Kernel Methods for Measuring Independence

Arthur Gretton ARTHUR@TUEBINGEN.MPG.DE
MPI for Biological Cybernetics, Spemannstr 38, 72076, Tübingen, Germany

Ralf Herbrich RHERB@MICROSOFT.COM
Microsoft Research Cambridge, 7 J. J. Thomson Avenue, Cambridge CB3 0FB, United Kingdom

Alexander Smola ALEX.SMOLA@NICTA.COM.AU
National ICT Australia, Canberra, ACT 0200, Australia

Olivier Bousquet OLIVIER.BOUSQUET@PERTINENCE.COM
Pertinence, 32, Rue des Jeûneurs, 75002 Paris, France

Bernhard Schölkopf BERNHARD.SCHOELKOPF@TUEBINGEN.MPG.DE
MPI for Biological Cybernetics, Spemannstr 38, 72076, Tübingen, Germany

Editor: A. Hyvärinen

Abstract

We introduce two new functionals, the constrained covariance and the kernel mutual information, to measure the degree of independence of random variables. These quantities are both based on the covariance between functions of the random variables in reproducing kernel Hilbert spaces (RKHSs). We prove that when the RKHSs are universal, both functionals are zero if and only if the random variables are pairwise independent. We also show that the kernel mutual information is an upper bound near independence on the Parzen window estimate of the mutual information. Analogous results apply for two correlation-based dependence functionals introduced earlier: we show the kernel canonical correlation and the kernel generalised variance to be independence measures for universal kernels, and prove the latter to be an upper bound on the mutual information near independence. The performance of the kernel dependence functionals in measuring independence is verified in the context of independent component analysis.

Keywords: independence, covariance operator, mutual information, kernel, parzen window estimate, independent component analysis

1. Introduction

Measures to determine the dependence or independence of random variables are well established in statistical analysis. For instance, one well known measure of statistical dependence between two random variables is the *mutual information* (Cover and Thomas, 1991), which for random vectors $\mathbf{x}, \mathbf{y}$ is zero if and only if the random vectors are independent. This may also be interpreted as the KL divergence $D_{\mathrm{KL}}\left(\mathbf{p}_{\mathbf{x},\mathbf{y}} || \mathbf{p}_{\mathbf{x}}\mathbf{p}_{\mathbf{y}}\right)$ between the joint density and the product of the marginal densities; the latter quantity generalises readily to distributions of more than two random variables (there exist other methods for independence measurement: see for instance Ingster, 1989).

There has recently been considerable interest in using criteria based on functions in reproducing kernel Hilbert spaces to measure dependence, notably in the context of independent component analysis. [1] This was first accomplished by Bach and Jordan (2002a), who introduced kernel dependence functionals that signifi-

1. The problem of instantaneous independent component analysis involves the recovery of linearly mixed, i.i.d. sources, in the absence of information about the source distributions beyond their mutual independence (Hyvärinen et al., 2001).

Table 1: Table of kernel dependence functionals. Columns show whether the functional is covariance or correlation based, and rows indicate whether the dependence measure is the maximum singular value of the covariance/correlation operator, or a bound on the mutual information.

	Covariance	Correlation
Max. singular value	COCO (Gretton et al., 2005b)	KCC (Bach and Jordan, 2002a)
MI bound	KMI	KGV (Bach and Jordan, 2002a)

cantly outperformed alternative approaches, including for source distributions that are difficult for standard ICA methods to deal with. In the present study, we build on this work with the introduction of two novel kernel-based independence measures. The first, which we call the constrained covariance (COCO), is simply the spectral norm of the covariance operator between reproducing kernel Hilbert spaces. We prove COCO to be zero if and only if the random variables being tested are independent, as long as the RKHSs used to compute it are universal. The second functional, called the kernel mutual information (KMI), is a more sophisticated measure of dependence, being a function of the entire spectrum of the covariance operator. We show that the KMI is an upper bound near independence on a Parzen window estimate of the mutual information, which becomes tight (*i.e.*, zero) when the random variables are independent, again assuming universal RKHSs. Note that Gretton et al. (2003a,b) attempted to show a link with the Parzen window estimate, although this earlier proof is wrong - the reader may compare Section 3 in the present document with the corresponding section of the original technical report, since the differences are fairly obvious. [2]

The constrained covariance has substantial precedent in the dependence testing literature. Indeed, Rényi (1959) suggested using the functional covariance or correlation to measure the dependence of random variables (implementation details depend on the nature of the function spaces chosen: the use of RKHSs is a more recent innovation). Thus, rather than using the covariance, we may consider a kernelised canonical correlation (KCC) (Bach and Jordan, 2002a; Leurgans et al., 1993), which is a regularised estimate of the spectral norm of the *correlation* operator between reproducing kernel Hilbert spaces. It follows from the properties of COCO that the KCC is zero at independence for universal kernels, since the correlation differs from the covariance only in its normalisation: at independence, where both the KCC and COCO are zero, this normalisation is immaterial. The introduction of a regulariser requires a new parameter that must be tuned, however, which was not needed for COCO or the KMI.

Another kernel method for dependence measurement, the kernel generalised variance (KGV) (Bach and Jordan, 2002a), extends the KCC by incorporating the entire spectrum of its associated correlation operator: in this respect, the KGV and KMI are analogous (see Table 1). Indeed, we prove here that under certain reasonable and easily enforced conditions, the KGV is an upper bound on the KMI (and hence on the mutual information near independence), which also becomes tight at independence. A relation between the KGV and the mutual information is also proposed by Bach and Jordan (2002a), who rely on a limiting argument in which the RKHS kernel size approaches zero (no Parzen window estimate is invoked): our discussion of this proof is given in Appendix B.2.

We should warn the reader that results presented in this study have a conceptual emphasis: we attempt to build on the work of Bach and Jordan (2002a) by on one hand exploring the mechanism by which kernel covariance operator-based functionals measure independence (including a characterisation of all kernels that induce independence measures), and on the other hand demonstrating the link between kernel dependence

2. Briefly, we now use Lemma 27 as a basis for our proof, which applies to every singular value of a matrix product; our earlier proof relied on Theorem 4.2.2 of Gretton et al. (2003a), which implies a result only for the largest singular value, and is therefore insufficient. On the other hand, we believe that the proof given by Gretton (2003) in Chapter 9 is correct, but the approach is a bit clumsy, and much longer than it needs to be.

Table 2: Table of acronyms

Acronym	Description
COCO	Constrained covariance
ICA	Independent component analysis
KCC	Kernel canonical correlation
KGV	Kernel generalised variance
KMI	Kernel mutual information
RKHS	Reproducing kernel Hilbert space

functionals and the mutual information. That said, we observe differences in practice when the various kernel methods are applied in ICA: the KMI generally outperforms the KGV for many sources/large sample sizes, whereas the KGV gives best performance for small sample sizes. The choice of regulariser for the KGV (and KCC) is also crucial, since a badly chosen regularisation is severely detrimental to performance when outlier noise is present. The KMI and COCO are robust to outliers, and yield experimental performance equivalent to the KGV and KCC with optimal regulariser choice, but without any tuning required.

The COCO and KCC dependence functionals for the 2-variable case are described in Section 2, and it is shown that these measure independence when the associated kernels are universal. The main results in this section are Definition 2, which presents both the population COCO and its empirical counterpart, and Theorem 6, which shows that COCO is an independence measure. Section 3 contains derivations of the kernel-based upper bounds on the mutual information, and proofs that these latter quantities likewise measure independence. In particular, the kernel mutual information is introduced in Definition 14, its use as an independence measure is justified by Theorem 15, and its relation to the mutual information is provided in Theorem 16. A generalisation to more than two variables, which permits the measurement of pairwise independence, is also presented. Section 4 addresses the application of kernel dependence measures to independent component analysis, including a method for reducing computational cost and a gradient descent technique (these being adapted straightforwardly from Bach and Jordan, 2002a). Finally, Section 5 describes our experiments: these demonstrate that the performance of the KMI and COCO, when used in ICA, is competitive with the KGV and KCC, respectively. The kernel methods also compare favourably with both standard and recent specialised ICA algorithms (RADICAL, CFICA, Fast ICA, Jade, and Infomax), and outperform these methods when demixing music sources (where the sample size is large). Most interestingly, when the KGV is made to approach the KMI by an appropriate choice of regularisation, its resistance to outlier noise is improved — moreover, kernel methods perform substantially better than the other algorithms tested when outliers are present.[3] We list our most commonly used acronyms in Table 2.

2. Constrained covariance, kernel canonical correlation

In this section, we focus on the formulation of measures of independence for two random variables. This reasoning uses well established principles, going back to Rényi (1959), who gave a list of desirable properties for a measure of statistical dependence $\mathcal{Q}(\mathbf{P}_{\mathsf{x},\mathsf{y}})$ between random variables x,y with distribution $\mathbf{P}_{\mathsf{x},\mathsf{y}}$. These include

1. $\mathcal{Q}(\mathbf{P}_{\mathsf{x},\mathsf{y}})$ is well defined,
2. $0 \leq \mathcal{Q}(\mathbf{P}_{\mathsf{x},\mathsf{y}}) \leq 1$,
3. $\mathcal{Q}(\mathbf{P}_{\mathsf{x},\mathsf{y}}) = 0$ if and only if x,y independent,
4. $\mathcal{Q}(\mathbf{P}_{\mathsf{x},\mathsf{y}}) = 1$ if and only if $\mathsf{y} = f(\mathsf{x})$ or $\mathsf{x} = g(\mathsf{y})$, where f and g are Borel measurable functions.

3. The performance reported here improves on that obtained by Bach and Jordan (2002a); Miller and Fisher III (2003) due to better tuning of the KGV and KCC regularisation.

Rényi (1959) shows that one measure satisfying these constraints is

$$\mathcal{Q}(\mathbf{P}_{\mathsf{x},\mathsf{y}}) \quad = \quad \sup_{f,g} \operatorname{corr}(f(\mathsf{x}), g(\mathsf{y})),$$

where $f(\mathsf{x}), g(\mathsf{y})$ must have finite positive variance, and f, g are Borel measurable. This is similar to the kernel canonical correlation (KCC) introduced by Bach and Jordan (2002a), although we shall see that the latter is more restrictive in its choice of f, g. We propose a different measure, the *constrained covariance* (COCO), which omits the fourth property and the upper bound in the second property; in the context of independence measurement, however, the first and third properties are adequate. [4]

We begin in Section 2.1 by defining RKHSs and covariance operators between them. In Section 2.2, we introduce the constrained covariance, and we demonstrate in Section 2.3 that this quantity is a measure of independence when computed in universal RKHSs (it follows that the KCC also requires a universal RKHS, as do all independence criteria that are based on the covariance in RKHSs). Finally, we describe the canonical correlation in Section 2.4, and its RKHS-based variant.

2.1 Covariance in function spaces

In this section, we provide the functional analytic background necessary in describing covariance operators between RKHSs. Our presentation follows and extends the work of Zwald et al. (2004); Hein and Bousquet (2004), who deal with covariance operators from a space to itself rather than from one space to another, and Fukumizu et al. (2004), who use covariance operators as a means of defining conditional covariance operators. Functional covariance operators were investigated earlier by Baker (1973), who characterises these operators for general Hilbert spaces.

Consider a Hilbert space F of functions from X to $\mathbb{R}$, where X is a separable metric space. The Hilbert space F is an RKHS if at each $x \in X$, the point evaluation operator $\delta_x : F \to \mathbb{R}$, which maps $f \in F$ to $f(x) \in \mathbb{R}$, is a bounded linear functional. To each point $x \in X$, there corresponds an element $\mathbf{x} := \phi(x) \in F$ (we call ϕ the *feature map*) such that $\langle \phi(x), \phi(x') \rangle_F = k(x, x')$, where $k : X \times X \to \mathbb{R}$ is a unique positive definite kernel. We also define a second RKHS G with respect to the separable metric space Y, with feature map ψ and kernel $\langle \psi(y), \psi(y') \rangle_G = l(y, y')$.

