequencies in the probability density function, which need not be related to the kurtosis — see [11–Section 4.2].

Table 2. Demixing of n randomly chosen i.i.d. samples of length m, where n varies from 2 to 16. The Gaussian kernel results are denoted *g*, and the Laplace kernel results *l*. The column *Rep.* gives the number of runs over which the average performance was measured. Note that some algorithm names are truncated: Fica is Fast ICA, IMAX is Infomax, RAD is RADICAL, CFIC is CFICA, CO is COCO, and HS is HSIC. Performance is measured using the Amari divergence (smaller is bettter).

n	m	Rep.	FICA	Jade	IMAX	RAD	CFIC	KCC	COg	COl	KGV	KMIg	KMIl	HSg	HSl
2	250	1000	10.5± 0.4	9.5 ± 0.4	44.4± 0.9	**5.4 ± 0.2**	7.2 ± 0.3	7.0 ± 0.3	7.8 ± 0.3	7.0 ± 0.3	**5.3 ± 0.2**	6.0 ± 0.2	5.7 ± 0.2	5.9 ± 0.2	5.8 ± 0.3
2	1000	1000	6.0 ± 0.3	5.1 ± 0.2	11.3± 0.6	**2.4 ± 0.1**	3.2 ± 0.1	3.3 ± 0.1	3.5 ± 0.1	2.9 ± 0.1	**2.3 ± 0.1**	2.6 ± 0.1	**2.3 ± 0.1**	2.6 ± 0.1	**2.4 ± 0.1**
4	1000	100	5.7 ± 0.4	5.6 ± 0.4	13.3± 1.1	**2.5 ± 0.1**	3.3 ± 0.2	4.5 ± 0.4	4.2 ± 0.3	4.6 ± 0.6	3.1 ± 0.6	4.0 ± 0.7	3.5 ± 0.7	2.7 ± 0.1	**2.5 ± 0.2**
4	4000	100	3.1 ± 0.2	2.3 ± 0.1	5.9 ± 0.7	**1.3 ± 0.1**	1.5 ± 0.1	2.4 ± 0.5	1.9 ± 0.1	1.6 ± 0.1	1.4 ± 0.1	1.4 ± 0.05	**1.2 ± 0.05**	**1.3 ± 0.05**	**1.2 ± 0.05**
8	2000	50	4.1 ± 0.2	3.6 ± 0.2	9.3 ± 0.9	**1.8 ± 0.1**	2.4 ± 0.1	4.8 ± 0.9	3.7 ± 0.9	5.2 ± 1.3	2.6 ± 0.3	2.1 ± 0.1	**1.9 ± 0.1**	**1.9 ± 0.1**	**1.8 ± 0.1**
8	4000	50	3.2 ± 0.2	2.7 ± 0.1	6.4 ± 0.9	**1.3 ± 0.05**	1.6 ± 0.1	2.1 ± 0.2	2.0 ± 0.1	1.9 ± 0.1	1.7 ± 0.2	**1.4 ± 0.1**	**1.3 ± 0.05**	**1.4 ± 0.05**	**1.3 ± 0.05**
16	5000	25	2.9 ± 0.1	3.1 ± 0.3	9.4 ± 1.1	**1.2 ± 0.05**	1.7 ± 0.1	3.7 ± 0.6	2.4 ± 0.1	2.6 ± 0.2	1.7 ± 0.1	1.5 ± 0.1	1.5 ± 0.1	**1.3 ± 0.05**	**1.3 ± 0.05**

In our next experiment, we investigated the effect of outlier noise added to the observations. We selected two generating distributions from Table 1, randomly and with replacement. After combining $m = 1000$ samples from these distributions with a randomly generated matrix having condition number between 1 and 2, we generated a varying number of outliers by adding ± 5 (with equal probability) to *both* signals at random locations. All kernels used were Gaussian with size $\sigma = 1$; Laplace kernels resulted in decreased performance for this noisy data. Results are shown in Figure 1. Note that we used $\kappa = 0.11$ for the KGV and KCC in this plot, which is an order of magnitude above the level recommended in [3]: this resulted in an improvement in performance (broadly

[10] COCO is referred to in this table as KC.