Let $\mathbf{P}_{\mathsf{x},\mathsf{y}}(x, y)$ be a joint measure[5] on $(X \times Y, \Gamma \times \Lambda)$ (here Γ and Λ are the Borel σ-algebras on X and Y, respectively, as required in Theorem 4 below), with associated marginal measures $\mathbf{P}_{\mathsf{x}}$ and $\mathbf{P}_{\mathsf{y}}$ and random variables x and y. Then following Baker (1973); Fukumizu et al. (2004), the covariance operator $C_{xy} : G \to F$ is defined[6] such that for all $f \in F$ and $g \in G$,

$$\langle f, C_{xy} g \rangle_F \quad = \quad \mathbf{E}_{\mathsf{x},\mathsf{y}} \left([f(\mathsf{x}) - \mathbf{E}_{\mathsf{x}}(f(\mathsf{x}))] [g(\mathsf{y}) - \mathbf{E}_{\mathsf{y}}(g(\mathsf{y}))] \right).$$

In practice, we do not deal with the measure $\mathbf{P}_{\mathsf{x},\mathsf{y}}$ itself, but instead observe samples drawn independently according to it. We write an i.i.d. sample of size m from $\mathbf{P}_{\mathbf{x},\mathbf{y}}$ as $\boldsymbol{z} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, and likewise $\boldsymbol{x} := \{x_1, \ldots, x_m\}$ and $\boldsymbol{y} := \{y_1, \ldots y_m\}$. Finally, we define the Gram matrices $\mathbf{K}$ and $\mathbf{L}$ of inner products in F and G, respectively, between the mapped observations above: here $\mathbf{K}$ has $(i.j)$th entry $k(x_i, x_j)$ and $\mathbf{L}$ has (i, j)th entry $l(y_i, y_j)$. The Gram matrices for the variables centred in their respective feature spaces are shown by Schölkopf et al. (1998) to be

$$\widetilde{\mathbf{K}} := \mathbf{HKH}, \quad \widetilde{\mathbf{L}} := \mathbf{HLH},$$

where

$$\mathbf{H} = \mathbf{I} - \frac{1}{m} \mathbf{1}_m \mathbf{1}_m^\top, \tag{1}$$

and $\mathbf{1}_m$ is an $m \times 1$ vector of ones.

4. The fourth property is required for $\mathcal{Q}$ to identify deterministic dependence, which an *in*dependence measure should not be concerned with.
5. We do not require this to have a density with respect to a reference measure $dx \times dy$ in this section. Note that we will need a density in Section 3, however.
6. Our operator (and that of Fukumizu et al., 2004) differs from Baker's in that Baker defines all measures directly on the function spaces.

2.2 The constrained covariance

In this section, we define the constrained covariance (COCO), and describe the properties of the kernelised version. The covariance between x and y is defined as follows.

Definition 1 (Covariance) *The covariance of two random variables* x,y *is given as*

$$\mathrm{cov}(\mathsf{x},\mathsf{y}) := \mathbf{E}_{\mathsf{x},\mathsf{y}}[\mathsf{x}\mathsf{y}] - \mathbf{E}_{\mathsf{x}}[\mathsf{x}]\mathbf{E}_{\mathsf{y}}[\mathsf{y}].$$

We next define the constrained covariance.

Definition 2 (Constrained Covariance (COCO)) *Given function classes* F, G *and a probability measure* $\mathbf{P}_{\mathsf{x},\mathsf{y}}$*, we define the* constrained covariance *as*

$$\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}}; F, G) := \sup_{f\in F, g\in G} [\mathrm{cov}(f(\mathsf{x}), g(\mathsf{y}))]. \tag{2}$$

If F *and* G *are unit balls in their respective vector spaces, then this is just the norm of the covariance operator: see Mourier (1953). Given* m *independent observations* $\boldsymbol{z} := ((x_1,y_1),\ldots,(x_m,y_m)) \subset (X\times Y)^m$*, the empirical estimate of COCO is defined as*

$$\mathrm{COCO}(\boldsymbol{z}; F, G) := \sup_{f\in F, g\in G} \left[\frac{1}{m}\sum_{i=1}^m f(x_i)g(y_i) - \frac{1}{m^2}\sum_{i=1}^m f(x_i)\sum_{j=1}^m g(y_j)\right].$$

When F and G are RKHSs, with F and G their respective unit balls, then $\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}}; F, G)$ is guaranteed to exist as long as the kernels k and l are bounded, since the covariance operator is then Hilbert-Schmidt (as shown by Gretton et al., 2005a). The empirical estimate $\mathrm{COCO}(\boldsymbol{z}; F, G)$ is also simplified when F and G are unit balls in RKHSs, since the representer theorem (Schölkopf and Smola, 2002) holds: this states that a solution of an optimisation problem, dependent only on the function evaluations on a set of observations and on RKHS norms, lies in the span of the kernel functions evaluated on the observations. This leads to the following lemma:

Lemma 3 (Value of $\mathrm{COCO}(\boldsymbol{z}; F, G)$**)** *Denote by* F, G *RKHSs on the domains* X *and* Y *respectively, and let* F, G *be the unit balls in the corresponding RKHSs. Then*

$$\mathrm{COCO}(z; F, G) = \frac{1}{m}\sqrt{\|\widetilde{\mathbf{K}}\widetilde{\mathbf{L}}\|_2}, \tag{3}$$

where the matrix norm $\|\cdot\|_2$ *denotes the largest singular value. An equivalent* unnormalised *form (which we will refer back to in Section 3) is* $\mathrm{COCO}(\boldsymbol{z}; F, G) = \max_i \gamma_i$*, where* γ_i *are the solutions to the generalised eigenvalue problem*

$$\begin{bmatrix} \mathbf{0} & \widetilde{\mathbf{K}}\widetilde{\mathbf{L}} \\ \widetilde{\mathbf{L}}\widetilde{\mathbf{K}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i \end{bmatrix} = \gamma_i \begin{bmatrix} \widetilde{\mathbf{K}} & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{L}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i \end{bmatrix}. \tag{4}$$

Proof By the representer theorem, the solution of the maximisation problem arising from $\mathrm{COCO}(\boldsymbol{z}; F, G)$ is given by $f(x) = \sum_{i=1}^m \alpha_i k(x_i, x)$ and $g(y) = \sum_{j=1}^m \beta_j l(y_j, y)$. Hence

$$\begin{aligned}
\mathrm{COCO}(\boldsymbol{z}; F, G) &= \sup_{\boldsymbol{\alpha}^\top\mathbf{K}\boldsymbol{\alpha}\le 1, \boldsymbol{\beta}^\top\mathbf{L}\boldsymbol{\beta}\le 1} \frac{1}{m}\boldsymbol{\alpha}^\top\mathbf{K}\mathbf{L}\boldsymbol{\beta} - \frac{1}{m^2}\boldsymbol{\alpha}^\top\mathbf{K}\mathbf{1}_m\mathbf{1}_m^\top\mathbf{L}\boldsymbol{\beta} \\
&= \sup_{\|\boldsymbol{\alpha}\|,\|\boldsymbol{\beta}\|\le 1} \frac{1}{m}\boldsymbol{\alpha}^\top\mathbf{K}^{1/2}\mathbf{H}\mathbf{L}^{1/2}\boldsymbol{\beta} \\
&= \frac{1}{m}\|\mathbf{K}^{1/2}\mathbf{H}\mathbf{L}^{1/2}\|_2.
\end{aligned}$$

Squaring the argument in the norm, rearranging, and using the fact that $\mathbf{H} = \mathbf{HH}$ proves the lemma. ■

The constrained covariance turns out to be similar in certain respects to a number of kernel algorithms, for an appropriate choice of F, G. By contrast with independence measurement, however, these methods seek to *maximise* the constrained covariance through the correct choice of feature space elements. First, and most obvious, is kernel partial least squares (kPLS) (Rosipal and Trejo, 2001), which at each stage maximises the constrained covariance directly (see Bakır et al., 2004). COCO is also optimised when obtaining the first principal component in kernel principal component analysis (kPCA), as described by Schölkopf et al. (1998), and is the criterion optimised in the spectral clustering/kernel target alignment framework of Cristianini et al. (2002). Details may be found in Appendix A.1.

Finally, we remark that alternative norms of the covariance operator should also be suited to measuring independence. Indeed, the Hilbert-Schmidt (HS) norm is proposed in this context by Gretton et al. (2005a): like the KMI, it exploits the entire spectrum of the empirical covariance operator, and gives experimental performance superior to COCO in ICA. The HS norm has the additional advantage of a well-defined population counterpart, and guarantees of $O(1/\sqrt{m})$ convergence of the empirical to the population quantity. The connection between the HS norm and the mutual information remains unknown, however.

2.3 Independence measurement with the constrained covariance

We now describe how COCO is used as a measure of independence. For our purposes, the notion of independence of random variables is best characterised by Jacod and Protter (2000, Theorem 10.1(e)):

Theorem 4 (Independence) *Let* x *and* y *be random variables on* $(X \times Y, \Gamma \times \Lambda)$ *with joint measure* $\mathbf{P}_{\mathsf{x},\mathsf{y}}(x,y)$*, where* Γ *and* Λ *are Borel* σ*-algebras on* X *and* Y*, respectively. Then the random variables* x *and* y *are independent if and only if* $\mathrm{cov}(f(\mathsf{x}), g(\mathsf{y})) = 0$ *for any pair* (f,g) *of bounded, continuous functions.*

It follows from Theorem 4 that if F, G are the sets of bounded continuous functions, then $\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}}; F, G) = 0$ if and only if x and y are independent. In other words, $\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}}; F, G)$ and $\mathrm{COCO}(z; F, G)$ are criteria which can be tested *directly* without the need for an intermediate density estimator (in general, the distributions may not even have densities). It is also clear, however, that unless F, G are restricted in further ways, $\mathrm{COCO}(\boldsymbol{z}; F, G)$ will always be large, due to the rich choice of functions available. A *non-trivial dependence functional* is thus obtained using function classes that do not give an everywhere-zero empirical average, yet which still guarantee that COCO is zero if and only if its arguments are independent. A tradeoff between the restrictiveness of the function classes and the convergence of $\mathrm{COCO}(\boldsymbol{z}; F, G)$ to $\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}}; F, G)$ can be accomplished using standard tools from uniform convergence theory (see Gretton et al., 2005b). It turns out that unit-radius balls in *universal* reproducing kernel Hilbert spaces constitute function classes that yield non-trivial dependence estimates. Universality is defined by Steinwart (2001) as follows:

Definition 5 (Universal kernel) *A* continuous kernel $k(\cdot,\cdot)$ *on a compact metric space* (X,d) *is called universal if and only if the RKHS* F *induced by the kernel is dense in* $C(X)$*, the space of continuous functions on* X*, with respect to the infinity norm* $\|f - g\|_\infty$*.*

Steinwart (2001) shows the following two kernels are universal on compact subsets of $\mathbb{R}^d$:

$$k(x,x') = \exp\left(-\lambda \|x - x'\|^2\right) \text{ and}$$
$$k(x,x') = \exp\left(-\lambda \|x - x'\|\right) \text{ for } \lambda > 0.$$

We now state our main result for this section.

Theorem 6 ($\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}}; F, G)$ **is only zero at independence for universal kernels)** *Denote by* F, G *RKHSs with universal kernels on the compact metric spaces* X *and* Y*, respectively, and let* F, G *be the unit balls in* F *and* G*. Then* $\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}}; F, G) = 0$ *if and only if* x, y *are independent.*

Proof It is clear that $\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}};F,G)$ is zero if x and y are independent. We prove the converse by showing that[7] $\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}};B(X),B(Y))=c$ for some $c>0$ implies $\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}};F,G)=d$ for $d>0$: this is equivalent to $\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}};F,G)=0$ implying $\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}};B(X),B(Y))=0$ (where this last result implies independence by Theorem 4). There exist two sequences of functions $f_n \in C(X)$ and $g_n \in C(Y)$, satisfying $\|f_n\|_\infty \leq 1, \|g_n\|_\infty \leq 1$, for which

$$\lim_{n\to\infty} \mathrm{cov}(f_n(\mathsf{x}),g_n(\mathsf{y}))=c.$$

More to the point, there exists an n^* for which $\mathrm{cov}(f_{n^*}(\mathsf{x}),g_{n^*}(\mathsf{y})) \geq c/2$. We know that F and G are respectively dense in $C(X)$ and $C(Y)$ with respect to the L_∞ norm: this means that for all $\frac{c}{24}>\varepsilon>0$, we can find some $f^* \in F$ (and an analogous $g^* \in G$) satisfying $\|f^*-f_{n^*}\|_\infty < \varepsilon$. Thus, we obtain

$$\begin{aligned}
\mathrm{cov}(f^*(\mathsf{x}),g^*(\mathsf{y})) &= \mathrm{cov}\left(f^*(\mathsf{x})-f_{n^*}(\mathsf{x})+f_{n^*}(\mathsf{x}),g^*(\mathsf{x})-g_{n^*}(\mathsf{x})+g_{n^*}(\mathsf{x})\right)\\
&= \mathbf{E}_{\mathsf{x},\mathsf{y}}\left[\left(f^*(\mathsf{x})-f_{n^*}(\mathsf{x})+f_{n^*}(\mathsf{x})\right)\left(g^*(\mathsf{y})-g_{n^*}(\mathsf{y})+g_{n^*}(\mathsf{y})\right)\right]\\
&\qquad -\mathbf{E}_{\mathsf{x}}\left(f^*(\mathsf{x})-f_{n^*}(\mathsf{x})+f_{n^*}(\mathsf{x})\right)\mathbf{E}_{\mathsf{y}}\left(g^*(\mathsf{y})-g_{n^*}(\mathsf{y})+g_{n^*}(\mathsf{y})\right)\\
&\geq \mathrm{cov}(f_{n^*}(\mathsf{x}),g_{n^*}(\mathsf{y}))-2\varepsilon\left|\mathbf{E}_{\mathsf{x}}\left(f_{n^*}(\mathsf{x})\right)\right|-2\varepsilon\left|\mathbf{E}_{\mathsf{y}}\left(g_{n^*}(\mathsf{y})\right)\right|-2\varepsilon^2\\
&\geq \frac{c}{2}-6\frac{c}{24}=\frac{c}{4}>0.
\end{aligned}$$

Finally, bearing in mind that $\|f^*(\mathsf{x})\|_F < \infty$ and $\|g^*(\mathsf{x})\|_G < \infty$, we have

$$\mathrm{cov}\left(\frac{f^*(\mathsf{x})}{\|f^*(\mathsf{x})\|_F},\frac{g^*(\mathsf{y})}{\|g^*(\mathsf{x})\|_G}\right) \geq \frac{c}{4\|f^*(\mathsf{x})\|_F\|g^*(\mathsf{x})\|_G}>0,$$

and hence $\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}};F,G)>0$. ■

The constrained covariance is further explored by Gretton et al. (2005b, 2004). We prove two main results in these studies, which are not covered in the present work:

- Theorems 10 and 11 of Gretton et al. (2005b) give upper bounds on the probability of large deviations of the empirical COCO from the population COCO: Theorem 10 covers negative deviations of the empirical COCO from the population COCO, and Theorem 11 describes positive deviations. For a fixed probability of deviation, the amount by which the empirical COCO differs from the population COCO decreases at rate $1/\sqrt{m}$ (for shifts in either direction). These bounds are necessary if we are to formulate *statistical tests* of independence based on the *measure* of independence that COCO provides. In particular, Gretton et al. (2005b, Section 5) give one such test .
- Theorem 8 of Gretton et al. (2005b) describes the behaviour of the population COCO when the random variables are not independent, for a simple family of probability densities represented as orthogonal series expansions. This is used to illustrate two concepts: first, that dependence can sometimes be hard to detect without a large number of samples (since the deviation of the population COCO from zero can be very small, even for dependent random variables); and second, that one type of hard-to-detect dependence is encoded in high frequencies of the probability density function.

We also apply COCO in these studies to detecting dependence in fMRI scans of the Macaque visual cortex. We refer the reader to these references for further detail on COCO.

2.4 The canonical correlation

The kernelised canonical correlation (KCC) — i.e., the norm of the *correlation operator* between RKHSs — was proposed as a measure of independence by Bach and Jordan (2002a). Consistency of the KCC was

7. Here $B(X)$ denotes the subset of $C(X)$ of continuous functions bounded by 1 in $L_\infty(X)$, and $B(Y)$ is defined in an analogous manner.

shown by Leurgans et al. (1993) for the operator norm, and by Fukumizu et al. (2005) for the functions in F and G that define it (in accordance with Definition 7 below). Further discussion and applications of the kernel canonical correlation include Akaho (2001); Bach and Jordan (2002a); Hardoon et al. (2004); Kuss (2001); Lai and Fyfe (2000); Melzer et al. (2001); Shawe-Taylor and Cristianini (2004); van Gestel et al. (2001). In particular, a much more extensive discussion of the properties of canonical correlation analysis and its kernelisation may be found in these studies, and this section simply summarises the properties and derivations relevant to our requirements for independence measurement.

The idea underlying the KCC is to find the functions $f \in F$ and $g \in G$ with largest *correlation* (as opposed to covariance, which we covered in the previous section). This leads to the following definition.

Definition 7 (Kernel canonical correlation (KCC)) *The kernel canonical correlation is defined as*

$$\begin{aligned} \mathrm{KCC}\left(\mathbf{P}_{\mathsf{x},\mathsf{y}}; F, G\right) &= \sup_{f \in F, g \in G} \mathrm{corr}\left(f(\mathsf{x}), g(\mathsf{y})\right) \\ &= \sup_{f \in F, g \in G} \frac{\mathbf{E}\left(f(\mathsf{x})g(\mathsf{y})\right) - \mathbf{E}_{\mathsf{x}}\left(f(\mathsf{x})\right)\mathbf{E}_{\mathsf{y}}\left(g(\mathsf{y})\right)}{\sqrt{\mathbf{E}_{\mathsf{x}}\left(f^2(\mathsf{x})\right) - \mathbf{E}_{\mathsf{x}}^2\left(f(\mathsf{x})\right)}\sqrt{\mathbf{E}_{\mathsf{y}}\left(g^2(\mathsf{y})\right) - \mathbf{E}_{\mathsf{y}}^2\left(g(\mathsf{y})\right)}}. \end{aligned}$$

As in the case of the constrained covariance, we may specify an empirical estimate similar to that in Lemma 3:

Lemma 8 (Empirical KCC) *The empirical kernel canonical correlation is given by* $\mathrm{KCC}\left(\boldsymbol{z}; F, G\right) := \max_i\left(\rho_i\right)$*, where* ρ_i *are the solutions to the generalised eigenvalue problem*

$$\begin{bmatrix} \mathbf{0} & \widetilde{\mathbf{K}}\widetilde{\mathbf{L}} \\ \widetilde{\mathbf{L}}\widetilde{\mathbf{K}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} = \rho_i \begin{bmatrix} \widetilde{\mathbf{K}}^2 & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{L}}^2 \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix}. \tag{5}$$

Bach and Jordan (2002a) point out that the first canonical correlation is very similar to the function maximised by the *alternating conditional expectation* algorithm of Breiman and Friedman (1985), although in the latter case $f(x)$ may be replaced with a linear combination of several functions of x.

We note that the numerator of the functional in Definition 7 is just the functional covariance, which suggests that the kernel canonical correlation might also be a useful measure of independence: this was proposed by Bach and Jordan (2002a) (the functional correlation was also analysed as an independence measure by Dauxois and Nkiet (1998), although this approach did not make use of RKHSs). A problem with using the kernel canonical correlation to measure independence is discussed in various forms by Bach and Jordan (2002a); Fukumizu et al. (2005); Greenacre (1984); Kuss (2001); Leurgans et al. (1993); we now describe one formulation of problem, and the two main ways in which it has been solved.

Lemma 9 (Without regularisation, the empirical KCC is independent of the data) *Suppose that the Gram matrices* $\mathbf{K}$ *and* $\mathbf{L}$ *have full rank. The* $2(m-1)$ *non-zero solutions to (5) are then* $\rho_i = \pm 1$*, regardless of* $\boldsymbol{z}$*.*

The proof is in Appendix B.1. This argument is used by Bach and Jordan (2002a); Fukumizu et al. (2005); Leurgans et al. (1993) to justify a regularised canonical correlation,

$$\mathrm{KCC}\left(\mathbf{P}_{\mathsf{x},\mathsf{y}}; F, G, \kappa\right) := \sup_{f \in F, g \in G} \frac{\mathrm{cov}\left(f(\mathsf{x}), g(\mathsf{y})\right)}{\left(\mathrm{var}\left(f\left(\mathsf{x}\right)\right) + \kappa \left\|f\right\|_F^2\right)^{1/2} \left(\mathrm{var}\left(g\left(\mathsf{y}\right)\right) + \kappa \left\|g\right\|_G^2\right)^{1/2}}, \tag{6}$$

although this requires an additional parameter κ, which complicates the model selection problem. As the number of observations increases, κ must approach zero to ensure consistency of the estimated KCC, and of the associated functions f and g that achieve the supremum. The rate of decrease of κ for consistency of KCC is derived by Leurgans et al. (1993) (for RKHSs based on spline kernels), and the rate required for consistency in the L_2 norm of f and g is obtained by Fukumizu et al. (2005) (for all RKHSs).

An alternative solution to the problem described in Lemma 9 is given by Kuss (2001), in which the projection directions used to compute the canonical correlations are expressed in terms of a more restricted set of basis functions, rather than the respective subspaces of F and G spanned by the entire set of mapped observations. These basis functions can be chosen using kernel PCA, for instance.

Finally, we show that the regularised kernel canonical correlation is a measure of independence, as long as the functions attaining the supremum have bounded variance.

Theorem 10 ($\mathrm{KCC}(\mathbf{P}_{\mathsf{x},\mathsf{y}};F,G,\kappa)=0$ **only at independence for universal kernels)** *Denote by* F,G *RKHSs with universal kernels on the compact metric spaces* X *and* Y*, respectively, and assume that* $\mathrm{var}(f(\mathsf{x}))<\infty$ *and* $\mathrm{var}(g(\mathsf{y}))<\infty$*. Then* $\mathrm{KCC}(\mathbf{P}_{\mathsf{x},\mathsf{y}};F,G,\kappa)=0$ *if and only if* x,y *are independent.*

Proof The proof is almost identical to the proof of Theorem 6. First, it is clear that x and y being independent implies $\mathrm{KCC}(\mathbf{P}_{\mathsf{x},\mathsf{y}};F,G,\kappa)=0$. Next, assume $\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}};B(X),B(Y))=c$ for $c>0$. We can then define $f^*\in F$ and $g^*\in G$ as before, such that

$$\mathrm{cov}(f^*(\mathsf{x}),g^*(\mathsf{y}))\geq\frac{c}{4}.$$

Finally, assuming $\mathrm{var}(f(\mathsf{x}))$ and $\mathrm{var}(g(\mathsf{y}))$ to be bounded, we get

$$\begin{aligned}&\mathrm{cov}\left(\frac{f^*(\mathsf{x})}{\left(\mathrm{var}(f^*(\mathsf{x}))+\kappa\|f^*\|_F^2\right)^{1/2}},\frac{g^*(\mathsf{y})}{\left(\mathrm{var}(g^*(\mathsf{y}))+\kappa\|g^*\|_G^2\right)^{1/2}}\right)\\ \geq\;&\frac{c}{4\left(\mathrm{var}(f^*(\mathsf{x}))+\kappa\|f^*\|_F^2\right)^{1/2}\left(\mathrm{var}(g^*(\mathsf{y}))+\kappa\|g^*\|_G^2\right)^{1/2}}\\ >\;&0.\end{aligned}$$

The requirement of bounded variance is not onerous: indeed, as in the case of the covariance operator, we are guaranteed that $\mathrm{var}(f(\mathsf{x}))$ and $\mathrm{var}(g(\mathsf{y}))$ are bounded when k and l are bounded. ■

3. Kernel approximations to the mutual information

In this section, we investigate approximations to the mutual information which can be used for measuring independence, in that they are still zero only when the random variables being tested are independent. We begin in Section 3.1 by introducing the mutual information between two multivariate Gaussian random variables, for which a closed form solution exists. We then describe a discrete approximation to the mutual information between *two continuous, univariate* random variables with an *arbitrary* joint density function, which is defined via a partitioning of the continuous space into a uniform grid of bins; it is well established that this approximation approaches the continuous mutual information as the grid becomes infinitely fine (Cover and Thomas, 1991). Next, we show that the discrete mutual information may be approximated by the Gaussian mutual information (GMI), by doing a Taylor expansion of both quantities to second order around independence.