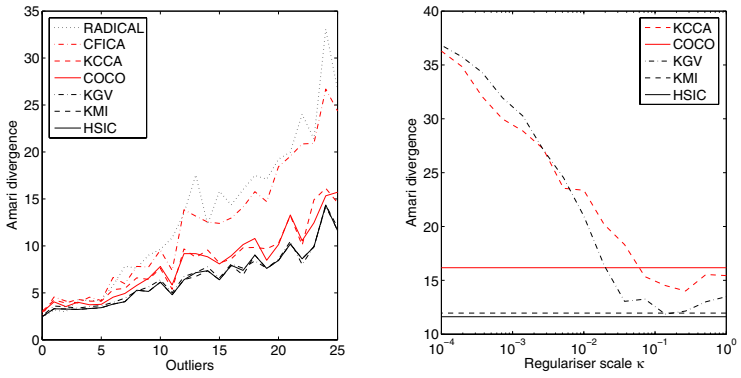

Fig. 1. Left: Effect of outliers on the performance of the ICA algorithms. Each point represents an average Amari divergence over 100 independent experiments (smaller is better). The number of corrupted observations in *both* signals is given on the horizontal axis. **Right:** Performance of the KCC and KGV as a function of κ for two sources of size $m = 1000$, where 25 outliers were added to each source following the mixing procedure.

speaking, an increase in κ causes the KGV to approach the KMI, and the KCC to approach COCO [10]).[11]

An additional experiment was also carried out on the same data, to test the sensitivity of the KCC and KGV to the choice of the regularisation constant κ. We observe in Figure 1 that too small a κ can cause severe underperformance for the KCC and KGV. On the other hand, κ is required to be small for good performance at large sample sizes in Table 2. A major advantage of HSIC, COCO, and the KMI is that these do not require any additional tuning beyond the selection of a kernel.

In conclusion, we emphasise that ICA based on HSIC, despite using a more general dependence test than in specialised ICA algorithms, nonetheless gives joint best performance on all but the smallest sample size, and is much more robust to outliers. Comparing with other kernel algorithms (which are also based on general dependence criteria), HSIC is simpler to define, requires no regularisation or tuning beyond kernel selection, and has performance that meets or exceeds the best alternative on all data sets besides the $m = 250$ case.

Acknowledgements. The authors would like to thank Kenji Fukumizu and Matthias Hein for helpful discussions. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. National ICT Australia is funded through the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council.

[11] The results presented here for the KCC and KGV also improve on those in [16, 3] since they include a polishing step for the KCC and KGV, which was not carried out in these earlier studies.

References

[1] S. Achard, D.-T. Pham, and C. Jutten, *Quadratic dependence measure for nonlinear blind source separation*, 4th International Conference on ICA and BSS, 2003.

[2] S.-I. Amari, A. Cichoki, and Yang H., *A new learning algorithm for blind signal separation*, Advances in Neural Information Processing Systems, vol. 8, MIT Press, 1996, pp. 757–763.

[3] F. Bach and M. Jordan, *Kernel independent component analysis*, Journal of Machine Learning Research **3** (2002), 1–48.

[4] C. R. Baker, *Joint measures and cross-covariance operators*, Transactions of the American Mathematical Society **186** (1973), 273–289.

[5] A. Bell and T. Sejnowski, *An information-maximization approach to blind separation and blind deconvolution*, Neural Computation **7** (1995), no. 6, 1129–1159.

[6] J.-F. Cardoso, *Blind signal separation: statistical principles*, Proceedings of the IEEE **90** (1998), no. 8, 2009–2026.

[7] A. Chen and P. Bickel, *Consistent independent component analysis and prewhitening*, Tech. report, Berkeley, 2004.

[8] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*, Applications of mathematics, vol. 31, Springer, New York, 1996.

[9] K. Fukumizu, F. R. Bach, and M. I. Jordan, *Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces*, Journal of Machine Learning Research **5** (2004), 73–99.

[10] A. Gretton, R. Herbrich, and A. Smola, *The kernel mutual information*, Tech. report, Cambridge University Engineering Department and Max Planck Institute for Biological Cybernetics, 2003.

[11] A. Gretton, A. Smola, O. Bousquet, R. Herbrich, A. Belitski, M. Augath, Y. Murayama, J. Pauls, B. Schölkopf, and N. Logothetis, *Kernel constrained covariance for dependence measurement*, AISTATS, vol. 10, 2005.

[12] M. Hein and O. Bousquet, *Kernels, associated structures, and generalizations*, Tech. Report 127, Max Planck Institute for Biological Cybernetics, 2004.

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

[14] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*, John Wiley and Sons, New York, 2001.

[15] S. E. Leurgans, R. A. Moyeed, and B. W. Silverman, *Canonical correlation analysis when the data are curves*, Journal of the Royal Statistical Society, Series B (Methodological) **55** (1993), no. 3, 725–740.

[16] E. Miller and J. Fisher III, *ICA using spacings estimates of entropy*, JMLR **4** (2003), 1271–1295.

[17] A. Rényi, *On measures of dependence*, Acta Math. Acad. Sci. Hungar. **10** (1959), 441–451.

[18] A. Samarov and A. Tsybakov, *Nonparametric independent component analysis*, Bernoulli **10** (2004), 565–582.

[19] I. Steinwart, *On the influence of the kernel on the consistency of support vector machines*, JMLR **2** (2001).

[20] Y. Yamanishi, J.-P. Vert, and M. Kanehisa, *Heterogeneous data comparison and gene selection with kernel canonical correlation analysis*, Kernel Methods in Computational Biology (Cambridge, MA) (B. Schölkopf, K. Tsuda, and J.-P. Vert, eds.), MIT Press, 2004, pp. 209–229.

[21] L. Zwald, O. Bousquet, and G. Blanchard, *Statistical properties of kernel principal component analysis*, Proceedings of the 17th Conference on Computational Learning Theory (COLT), 2004.

A Proofs

A.1 Proof of Lemma 1

We expand C_{xy} via (6) and using (3):

$$\begin{aligned}\|C_{xy}\|_{\mathrm{HS}}^2 &= \langle \tilde{C}_{xy} - M_{xy}, \tilde{C}_{xy} - M_{xy} \rangle_{\mathrm{HS}} \\ &= \mathbf{E}_{x,y,x',y'} \left[\langle \phi(x) \otimes \psi(y), \phi(x) \otimes \psi(y) \rangle_{\mathrm{HS}} \right] \\ &\quad -2\mathbf{E}_{x,y} \left[\langle \mu_x \otimes \mu_y, \phi(x) \otimes \psi(y) \rangle_{\mathrm{HS}} \right] + \langle \mu_x \otimes \mu_y, \mu_x \otimes \mu_y \rangle_{\mathrm{HS}}\end{aligned}$$

Substituting the definition of μ_x and μ_y and using that $\langle \phi(x), \phi(x') \rangle_{\mathcal{F}} = k(x, x')$ (and likewise for $l(y, y')$) proves the claim.

A.2 Proof of Theorem 1

The idea underlying this proof is to expand $\mathbf{tr} HKHL$ into terms depending on pairs, triples, and quadruples (i, j), (i, j, q) and (i, j, q, r) of non-repeated terms, for which we can apply uniform convergence bounds with U-statistics.

By definition of H we can write

$$\mathbf{tr} KHLH = \underbrace{\mathbf{tr} KL}_{(a)} - 2m^{-1} \underbrace{\mathbf{1}^\top KL\mathbf{1}}_{(b)} + m^{-2} \underbrace{\mathbf{tr} K \mathbf{tr} L}_{(c)}$$

where $\mathbf{1}$ is the vector of all ones, since $H = \mathbf{1} - m^{-1}\mathbf{1}\mathbf{1}^\top$ and since K, L are symmetric. We now expand each of the terms separately and take expectations with respect to Z.