We next address how to go about estimating this Gaussian approximation of the discrete mutual information, given observations drawn according to some probability density. In Section 3.1.4, we derive a Parzen window estimate of the GMI. Next, in Section 3.1.5, we give an upper bound on the empirical GMI, which we call the *kernel mutual information* (KMI) (Definition 14 presents the KMI, and Theorem 16 contains the bound: these are the main results of this section). We show in Theorem 15 that the KMI is zero if and only if the empirical COCO is zero, which justifies using the KMI as a measure of independence. An important property of this bound is that it does *not* require numerical integration, or indeed any space partitioning or grid-based approximations (see e.g. Paninski (2003) and references therein). Rather, we are able to obtain a

closed form expression when the grid[8] becomes *infinitely fine*. Finally, we demonstrate in Section 3.1.6 that the regularised kernel generalised variance (KGV) proposed by Bach and Jordan (2002a) is an upper bound on the KMI, and hence on the Gaussian mutual information, under certain circumstances. A comparison with the link originally proposed between the KGV and the mutual information is given in Appendix B.2.

We should emphasise at this point an important distinction between the KMI and KGV on one hand, and COCO and the KCC on the other. We recall that the empirical COCO in Lemma 3 is a finite sample estimate of the population quantity in Definition 2, and the empirical KCC in Lemma 8 has a population equivalent in Definition 7 (convergence of the empirical estimates to the population quantities is guaranteed in both cases, as described in the discussion of Section 2). The KMI and KGV, on the other hand, are bounds on particular sample-based quantities, and are *not* defined here with respect to corresponding population expressions. That said, the KGV appears to be a regularised empirical estimate of the mutual information for Gaussian processes of Baker (1970), although to our knowledge the convergence of the KGV to this population quantity is not yet established.

In Section 3.2, we derive generalisations of the COCO and KMI to more than two univariate random variables. We prove the high dimensional COCO and KMI are zero if and only if the associated pairwise empirical constrained covariances are zero, which makes them suited for application in ICA (see Theorem 24).

3.1 The KMI, the KGV, and the mutual information

3.1.1 Mutual information between two multivariate Gaussian random variables

We begin by introducing the Gaussian mutual information and its relation with the canonical correlation. Thus, the present section should be taken as background material which we will refer back to in the discussion that follows. Cover and Thomas (1991) provide a more detailed and general discussion of these principles. If $\mathbf{x}_G, \mathbf{y}_G$ are Gaussian random vectors[9] in $\mathbb{R}^{l_x}, \mathbb{R}^{l_y}$ respectively, with joint covariance matrix $\mathbf{C} := \begin{bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{xy}^\top & \mathbf{C}_{yy} \end{bmatrix}$, then the mutual information between them can be written

$$I(\mathbf{x}_G; \mathbf{y}_G) = -\frac{1}{2} \log \left(\frac{|\mathbf{C}|}{|\mathbf{C}_{xx}|\,|\mathbf{C}_{yy}|} \right), \tag{7}$$

where $|\cdot|$ is the determinant. We note that the Gaussian mutual information takes the distinctive form of a log ratio of determinants: we will encounter this expression repeatedly in the subsequent reasoning, under various guises. For this reason, we now present a theorem which describes several alternative expressions for this ratio.

Theorem 11 (Ratio of determinants) *Given a partitioned matrix*[10]

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{C} \end{bmatrix} \succ \mathbf{0}, \tag{8}$$

8. Introduced in the discrete approximation to the mutual information
9. The subscripts G are used to emphasise that $\mathbf{x}_G, \mathbf{y}_G$ are Gaussian; this notation is introduced here to make the reasoning clearer in subsequent sections.
10. We use $\mathbf{X} \succ \mathbf{0}$ to indicate that $\mathbf{X}$ is positive definite.

we can write

$$\begin{aligned}\frac{\left|\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^\top & \mathbf{C}\end{bmatrix}\right|}{|\mathbf{A}|\,|\mathbf{C}|} &= \left|\begin{bmatrix}\mathbf{I} & \mathbf{A}^{-1/2}\mathbf{B}\mathbf{C}^{-1/2}\\ \mathbf{C}^{-1/2}\mathbf{B}^\top\mathbf{A}^{-1/2} & \mathbf{I}\end{bmatrix}\right| \\ &= \left|\mathbf{I}-\mathbf{A}^{-1/2}\mathbf{B}\mathbf{C}^{-1}\mathbf{B}^\top\mathbf{A}^{-1/2}\right| \\ &= \prod_i(1-\rho_i^2) \\ &> 0\end{aligned}$$

where ρ_i are the singular values of $\mathbf{A}^{-1/2}\mathbf{B}\mathbf{C}^{-1/2}$ (i.e. the positive square root of the eigenvalues of $\mathbf{A}^{-1/2}\mathbf{B}\mathbf{C}^{-1}\mathbf{B}^\top\mathbf{A}^{-1/2}$). Alternatively, we can write ρ_i as the positive solutions to the generalised eigenvalue problem

$$\begin{bmatrix}\mathbf{0} & \mathbf{B}\\ \mathbf{B}^\top & \mathbf{0}\end{bmatrix}\mathbf{a}_i = \rho_i\begin{bmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \mathbf{C}\end{bmatrix}\mathbf{a}_i.$$

The proof is in Appendix A.2. Using this result, we may rewrite (7) as

$$I(\mathbf{x}_G;\mathbf{y}_G) = -\frac{1}{2}\log\left(\prod_i(1-\rho_i^2)\right), \tag{9}$$

where ρ_i are the singular values of $\mathbf{C}_{xx}^{-1/2}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1/2}$; or alternatively, the positive solutions to the generalised eigenvalue problem

$$\begin{bmatrix}\mathbf{0} & \mathbf{C}_{xy}\\ \mathbf{C}_{xy}^\top & \mathbf{0}\end{bmatrix}\mathbf{a}_i = \rho_i\begin{bmatrix}\mathbf{C}_{xx} & \mathbf{0}\\ \mathbf{0} & \mathbf{C}_{yy}\end{bmatrix}\mathbf{a}_i. \tag{10}$$

In this final configuration, it is apparent that ρ_i are the canonical correlates of the Gaussian random variables $\mathbf{x}_G$ and $\mathbf{y}_G$. We note that the definition of the Gaussian mutual information provided by (9) and (10) holds even when $\mathbf{C}$ does not have full rank (which indicates that $\begin{bmatrix}\mathbf{x}_G^\top & \mathbf{y}_G^\top\end{bmatrix}^\top$ spans a subspace of $\mathbb{R}^{l_x+l_y}$), since for $\mathbf{C}\succeq\mathbf{0}$ we require $\mathbf{C}_{xy}$ to have the same nullspace as $\mathbf{C}_{yy}$, and $\mathbf{C}_{xy}^\top$ to have the same nullspace as $\mathbf{C}_{xx}$. Alternatively, we could make a change of variables to a lower dimensional space in which the resulting covariance has full rank, and then use the ratio of determinants (7) with this new covariance.

3.1.2 Mutual information between discretised univariate random variables

In this section, and in the sections that follow, we consider only the case where X and Y are closed, bounded subsets of $\mathbb{R}$, and require $(\mathsf{x},\mathsf{y})\in X\times Y$ to have the joint density $\mathbf{p}_{\mathsf{x},\mathsf{y}}$ (this is by contrast with the discussion in Section 2, in which X and Y were defined simply as separable metric spaces, and the measure $\mathbf{P}_{\mathsf{x},\mathsf{y}}$ did not necessarily admit a density). We will also assume $X\times Y$ represents the support of $\mathbf{p}_{\mathsf{x},\mathsf{y}}$. The present section introduces a discrete approximation to the mutual information between x and y, as described by Cover and Thomas (1991). Consider a grid of size $l_x\times l_y$ over $X\times Y$. Let the indices i,j denote the point $(q_i,r_j)\in X\times Y$ on this grid, and let $\boldsymbol{q}=(q_1,\ldots,q_{l_x}),\boldsymbol{r}=(r_1,\ldots,r_{l_y})$ be the complete sequences of grid coordinates. Assume, further, that the spacing between points along the x and y axes is respectively Δ_x and Δ_y (the bins being evenly spaced). We define two multinomial random variables $\hat{\mathsf{x}},\hat{\mathsf{y}}$ with a distribution $\mathbf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j)$ over the grid (the complete $l_x\times l_y$ matrix of such probabilities is $\mathbf{P}_{xy}$); this corresponds to the probability that x,y is within a small interval surrounding the grid position q_i,r_j, so

$$\begin{aligned}\mathbf{P}_{\hat{\mathsf{x}}}(i) &= \int_{q_i}^{q_i+\Delta_x}\mathbf{p}_{\mathsf{x}}(x)dx, \qquad \mathbf{P}_{\hat{\mathsf{y}}}(j)=\int_{r_j}^{r_j+\Delta_y}\mathbf{p}_{\mathsf{y}}(y)dy,\\ \mathbf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j) &= \int_{q_i}^{q_i+\Delta_x}\int_{r_j}^{r_j+\Delta_y}\mathbf{p}_{\mathsf{x},\mathsf{y}}(x,y)dxdy.\end{aligned}$$

Thus $\mathbf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j)$ is a discretisation of $\mathbf{p}_{\mathsf{x},\mathsf{y}}$. Finally, we denote as $\mathbf{p}_x$ the vector for which $(\mathbf{p}_x)_i = \mathbf{P}_{\hat{\mathsf{x}}}(i)$, with a similar $\mathbf{p}_y$ definition. The mutual information between $\hat{\mathsf{x}}$ and $\hat{\mathsf{y}}$ is defined as

$$I(\hat{\mathsf{x}};\hat{\mathsf{y}}) = \sum_{i=1}^{l_x}\sum_{j=1}^{l_y} \mathbf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j)\log\left(\frac{\mathbf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j)}{\mathbf{P}_{\hat{\mathsf{x}}}(i)\,\mathbf{P}_{\hat{\mathsf{y}}}(j)}\right). \tag{11}$$

It is well known that $I(\mathsf{x},\mathsf{y})$ is the limit of $I(\hat{\mathsf{x}};\hat{\mathsf{y}})$ as the discretisation becomes infinitely fine (Cover and Thomas, 1991, Section 9.5).

3.1.3 Multivariate Gaussian approximation to the discretised mutual information

In this section, we draw together results from the two previous sections, showing it is possible to approximate the *discrete* mutual information in Section 3.1.2 with a *Gaussian* mutual information between vectors of sufficiently high dimension, as long as we are close to independence. The results in this section are due to Bach and Jordan (2002a), although the proof of (18) below is novel. We begin by defining an equivalent multidimensional representation $\check{\mathbf{x}},\check{\mathbf{y}}$ of $\hat{\mathsf{x}},\hat{\mathsf{y}}$ in the previous section, where $\check{\mathbf{x}} \in \mathbb{R}^{l_x}$ and $\check{\mathbf{y}} \in \mathbb{R}^{l_y}$, such that $\hat{\mathsf{x}} = i$ is equivalent to $(\check{\mathbf{x}})_i = 1$ and $(\check{\mathbf{x}})_{j:\,j\neq i} = 0$. To be precise, we define the functions[11]

$$\mathfrak{K}_i(x) = \begin{cases} 1 & x \in [q_i, q_i+\Delta_x) \\ 0 & \text{otherwise} \end{cases}, \qquad \mathfrak{K}_j(y) = \begin{cases} 1 & x \in [r_j, r_j+\Delta_y) \\ 0 & \text{otherwise} \end{cases},$$

such that

$$\mathbf{E}_{\mathsf{x}}(\mathfrak{K}_i(\mathsf{x})) = \mathbf{E}_{\mathsf{x}}((\check{\mathbf{x}})_i) = \int_{-\infty}^{\infty} \mathfrak{K}_i(x)\mathbf{p}_{\mathsf{x}}(x)dx = \mathbf{P}_{\hat{\mathsf{x}}}(i)$$

and

$$\mathbf{E}_{\mathsf{x},\mathsf{y}}(\mathfrak{K}_i(\mathsf{x})\mathfrak{K}_j(\mathsf{y})) = \mathbf{E}_{\mathsf{x},\mathsf{y}}\left((\check{\mathbf{x}})_i(\check{\mathbf{y}})_j\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathfrak{K}_i(x)\mathfrak{K}_j(y)\mathbf{p}_{\mathsf{x},\mathsf{y}}(x,y)dxdy = \mathbf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j).$$

A specific instance of the second formula is when $\mathsf{y} = \mathsf{x}$, $\mathfrak{K}_i(x) = \mathfrak{K}_i(y)$, and $\mathbf{p}_{\mathsf{x},\mathsf{y}}(x,y) = \delta_x(y)\mathbf{p}_{\mathsf{x}}(x)$, where $\delta_x(y)$ is a delta function centred at x. Then

$$\begin{aligned} \mathbf{E}_{\mathsf{x}}(\mathfrak{K}_i(\mathsf{x})\mathfrak{K}_j(\mathsf{x})) = \mathbf{E}_{\mathsf{x}}\left(\left(\check{\mathbf{x}}\check{\mathbf{x}}^\top\right)_{i,j}\right) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathfrak{K}_i(x)\mathfrak{K}_j(y)\mathbf{p}_{\mathsf{x}}(x)\delta_x(y)dxdy \\ &= \begin{cases} \mathbf{P}_{\hat{\mathsf{x}}}(i) & i=j \\ 0 & \text{otherwise} \end{cases}. \end{aligned}$$

In summary,

$$\mathbf{E}_{\mathsf{x},\mathsf{y}}\left(\check{\mathbf{x}}\check{\mathbf{y}}^\top\right) = \mathbf{P}_{xy} \tag{12}$$

$$\mathbf{E}_{\mathsf{x}}(\check{\mathbf{x}}) = \mathbf{p}_x \tag{13}$$

$$\mathbf{E}_{\mathsf{x}}\left(\check{\mathbf{x}}\check{\mathbf{x}}^\top\right) = \mathbf{D}_x \tag{14}$$

where $\mathbf{D}_x = \mathrm{diag}(\mathbf{p}_x)$. Using these results, it is possible to define the covariances

$$\mathbf{C}_{xy} = \mathbf{E}_{\mathsf{x},\mathsf{y}}\left(\check{\mathbf{x}}\check{\mathbf{y}}^\top\right) - \mathbf{E}_{\mathsf{x}}(\check{\mathbf{x}})\,\mathbf{E}_{\mathsf{y}}(\check{\mathbf{y}})^\top = \mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top, \tag{15}$$

$$\mathbf{C}_{xx} = \mathbf{E}_{\mathsf{x}}\left(\check{\mathbf{x}}\check{\mathbf{x}}^\top\right) - \mathbf{E}_{\mathsf{x}}(\check{\mathbf{x}})\,\mathbf{E}_{\mathsf{x}}(\check{\mathbf{x}})^\top = \mathbf{D}_x - \mathbf{p}_x\mathbf{p}_x^\top, \tag{16}$$

$$\mathbf{C}_{yy} = \mathbf{E}_{\mathsf{y}}\left(\check{\mathbf{y}}\check{\mathbf{y}}^\top\right) - \mathbf{E}_{\mathsf{y}}(\check{\mathbf{y}})\,\mathbf{E}_{\mathsf{y}}(\check{\mathbf{y}})^\top = \mathbf{D}_y - \mathbf{p}_y\mathbf{p}_y^\top. \tag{17}$$

11. Note that we do *not* require $\Delta_x = \Delta_y$: thus the functions $\mathfrak{K}_i(x)$ and $\mathfrak{K}_j(y)$ below may not be identical (the argument of the function specifies whether Δ_x or Δ_y is used, to simplify notation).

We may therefore define Gaussian random variables $\mathbf{x}_G, \mathbf{y}_G$ with the same covariance structure as $\check{\mathbf{x}}, \check{\mathbf{y}}$, and with mutual information given by (7). We prove in Appendix A.3 that the mutual information for this Gaussian case is

$$I(\mathbf{x}_G; \mathbf{y}_G) = -\frac{1}{2} \log \left(\left| \mathbf{I}_{l_y} - \left(\mathbf{P}_{xy} - \mathbf{p}_x \mathbf{p}_y^\top \right)^\top \mathbf{D}_x^{-1} \left(\mathbf{P}_{xy} - \mathbf{p}_x \mathbf{p}_y^\top \right) \mathbf{D}_y^{-1} \right| \right), \tag{18}$$

which can also be expressed in the singular value form (9). The relation between (18) and (11) is given in the following lemma, which is proved by Bach and Jordan (2002a, Appendix. B.1).

Lemma 12 (The discrete MI approximates the Gaussian MI near independence)
Let $\mathbf{P}_{\hat{\mathrm{x}},\hat{\mathrm{y}}}(i,j) = \mathbf{P}_{\hat{\mathrm{x}}}(i)\,\mathbf{P}_{\hat{\mathrm{y}}}(j)\,(1+\varepsilon_{i,j})$ for an appropriate choice of $\varepsilon_{i,j}$, where $\varepsilon_{i,j}$ is small near independence. Then the second order Taylor expansion of the discrete mutual information in (11) is

$$I(\hat{\mathrm{x}}; \hat{\mathrm{y}}) \approx \frac{1}{2} \sum_{i=1}^{l_x} \sum_{j=1}^{l_y} \mathbf{P}_{\hat{\mathrm{x}}}(i)\,\mathbf{P}_{\hat{\mathrm{y}}}(j)\,\varepsilon_{i,j}^2,$$

which is equal to the second order Taylor expansion of the Gaussian mutual information in (18), namely

$$I(\mathbf{x}_G; \mathbf{y}_G) \approx \frac{1}{2} \sum_{i=1}^{l_x} \sum_{j=1}^{l_y} \mathbf{P}_{\hat{\mathrm{x}}}(i)\,\mathbf{P}_{\hat{\mathrm{y}}}(j)\,\varepsilon_{i,j}^2.$$

3.1.4 Kernel density estimates of the Gaussian mutual information

In this section, we describe a kernel density estimate of the approximate mutual information in (18): this is the point at which our reasoning diverges from the approach of Bach and Jordan (2002a). Before proceeding, we motivate this discussion with a short overview of the Parzen window estimate and its properties, as drawn from Silverman (1986); Duda et al. (2001) (this discussion pertains to the general case of multivariate x, although our application requires only univariate random variables). Given a sample $\boldsymbol{x}$ of size m, each point x_l of which is assumed generated i.i.d. according to some unknown distribution with density $\mathbf{p}_{\mathrm{x}}$, the associated Parzen window estimate of this density is written

$$\widehat{\mathbf{p}}_{\mathrm{x}}(x) = \frac{1}{m} \sum_{l=1}^{m} \kappa(x_l - x).$$

The kernel function[12] $\kappa(x_l - x)$ must be a legitimate probability density function, in that it should be correctly normalised,

$$\int_X \kappa(x)\,dx = 1, \tag{19}$$

and $\kappa(x) \geq 0$. We may rescale the kernel according to $\frac{1}{V_x}\kappa\left(\frac{x}{\sigma_x}\right)$, where the term V_x is needed to preserve (19). Denoting as $V_{x,m}$ the normalisation for a sample size m, then we are guaranteed that the Parzen window estimate converges to the true probability density as long as

$$\begin{aligned} \lim_{m\to\infty} V_{x,m} &= 0, \\ \lim_{m\to\infty} m V_{x,m} &= \infty. \end{aligned}$$

This method requires an initial choice of σ_x for the sample size we start with, which can be obtained by cross validation.

We return now to the problem of empirically estimating the mutual information described in Sections 3.1.2 and 3.1.3. Our estimate is described in the following definition.

12. The reader should not confuse the present kernel with the RKHS kernels introduced earlier. That said, we shall see later that the two kernels are linked.

Definition 13 (Parzen window estimate of the Gaussian mutual information) *A Parzen window estimate of the Gaussian mutual information in (18) is defined as*

$$\widehat{I}(\hat{\mathrm{x}};\hat{\mathrm{y}}) \quad = \quad -\frac{1}{2}\log\left(\prod_{i=1}^{\min(l_x,l_y)}(1+\hat{\rho}_i)(1-\hat{\rho}_i)\right), \tag{20}$$

where $\hat{\rho}_i$ are the singular values of

$$\left(\mathbf{D}_l^{(x)}\right)^{-1/2}\left(\mathbf{K}_l\mathbf{H}(\mathbf{L}_l)^\top\right)\left(\mathbf{D}_l^{(y)}\right)^{-1/2}. \tag{21}$$

Of the four matrices in this definition, $\mathbf{D}_l^{(x)}$ is a diagonal matrix of unnormalised Parzen window estimates of $\mathbf{p}_x$ at the grid points,

$$\mathbf{D}_l^{(x)} = \frac{1}{\Delta_x}\begin{bmatrix} \sum_{l=1}^m \kappa(q_1-x_l) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sum_{l=1}^m \kappa(q_{l_x}-x_l) \end{bmatrix}, \tag{22}$$

$\mathbf{D}_l^{(y)}$ is the equivalent diagonal matrix for $\mathbf{p}_y$,[13] *and*

$$\mathbf{K}_l := \begin{bmatrix} \kappa(q_1-x_1) & \dots & \kappa(q_1-x_m) \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ \kappa(q_{l_x}-x_1) & \dots & \kappa(q_{l_x}-x_m) \end{bmatrix}, \quad \mathbf{L}_l := \begin{bmatrix} \kappa(r_1-y_1) & \dots & \kappa(r_1-y_m) \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ \kappa(r_{l_y}-y_1) & \dots & \kappa(r_{l_y}-y_m) \end{bmatrix}, \tag{23}$$

where we write the above in such a manner as to indicate $l_x \gg m$ and $l_y \gg m$.

Details of how we obtained this definition are given in Appendix A.4. The main disadvantage in using this approximation to the mutual information is that it is exceedingly computationally inefficient, in that it requires a kernel density estimate at each point in a fine grid. In the next section, we show that it is possible to eliminate this grid altogether when we take an upper bound.

3.1.5 The KMI: an upper bound on the mutual information

We now define the kernel mutual information, and show is both a valid dependence criterion (Theorem 15), and an upper bound on the Parzen GMI in Lemma 13 (Theorem 16).