For notational convenience we introduce the Pochhammer symbol $(m)_n := \frac{m!}{(m-n)!}$. One may check that $\frac{(m)_n}{m^n} = 1 + O(m^{-1})$. We also introduce the index set $\mathbf{i}_r^m$, which is the set of all r-tuples drawn without replacement from $\{1, \ldots, m\}$.

(a) We expand $\mathbf{E}_Z[\mathbf{tr} KL]$ into

$$\mathbf{E}_Z \left[\sum_i K_{ii} L_{ii} + \sum_{(i,j) \in \mathbf{i}_2^m} K_{ij} L_{ji} \right] = O(m) + (m)_2 \, \mathbf{E}_{x,y,x',y'} \left[k(x, x') l(y, y') \right] \tag{10}$$

Normalising terms by $\frac{1}{(m-1)^2}$ yields the first term in (8), since $\frac{m(m-1)}{(m-1)^2} = 1 + O(m^{-1})$.

(b) We expand $\mathbf{E}_Z[\mathbf{1}^\top KL\mathbf{1}]$ into

$$\begin{aligned}&\mathbf{E}_Z \left[\sum_i K_{ii} L_{ii} + \sum_{(i,j) \in \mathbf{i}_2^m} (K_{ii} L_{ij} + K_{ij} K_{jj}) \right] + \mathbf{E}_Z \left[\sum_{(i,j,r) \in \mathbf{i}_3^m} K_{ij} L_{jr} \right] \\ &= O(m^2) + (m)_3 \, \mathbf{E}_{x,y} \left[\mathbf{E}_{x'}[k(x, x')] \mathbf{E}_{y'}[l(y, y')] \right]\end{aligned}$$

Again, normalising terms by $\frac{2}{m(m-1)^2}$ yields the second term in (8). As with (a) we used that $\frac{m(m-1)(m-2)}{m(m-1)^2} = 1 + O(m^{-1})$.

(c) As before we expand $\mathbf{E}_Z[\mathbf{tr}K\mathbf{tr}L]$ into terms containing varying numbers of identical indices. By the same argument we obtain

$$O(m^3) + \mathbf{E}_Z\left[\sum_{(i,j,q,r)\in \mathbf{i}_4^m} K_{ij}L_{qr}\right] = O(m^3) + (m)_4\, \mathbf{E}_{x,x'}[k(x,x')]\mathbf{E}_{y,y'}[l(y,y')]. \tag{11}$$

Normalisation by $\frac{1}{m^2(m-1)^2}$ takes care of the last term in (8), which completes the proof.

A.3 Proof of Theorem 3

As in the proof in Appendix A.2, we deal separately with each of the three terms in (8), omitting for clarity those terms that decay as $O(m^{-1})$ or faster.[12] Denote by $\mathbf{P}_Z$ the probability with respect to m independent copies (x_i, y_i) drawn from p_{xy}. Moreover, we split t into $\alpha t + \beta t + (1-\alpha-\beta)t$ where $\alpha, \beta > 0$ and $\alpha + \beta < 1$. The probability of a positive deviation t has bound

$$\begin{aligned}
&\mathbf{P}_Z\{\mathrm{HSIC}(Z,\mathcal{F},\mathcal{G}) - \mathrm{HSIC}(p_{xy},\mathcal{F},\mathcal{G}) \geq t\} \\
&\leq \mathbf{P}_Z\left\{\mathbf{E}_{x,y,x',y'}[k(x,x')l(y,y')] - \frac{1}{(m)_2}\sum_{\mathbf{i}_2^m} K_{i_1i_2}L_{i_1i_2} \geq \alpha t\right\} \\
&\quad + \mathbf{P}_Z\left\{\mathbf{E}_{x,y}[\mathbf{E}_{x'}[k(x,x')]\mathbf{E}_{y'}[l(y,y')]] - \frac{1}{(m)_3}\sum_{\mathbf{i}_3^m} K_{i_1i_2}L_{i_2i_3} \geq \frac{\beta}{2}t\right\} \\
&\quad + \mathbf{P}_Z\left\{\mathbf{E}_{x,x'}[k(x,x')]\mathbf{E}_{y,y'}[l(y,y')] - \frac{1}{(m)_4}\sum_{\mathbf{i}_4^m} K_{i_1i_2}L_{i_3i_4} \geq \frac{1-\alpha-\beta}{t}\right\}
\end{aligned}$$