Definition 14 (The kernel mutual information) *The kernel mutual information is defined as*

$$\begin{aligned} \mathrm{KMI}(\boldsymbol{z};F,G) \quad &:= \quad -\frac{1}{2}\log\left(\left|\mathbf{I}-\nu_{\boldsymbol{z}}^{-2}\widetilde{\mathbf{K}}\widetilde{\mathbf{L}}\right|\right) \\ &= \quad -\frac{1}{2}\log\left(\prod_i\left(1-\frac{\gamma_i^2}{\nu_{\boldsymbol{z}}^2}\right)\right), \end{aligned}$$

where γ_i are the non-zero solutions[14] *to*

$$\begin{bmatrix} \mathbf{0} & \widetilde{\mathbf{K}}\widetilde{\mathbf{L}} \\ \widetilde{\mathbf{L}}\widetilde{\mathbf{K}} & \mathbf{0} \end{bmatrix}\begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} = \gamma_i \begin{bmatrix} \mathbf{0} & \widetilde{\mathbf{K}} \\ \widetilde{\mathbf{L}} & \mathbf{0} \end{bmatrix}\begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix}, \tag{24}$$

13. As in our Section 3.1.3 definition of $\mathfrak{K}_i(x)$ and $\mathfrak{K}_j(y)$, we use the notation $\kappa(x)$ and $\kappa(y)$ to denote the Parzen windows for the estimates $\widehat{\mathbf{p}}_x(x)$ and $\widehat{\mathbf{p}}_y(y)$, respectively, even though these may not be identical kernel functions. The argument again indicates which kernel is used.

14. Compare with (4).

the centred Gram matrices $\widetilde{\mathbf{K}}$ *and* $\widetilde{\mathbf{L}}$ *are defined using RKHS kernels obtained via convolution of the associated Parzen windows,*[15]

$$k(x_i,x_j) = \int_X \kappa(x_i - q)\kappa(x_j - q)dq \qquad \text{and} \qquad l(y_i,y_j) = \int_Y \kappa(y_i - r)\kappa(y_j - r)dr,$$

and

$$\nu_{\boldsymbol{z}} = \min\left\{ \min_{j \in \{1\ldots m\}} \sum_{i=1}^{m} \kappa(x_i - x_j), \min_{j \in \{1\ldots m\}} \sum_{i=1}^{m} \kappa(y_i - y_j) \right\}.$$

We note that the above definition bears some similarity to the estimate of Pham (2002). That said, we approximate the mutual information, rather than the entropy; in addition, the KMI is computed in the limit of infinitely small grid size, which removes the need for binning. Thus, we retain our original kernel, rather than using a spline kernel in all cases. This allows us greater freedom to choose a kernel density appropriate to the characteristics of the sources.

The KMI inherits the following important property from the constrained covariance.

Theorem 15 (The KMI is zero if and only if the empirical COCO is zero) *The KMI is zero,* $\mathrm{KMI}(\boldsymbol{z};F,G) = 0$*, if and only if the empirical constrained covariance is zero,* $\mathrm{COCO}(\boldsymbol{z};F,G) = 0$.

Proof This theorem follows from the constrained covariance being the largest eigenvalue γ_i of (24). ■

The relation of the KMI to the mutual information is given by the following theorem, which is the main result of Section 3.

Theorem 16 (The KMI upper bounds the GMI) *Assume that* $X \times Y$ *is chosen to be the support of* $\mathbf{p}_{x,y}$*, that* $\mathbf{p}_{x,y}$ *is bounded away from zero, and that*

$$\begin{aligned} \min_{x\in X} \sum_{i=1}^{m} \kappa(x - x_i) &\approx \min_{j \in \{1\ldots m\}} \sum_{i=1}^{m} \kappa(x_i - x_j) \quad \text{and} \\ \min_{y\in Y} \sum_{i=1}^{m} \kappa(y - y_i) &\approx \min_{j \in \{1\ldots m\}} \sum_{i=1}^{m} \kappa(y_i - y_j) \end{aligned}$$

(the expressions above are alternative, unnormalised estimates of $\min_{x\in X} \mathbf{p}_x(x)$ *and* $\min_{y\in Y} \mathbf{p}_y(y)$*, respectively; the right hand expressions are used so as to obtain the KMI entirely in terms of the sample* $\boldsymbol{z}$*). Then*

$$\mathrm{KMI}(\boldsymbol{z};F,G) \gtrapprox \widehat{I}(\hat{x};\hat{y}). \tag{25}$$

This theorem is proved in Appendix A.5. In particular, the approximate nature of the inequality (25) arises from our use of empirical estimates for lower bounds on $\mathbf{p}_x(x)$ and $\mathbf{p}_y(y)$ (see the proof for details).

3.1.6 The KGV: an alternative upper bound on the mutual information

Bach and Jordan (2002a) propose two related quantities as independence functionals: the kernel canonical correlation (KCC), as discussed in Section 2.4, and the kernel generalised variance (KGV). In this section, we demonstrate that the latter quantity is an upper bound on the KMI under certain conditions. This approach is different to the proof of Bach and Jordan, who employ a limit as the RKHS kernels become infinitely small, and do not make use of Parzen windows. In any event, there may be some problems with this limiting argument: see Appendix B.2 for further discussion. We begin by recalling the definition of the KGV.

15. Recall that $\kappa(x-q)$ may be different from $\kappa(y-r)$, and that the identity of the Parzen window is specified by its argument.

Definition 17 (The kernel generalised variance) *The empirical KGV is defined as*

$$\mathrm{KGV}(\boldsymbol{z}; F, G, \theta) = -\frac{1}{2}\log\left(\prod_i (1-\rho_i^2)\right), \tag{26}$$

where ρ_i *are the solutions to the generalised eigenvalue problem*[16]

$$\begin{bmatrix} \mathbf{0} & \widetilde{\mathbf{K}}\widetilde{\mathbf{L}} \\ \widetilde{\mathbf{L}}\widetilde{\mathbf{K}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} = \rho_i \begin{bmatrix} \theta\widetilde{\mathbf{K}}^2 + \nu_{\boldsymbol{z}}(1-\theta)\widetilde{\mathbf{K}} & \mathbf{0} \\ \mathbf{0} & \theta\widetilde{\mathbf{L}}^2 + \nu_{\boldsymbol{z}}(1-\theta)\widetilde{\mathbf{L}} \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix}, \tag{27}$$

and $\theta \in [0,1]$.

Next, we demonstrate the link between the KGV and the KMI.

Theorem 18 (The KGV upper bounds the KMI) *For all* $\theta \in [0,1]$,

$$\mathrm{KGV}(\boldsymbol{z}; F, G, \theta) \geq \mathrm{KMI}(\boldsymbol{z}; F, G),$$

with equality only at $\theta = 0$, *subject to the conditions*

$$\nu_{\boldsymbol{z}}\mathbf{I} - \widetilde{\mathbf{K}} \succ 0 \qquad \text{and} \qquad \nu_{\boldsymbol{z}}\mathbf{I} - \widetilde{\mathbf{L}} \succ 0. \tag{28}$$

This theorem is proved in Appendix A.6. The requirements (28) should be checked at the point of implementation to guarantee a bound, but we are assured of being able to enforce them: for example, when k is the convolution of (properly normalised) Gaussian kernels κ of size σ, then

$$k(x_i, x_j) = \frac{1}{\sqrt{2\pi(2\sigma^2)}} \exp\left(-\frac{1}{2(2\sigma^2)}(x_j - x_i)^2\right),$$

which is a Gaussian with twice the variance and $1/\sqrt{2}$ the peak amplitude of κ. An upper bound on the spectral norm of $\widetilde{\mathbf{K}}$ is $\max_j \sum_{i=1}^m k(x_i, x_j)$, which follows from Horn and Johnson (1985, Corollary 6.1.5).[17] In other words, even by this conservative estimate, we are assured there exists a $\sigma > 0$ small enough for (28) to hold (the requirements (28) are also sufficient to guarantee the existence of the KMI, since they cause the argument of the logarithm in Definition 14 to be positive).

3.2 Multivariate COCO and KMI

We now describe how our dependence functionals may be generalised to more than two random variables. Let us define the continuous univariate random variables $\mathsf{x}_1, \ldots, \mathsf{x}_n$ on $X_1, \ldots, X_n$, with joint distribution $\mathbf{P}_{\mathsf{x}_1,\ldots,\mathsf{x}_n}$. We also define the associated feature spaces $F_{X_1}, \ldots, F_{X_n}$, each with its corresponding kernel (as in the 2 variable case, the kernels may be different). We begin with a generalisation of the concept of constrained covariance. Our expression takes a similar form to that of Bach and Jordan (2002a, Appendix A.3), although they deal with canonical correlations rather than constrained covariances, which changes the discussion in some respects.

Definition 19 (Empirical multivariate COCO) *Let* $\boldsymbol{z} := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ *be an i.i.d. sample of size m from the joint distribution* $\mathbf{P}_{\mathsf{x}_1,\ldots,\mathsf{x}_n}$. *The multivariate COCO is defined as*

$$\mathrm{COCO}(\boldsymbol{z}; F_{X_1}, \ldots, F_{X_n}) \quad := \quad \max_j \left(|\lambda_j|\right),$$

16. See (5). Note that Bach and Jordan (2002a) handle the scaling differently: they replace the right hand matrix in (27) with $\begin{bmatrix} \widetilde{\mathbf{K}}^2 + \varsigma\widetilde{\mathbf{K}} & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{L}}^2 + \varsigma\widetilde{\mathbf{L}} \end{bmatrix}$ for a regularisation scale ς. We shall see that the form in (27) guarantees the KGV to upper bound the KMI (and hence $\widehat{I}(\hat{x}, \hat{y})$ in (20)).
17. Bearing in mind Lemma 27, and that $\mathbf{H}$ has singular values in $\{1, 0\}$.

where λ_j are the solutions to the generalised eigenvalue problem

$$\begin{bmatrix} \mathbf{0} & \widetilde{\mathbf{K}}_1\widetilde{\mathbf{K}}_2 & \dots & \widetilde{\mathbf{K}}_1\widetilde{\mathbf{K}}_n \\ \widetilde{\mathbf{K}}_2\widetilde{\mathbf{K}}_1 & \mathbf{0} & \dots & \widetilde{\mathbf{K}}_2\widetilde{\mathbf{K}}_n \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{\mathbf{K}}_n\widetilde{\mathbf{K}}_1 & \widetilde{\mathbf{K}}_n\widetilde{\mathbf{K}}_2 & \dots & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c}_{1,j} \\ \mathbf{c}_{2,j} \\ \vdots \\ \mathbf{c}_{n,j} \end{bmatrix} = \lambda_j \begin{bmatrix} \widetilde{\mathbf{K}}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{K}}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \widetilde{\mathbf{K}}_n \end{bmatrix} \begin{bmatrix} \mathbf{c}_{1,j} \\ \mathbf{c}_{2,j} \\ \vdots \\ \mathbf{c}_{n,j} \end{bmatrix}, \tag{29}$$

$\widetilde{\mathbf{K}}_i = \mathbf{H}\mathbf{K}_i\mathbf{H}$, and $\mathbf{K}_i$ is the uncentred Gram matrix of the observations $\boldsymbol{x}_i$ drawn from $\mathbf{P}_{\times_i}$.

This expression is obtained using reasoning analogous to the bivariate empirical COCO in Section 2. The following result justifies using the multivariate COCO as an independence measure.

Lemma 20 (The multivariate COCO measures pairwise independence) *The multivariate constrained covariance is zero if and only if all the empirical pairwise constrained covariances are zero:*
$\mathrm{COCO}(\boldsymbol{z}; F_{X_1}, \dots, F_{X_n}) = 0$ *iff* $\mathrm{COCO}\left(\boldsymbol{x}_i, \boldsymbol{x}_j; F_{X_i}, F_{X_j}\right) = 0$ *for all $i \neq j$.*

We note that although the multivariate COCO only verifies pairwise independence, this is nonetheless sufficient to recover mutually independent sources in the context of linear ICA: see Theorem 24. It is instructive to compare with the KCC-based dependence functional for more than two variables, which uses the smallest eigenvalue of a matrix of correlations (with diagonal terms equal to one, rather than zero), where this correlation matrix has only positive eigenvalues.

We next introduce a generalisation of the kernel mutual information to more than two variables. By analogy with the 2-variable case in Definition 14, we propose the following definition.

Definition 21 (Multivariate KMI) *The kernel mutual information for more than two random variables is defined as*

$$\mathrm{KMI}(\boldsymbol{z}; F_{X_1}, \dots, F_{X_n}) := -\frac{1}{2}\log \prod_{j=1}^{mn} \left(1 + \breve{\lambda}_j\right), \tag{30}$$

where $\nu_{\boldsymbol{z}}\breve{\lambda}_j = \lambda_j$, and

$$\begin{aligned} \nu_{\boldsymbol{z}} &:= \min_{i \in \{1,\dots,n\}} \nu_{\boldsymbol{x}_i}, \text{ where} \\ \nu_{\boldsymbol{x}_i} &:= \min_{j \in \{1 \dots m\}} \sum_{l=1}^{m} \kappa\left(x_{i,l} - x_{i,j}\right). \end{aligned} \tag{31}$$

For (30) to be defined, it is necessary that $1 + \breve{\lambda}_j > 0$ for all j, which is true near independence. The following lemma describes the sense in which the multivariate KMI measures independence.

Lemma 22 (The multivariate KMI measures pairwise independence) *The multivariate KMI is zero if and only if the empirical constrained covariance is zero for every pair of random variables: in other words,*

$$\mathrm{KMI}(\boldsymbol{z}; F_{X_1}, \dots, F_{X_n}) = 0$$

if and only if

$$\mathrm{COCO}\left(\boldsymbol{x}_i, \boldsymbol{x}_j; F_{X_i}, F_{X_j}\right) = 0$$

for all $i \neq j$.

The proof is in Appendix A.7. We now briefly outline how the dependence functional in (30) relates to the KL divergence. In the case of a Gaussian random vector $\mathbf{x}_G$, which can be segmented as $\mathbf{x}_G^\top :=$

$\left[\begin{array}{ccc} \mathbf{x}_{G,1}^\top & \cdots & \mathbf{x}_{G,n}^\top \end{array} \right]$, the KL divergence between the joint distribution of $\mathbf{x}_G$ and the product of the marginal distributions of the $\mathbf{x}_{G,i}$ can be written in terms of the relevant covariance matrices as

$$D_{\mathrm{KL}}\left(\mathsf{p}_{\mathbf{x}_G} \left\| \prod_{i=1}^{n} \mathsf{p}_{\mathbf{x}_{G,i}} \right.\right) = -\frac{1}{2}\log\left(\frac{|\mathbf{C}|}{\prod_{i=1}^{n}|\mathbf{C}_{ii}|}\right),$$

where

$$\begin{aligned} \mathbf{C} &= \mathsf{E}_{\mathbf{x}_G}\left(\mathbf{x}_G\mathbf{x}_G^\top\right) - \mathsf{E}_{\mathbf{x}_G}\left(\mathbf{x}_G\right)\mathsf{E}_{\mathbf{x}_G}\left(\mathbf{x}_G^\top\right), \\ \mathbf{C}_{ii} &= \mathsf{E}_{\mathbf{x}_{G,i}}\left(\mathbf{x}_{G,i}\mathbf{x}_{G,i}^\top\right) - \mathsf{E}_{\mathbf{x}_{G,i}}\left(\mathbf{x}_{G,i}\right)\mathsf{E}_{\mathbf{x}_{G,i}}\left(\mathbf{x}_{G,i}^\top\right). \end{aligned}$$

These results should allow us to generalise the reasoning in Section 3.1, substituting the kernel density estimates

$$\begin{aligned} \widehat{\mathsf{P}}_{\mathsf{x}_i}(x_i) &= \frac{1}{m}\sum_{l=1}^{m}\kappa\left(x_{i,l}-x_i\right), \\ \widehat{\mathsf{P}}_{\mathsf{x}_1,\ldots,\mathsf{x}_n}(x_1,\ldots,x_n) &= \frac{1}{m}\sum_{l=1}^{m}\prod_{i=1}^{n}\kappa\left(x_{i,l}-x_i\right), \end{aligned}$$

and applying the bounding technique of Section 3.1.5 to obtain the quantity in (30); this is a reason for our choosing $\nu_{\mathbf{z}}$ to scale $\breve{\lambda}_j$.[18] The details of this generalisation are beyond the scope of the present work.

4. Implementation and application to ICA

Any practical validation of the independence measures described above is best conducted with respect to some ground truth, in which genuinely independent random variables are tested using the proposed functionals (COCO, KMI). Thus, one test of performance is independent component analysis (ICA): this entails separating independent random variables that have been linearly mixed, using only their property of independence (specifically, we recover the coefficients that describe the linear mixing).

An ICA algorithm using COCO and the KMI comprises two components: the efficient computation of COCO and the KMI, using low rank approximations of the Gram matrices, and gradient descent on the space of linear mixing matrices. These results are summarised from the more detailed discussion by Bach and Jordan (2002a) (although the low rank decomposition is in our case made easier by the absence of the variance term used in the KCC and KGV).

4.1 Efficient computation of kernel dependence functionals

We note that COCO requires us to determine the eigenvalue of maximum magnitude for an $mn \times mn$ matrix (see (29)), and the KMI is a determinant of an $mn \times mn$ matrix, as specified in (30). For any reasonable sample size m, the cost of these computations is prohibitive. We now describe how the computational complexity of this problem may be substantially reduced. First, we note that any positive (semi)definite matrix can be written $\mathbf{K}_i = \mathbf{Z}_i\mathbf{Z}_i^\top$, where $\mathbf{Z}_i$ is lower triangular: this is known as the Cholesky decomposition. If the eigenvalues of the Gram matrix $\mathbf{K}_i$ decay sufficiently rapidly, however, we may make the approximation

$$\mathbf{K}_i \approx \mathbf{Z}_i\mathbf{Z}_i^\top \tag{32}$$

to the Gram matrix $\mathbf{K}_i$, where $\mathbf{Z}_i$ is an $m \times d_i$ matrix; the error due to this approach may be measured via the maximum eigenvalue μ_i of $\mathbf{K}_i - \mathbf{Z}_i\mathbf{Z}_i^\top$. The $\mathbf{Z}_i$ are determined via an *incomplete* Cholesky decomposition, in which the smaller pivots are skipped; symmetric permutation of the rows and columns of $\mathbf{K}_i$

18. On a more pragmatic note, the factor $\nu_{\mathbf{z}}$ generally causes $\left|\breve{\lambda}_j\right| < |\lambda_j|$, which results in $\mathrm{KMI}\left(\mathbf{z}; F_{X_1}, \ldots, F_{X_n}\right)$ being defined further from independence. This is not the only such scaling factor, however.

is used in the course of this process to increase the accuracy and numerical stability of the approximation. This method is applied by Fine and Scheinberg (2001) to decrease the storage and computational requirements of interior point methods in SVMs, and by Bach and Jordan (2002a) for faster computation of the KGV and KCC (pseudocode algorithms may be found in both references). Once the incomplete Cholesky decomposition is accomplished, we can compute the approximate *centred* Gram matrices according to $\widetilde{\mathbf{K}}_i := \mathbf{H}\mathbf{K}_i\mathbf{H} = (\mathbf{H}\mathbf{Z}_i)\left(\mathbf{H}\mathbf{Z}_i^\top\right) = \widetilde{\mathbf{Z}}_i\widetilde{\mathbf{Z}}_i^\top$.

We now show how this low rank decomposition may be used to more efficiently compute the constrained covariance in (29). Substituting

$$\mathbf{d}_{i,j} = \widetilde{\mathbf{Z}}_i^\top \mathbf{c}_{i,j},$$

we get

$$\begin{bmatrix} \mathbf{0} & \widetilde{\mathbf{Z}}_1\widetilde{\mathbf{Z}}_1^\top\widetilde{\mathbf{Z}}_2 & \dots & \widetilde{\mathbf{Z}}_1\widetilde{\mathbf{Z}}_1^\top\widetilde{\mathbf{Z}}_n \\ \widetilde{\mathbf{Z}}_2\widetilde{\mathbf{Z}}_2^\top\widetilde{\mathbf{Z}}_1 & \mathbf{0} & \dots & \widetilde{\mathbf{Z}}_2\widetilde{\mathbf{Z}}_2^\top\widetilde{\mathbf{Z}}_n \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{\mathbf{Z}}_n\widetilde{\mathbf{Z}}_n^\top\widetilde{\mathbf{Z}}_1 & \widetilde{\mathbf{Z}}_n\widetilde{\mathbf{Z}}_n^\top\widetilde{\mathbf{Z}}_2 & \dots & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{d}_{1,j} \\ \mathbf{d}_{2,j} \\ \vdots \\ \mathbf{d}_{n,j} \end{bmatrix} = \lambda_j \begin{bmatrix} \widetilde{\mathbf{Z}}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{Z}}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \widetilde{\mathbf{Z}}_n \end{bmatrix} \begin{bmatrix} \mathbf{d}_{1,j} \\ \mathbf{d}_{2,j} \\ \vdots \\ \mathbf{d}_{n,j} \end{bmatrix}.$$

We may premultiply both sides by[19] $\operatorname{diag}\left(\begin{bmatrix} \widetilde{\mathbf{Z}}_1^\top & \dots & \widetilde{\mathbf{Z}}_n^\top \end{bmatrix}\right)$ without increasing the nullspace of this generalised eigenvalue problem, and we then eliminate $\operatorname{diag}\left(\begin{bmatrix} \widetilde{\mathbf{Z}}_1^\top\widetilde{\mathbf{Z}}_1 & \dots & \widetilde{\mathbf{Z}}_n^\top\widetilde{\mathbf{Z}}_n \end{bmatrix}\right)$ from both sides. Making these changes, we are left with

$$\begin{bmatrix} \mathbf{0} & \widetilde{\mathbf{Z}}_1^\top\widetilde{\mathbf{Z}}_2 & \dots & \widetilde{\mathbf{Z}}_1^\top\widetilde{\mathbf{Z}}_n \\ \widetilde{\mathbf{Z}}_2^\top\widetilde{\mathbf{Z}}_1 & \mathbf{0} & \dots & \widetilde{\mathbf{Z}}_2^\top\widetilde{\mathbf{Z}}_n \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{\mathbf{Z}}_n^\top\widetilde{\mathbf{Z}}_1 & \widetilde{\mathbf{Z}}_n^\top\widetilde{\mathbf{Z}}_2 & \dots & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{d}_{1,j} \\ \mathbf{d}_{2,j} \\ \vdots \\ \mathbf{d}_{n,j} \end{bmatrix} = \lambda_j \begin{bmatrix} \mathbf{d}_{1,j} \\ \mathbf{d}_{2,j} \\ \vdots \\ \mathbf{d}_{n,j} \end{bmatrix}, \tag{33}$$

which is a much more tractable eigenvalue problem, having dimension $\sum_{i=1}^n d_i$. The same procedure may easily be used to recast (30) as the determinant of an $\left(\sum_{i=1}^n d_i\right) \times \left(\sum_{i=1}^n d_i\right)$ matrix. We now briefly consider how to choose the rank d_i for a given precision μ_i: this depends on both the density $\mathbf{p}_{x_i}$ and the kernel $k(x_i, x)$. For Gaussian kernels and densities with exponential decay rates, Bach and Jordan (2002a) show the required precision relates to the rank according to $d_i = O(\log(m/\mu_i))$, which demonstrates the slow increase in rank with sample size. In the case of the KGV and KCC, however, the form of the empirical estimate causes eigenvalues less than approximately $10^{-3}m\kappa/2$ to be discarded, which thus serves as a target precision to ensure the $\mathbf{Z}_i$ retain constant rank regardless of m. We also adopt this threshold in our simulations with the Gaussian kernel, although our motivation is purely a reduction of computational cost.