Using the shorthand $z := (x, y)$ we define the kernels of the U-statistics in the three expressions above as $g(z_i, z_j) = K_{ij}L_{ij}$, $g(z_i, z_j, z_r) = K_{ij}L_{jr}$ and $g(z_i, z_j, z_q, z_r) = K_{ij}L_{qr}$. Finally, employing Theorem 2 allows us to bound the three probabilities as

$$e^{-2mt^2\frac{\alpha^2}{2}}, e^{-2mt^2\frac{\beta^2}{3\times 4}}, \text{ and } e^{-2mt^2\frac{(1-\alpha-\beta)^2}{4}},$$

Setting the argument of all three exponentials equal yields $\alpha^2 > 0.24$: consequently, the positive deviation probability is bounded from above by $3e^{-\alpha^2 mt^2}$. The bound in Theorem 2 also holds for deviations in the opposite direction, thus the overall probability is bounded by doubling this quantity. Solving for t yields the desired result.

[12] These terms are either sample means or U-statistics scaled as m^{-1} or worse, and are thus guaranteed to converge at rate $m^{-1/2}$ according to reasoning analogous to that employed below. Thus, we incorporate them in the C/m term.

A.4 Proof of Lemma 2

Computing A and B costs $O(md_f^2)$ and $O(md_g^2)$ time respectively. Next note that

$$\begin{aligned}\mathbf{tr} H(AA^\top)H(BB^\top) &= \mathbf{tr}\left(B^\top(HA)\right)\left(B^\top(HA)\right)^\top \\ &= \|(HA)^\top B\|_{\mathrm{HS}}^2\end{aligned}$$

Here computing (HA) costs $O(md_f)$ time. The dominant term in the remainder is the matrix-matrix multiplication at $O(md_fd_g)$ cost. Hence we use

$$\widetilde{HSIC}(Z;\mathcal{F},\mathcal{G}) := (m-1)^{-2}\|(HA)^\top B\|_{\mathrm{HS}}^2.$$

B HSIC Derivation of Achard *et al.*

Achard *et al.* [1] motivate using HSIC to test independence by associating the empirical HSIC with a particular population quantity, which they claim is zero if and only if the random variables being tested are independent. We now examine their proof of this assertion. The derivation begins with [1–Lemma 2.1], which states the components x_i of the random vector $\mathbf{x}$ are mutually independent if and only if

$$\mathbf{E}_{\mathbf{x}}\left[\prod_{i=1}^n k(x_i - y_i)\right] = \prod_{i=1}^n \left[\mathbf{E}_{x_i} k(x_i - y_i)\right] \quad \forall y_1, \ldots, y_m, \tag{12}$$

as long as the kernel k has Fourier transform everywhere non-zero (here y_i are real valued offset terms). Achard *et al.* claim that testing the above is equivalent to testing whether $Q(\mathbf{x}) = 0$, where

$$Q(\mathbf{x}) = \frac{1}{2}\int\left(\mathbf{E}_{\mathbf{x}}\left[\prod_{i=1}^n k\left(\frac{x_i}{\sigma_i} - y_i\right)\right] - \prod_{i=1}^n\left[\mathbf{E}_{x_i} k\left(\frac{x_i}{\sigma_i} - y_i\right)\right]\right)^2 dy_1 \ldots dy_n, \tag{13}$$

for scale factors $\sigma_i > 0$ (the empirical HSIC is then recovered by replacing the population expectations with their empirical counterparts, and some additional manipulations). However $Q(\mathbf{x}) = 0$ tells us only that (12) holds almost surely, whereas a test of independence requires (12) to hold pointwise. In other words, $Q(\mathbf{x}) = 0$ does not imply $\mathbf{x}$ are mutually independent, even though mutual independence implies $Q(\mathbf{x}) = 0$.