4.2 Independent component analysis

We describe the goal of instantaneous independent component analysis (ICA), drawing on the numerous existing surveys of ICA and related methods, including those by Hyvärinen et al. (2001); Lee et al. (2000); Cichocki and Amari (2002); Haykin (1998); as well as the review by Comon (1994) of older literature on the topic. We are given m samples $\boldsymbol{t} := (\mathbf{t}_1, \dots, \mathbf{t}_m)$ of the n dimensional random vector $\mathbf{t}$, which are drawn independently and identically from the distribution $\mathsf{P}_\mathbf{t}$. The vector $\mathbf{t}$ is related to the random vector $\mathbf{s}$ (also of dimension n) by the linear mixing process

$$\mathbf{t} = \mathbf{B}\mathbf{s}, \tag{34}$$

where $\mathbf{B}$ is a matrix with full rank. We refer to our ICA problem as being *instantaneous* as a way of describing the dual assumptions that any observation $\mathbf{t}$ depends only on the sample $\mathbf{s}$ at that instant, and that the samples $\mathbf{s}$ are drawn independently and identically.

19. The notation $\operatorname{diag}\left(\begin{bmatrix} \widetilde{\mathbf{Z}}_1^\top & \dots & \widetilde{\mathbf{Z}}_n^\top \end{bmatrix}\right)$ defines a matrix with blocks $\widetilde{\mathbf{Z}}_i^\top$ along the diagonal, and zeros elsewhere. The matrix need not be square, however, and the diagonal is in this case defined in a manner consistent with the asymmetry of the $\widetilde{\mathbf{Z}}_i^\top$.

The components s_i of s are assumed to be mutually independent: this model codifies the assumption that the sources are generated by unrelated phenomena (for instance, one component might be an EEG signal from the brain, while another could be due to electrical noise from nearby equipment). Mutual independence (in the case where the random variables admit probability densities) has the following definition (Papoulis, 1991):

Definition 23 (Mutual independence) *Suppose we have a random vector* s *of dimension* n*. We say that the components* s_i *are* mutually *independent if and only if*

$$\mathbf{p}_{\mathsf{s}}(\mathbf{s}) = \prod_{i=1}^{n} \mathbf{p}_{\mathsf{s}_i}(s_i). \tag{35}$$

It follows easily that the random variables are pairwise *independent if they are* mutually *independent; i.e.* $\mathbf{p}_{\mathsf{s}_i}(s_i)\,\mathbf{p}_{\mathsf{s}_j}(s_j) = \mathbf{p}_{\mathsf{s}_i,\mathsf{s}_j}(s_i, s_j)$ *for all* $i \neq j$*. The reverse does* not *hold, however: pairwise independence does not guarantee mutual independence.*

Our goal is to recover s via an estimate $\mathbf{W}$ of the inverse of the matrix $\mathbf{B}$, such that the recovered vector $\mathsf{x} = \mathbf{WBs}$ has mutually independent components.[20] For the purpose of simplifying our discussion, we will assume that $\mathbf{B}$ (and hence $\mathbf{W}$) is an *orthogonal matrix*; in the case of arbitrary $\mathbf{B}$, the observations must first be decorrelated before an orthogonal $\mathbf{W}$ is applied (Hyvärinen et al., 2001). In our experiments, however, we will deal with general mixing matrices.

Mutual independence is generally difficult to determine. In the case of linear mixing, however, we are able to find a unique optimal unmixing matrix $\mathbf{W}$ using only the *pairwise* independence between elements of x, which is equivalent to recovering the *mutually* independent terms of s (up to permutation and scaling). This is due to the following theorem (Comon, 1994, Theorem 11).

Theorem 24 (Mutual independence in linear ICA) *Let* s *be a vector of dimension* n *with mutually independent components, of which at most one is Gaussian, and for which the underlying densities do not contain delta functions. Let* x *be a random vector related to* s *according to* $\mathsf{x} = \mathbf{As}$*, where* $\mathbf{A}$ *is an orthogonal* $n \times n$ *matrix.*[21] *Then the properties*

- *The components of* x *are pairwise independent*
- *The components of* x *are mutually independent*
- $\mathbf{A} = \mathbf{PS}$*, where* $\mathbf{P}$ *is a permutation matrix, and* $\mathbf{S}$ *a diagonal matrix*

are equivalent.

We acknowledge that the application of a general dependence function to linear ICA is not guaranteed to be an optimal non-parametric approach to the problem of estimating the entries in $\mathbf{B}$—for instance, Samarov and Tsybakov (2004) provide a method that guarantees $\sqrt{n}$-consistent estimates of the columns of $\mathbf{B}$ under certain smoothness assumptions on the source densities, which is a more natural goal in view of the mixing model (34). Indeed, most specialised ICA algorithms exploit the linear mixing structure of the problem to avoid having to employ a general measure of independence, which makes the task of recovering $\mathbf{B}$ 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). In addition, the kernel methods yield better experimental performance than other specialised ICA approaches

20. It turns out that the problem described above is indeterminate in certain respects. For instance, our measure of independence does not change when the ordering of elements in x is swapped, or when components of x are scaled by different constant amounts. Thus, source recovery takes place up to these invariances.
21. For the purposes of ICA, $\mathbf{A}$ combines both the mixing and unmixing processes, *i.e.*, $\mathbf{A} = \mathbf{WB}$.

(including recent state-of-the-art algorithms) in our tests of outlier resistance and musical source separation (see Section 5).

We also note at this point that if elements $\mathbf{t}_i, \mathbf{t}_j$ in the sample $\boldsymbol{t}$ are *not* drawn independently for $i \neq j$ (for instance, if they are generated by a random process with non-zero correlation between the outputs at different times), then an entirely different set of approaches can be brought to bear (see for instance Belouchrani et al., 1997; Pham and Garat, 1997).[22] Although the present study concentrates entirely on the i.i.d. case, we will briefly address random processes with time dependencies in Section 6, when describing possible extensions to our work. Finally, we draw attention to an alternative ICA setting, as described by Cardoso (1998b); Theis (2005), in which $\mathbf{s}$ is partitioned into mutually independent vectors (which might each have internal dependence structure): we wish to recover these vectors following linear mixing. As pointed out by Bach and Jordan (2002a), kernel dependence functionals are well suited to this problem, since they also apply straightforwardly to multivariate random variables: it suffices to define appropriate Gram matrices.

4.3 Gradient descent on the Stiefel manifold

We now describe the method used to minimise our kernel dependence functionals over possible choices of the orthogonal demixing matrix $\mathbf{W}$. The manifold described by $n \times p$ matrices $\mathbf{A}$ for which $\mathbf{A}^\top \mathbf{A} = \mathbf{I}$, where $n \geq p$, is known as the *Stiefel manifold*. Gradient descent for functions defined on this manifold is described by Edelman et al. (1998), and Bach and Jordan (2002a) applied this technique to kernel ICA. A clear and intuitive explanation of this procedure is also given by Hyvärinen and Plumbley (2002). Let $f(\mathbf{W},\boldsymbol{tt})$ be the particular dependence functional (COCO or KMI) on which we wish to do gradient descent, where $\boldsymbol{t} := (\mathbf{t}_1, \ldots, \mathbf{t}_m)$ are the whitened, mixed observations. A naive gradient descent algorithm would involve computing the derivative

$$\mathbf{G} := \frac{\partial f(\mathbf{W},\boldsymbol{t})}{\partial \mathbf{W}},$$

updating $\mathbf{W}$ according to $\mathbf{W} \rightarrow \mathbf{W} + \mu\mathbf{G}$ (where μ is chosen to minimise $f(\mathbf{W} + \mu\mathbf{G},\boldsymbol{t})$), and projecting the resulting matrix back onto the Stiefel manifold. This might not be particularly efficient, however, in that the update can largely be cancelled by the subsequent projection operation. Instead, we attempt to find the direction of steepest descent on the Stiefel manifold, and to perform our update with the constraint that we remain on this manifold. To achieve this, we first describe the set of perturbations to $\mathbf{W}$ that retain the orthogonality of $\mathbf{W}$, then choose the direction of steepest descent/ascent within this set, and finally give the expression that parameterises the shifts along the geodesic[23] in this direction.

Let $\boldsymbol{\Delta}$ be a perturbation with small norm to the orthogonal matrix $\mathbf{W}$, such that $\mathbf{W} + \boldsymbol{\Delta}$ remains on the Stiefel manifold. For this constraint to hold, we require

$$(\mathbf{W} + \boldsymbol{\Delta})^\top (\mathbf{W} + \boldsymbol{\Delta}) = \mathbf{I}, \text{ which implies} \tag{36}$$

$$\mathbf{W}^\top \boldsymbol{\Delta} + \boldsymbol{\Delta}^\top \mathbf{W} \approx \mathbf{0}; \tag{37}$$

in other words, $\mathbf{W}^\top \boldsymbol{\Delta}$ is skew-symmetric. To find the particular $\boldsymbol{\Delta}$ that gives the direction of steepest change of $f(\mathbf{W},\boldsymbol{t})$, we solve

$$\boldsymbol{\Delta}_{\max} := \arg\max_{\boldsymbol{\Delta}} f(\mathbf{W} + \boldsymbol{\Delta},\boldsymbol{t}),$$

subject to $\mathrm{tr}(\boldsymbol{\Delta}^\top \boldsymbol{\Delta}) = \text{const}$ and (37). This yields

$$\boldsymbol{\Delta}_{\max} = \mathbf{G} - \mathbf{W}\mathbf{G}^\top \mathbf{W},$$

where the proof is provided by Edelman et al. (1998); Hyvärinen and Plumbley (2002). Finally, if we use q to parameterise displacement along a geodesic in the direction $\boldsymbol{\Delta}_{\max}$ from an initial matrix $\mathbf{W}(0)$, then the resulting $\mathbf{W}(q)$ is given by

$$\mathbf{W}(q) = \mathbf{W}(0) \exp\left(q\mathbf{W}(0)^\top \boldsymbol{\Delta}_{\max}\right).$$

22. In particular, it becomes possible to separate Gaussian processes when they are correlated over time.
23. A geodesic represents the shortest path on a manifold between two points; equivalently, the acceleration involved in moving between two points along a geodesic is perpendicular to the manifold when constant velocity is maintained.

As in the implementation of Bach and Jordan (2002a), we determine an approximation of the gradient of $f(\mathbf{W},\boldsymbol{t})$ by making small perturbations to $\mathbf{W}$ about each possible Jacobi rotation, and recomputing f for each such perturbation. Gradient descent is then accomplished using a Golden search along this direction of steepest descent.

Finally, we note that procedures are given by Edelman et al. (1998) to compute the Hessian on the Stiefel manifold, as are implementations of Newton's method and conjugate gradient descent. In addition, an adaptive algorithm for gradient descent on the Stiefel manifold is proposed by Zhu and Zhang (2002). The application of these methods to improve the performance of our algorithm is beyond the scope of the present work.

4.4 Computational cost

We conclude this section with a summary of the overall computational cost of ICA based on COCO and the KMI: this analysis draws directly from the assessment of Bach and Jordan (2002a, Section 6), since COCO and the KMI cost effectively the same as the KCC and KGV, respectively. The first step in ICA, which is not discussed here, is the decorrelation of the sources (as described for instance by Hyvärinen et al., 2001), which has a cost $\mathrm{O}(mn^2)$. We next consider the cost of computing the multivariate COCO and KMI. In both approaches, each of the n sources requires an estimate of its $m \times m$ Gram matrix using incomplete Cholesky decomposition, which costs $\mathrm{O}(md^2)$, where d is the largest rank retained in the computation of the $\mathbf{Z}_i$ in (32): the net cost is $\mathrm{O}(mnd^2)$. These $\mathbf{Z}_i$ are then centred and assembled into the matrix in (33), which entails $n(n-1)/2$ operations each costing $\mathrm{O}(md^2)$, for an overall cost $\mathrm{O}(mn^2d^2)$. COCO is given by the largest eigenvalue of this matrix, and costs $\mathrm{O}(n^2d^2)$; the KMI is a determinant, and costs $\mathrm{O}(n^3d^3)$.

We compute the gradient of the kernel dependence measures using the method of finite differences (as described in the previous section), which necessitates $n(n-1)/2$ evaluations of the measure used. In each evaluation, we need only compute two incomplete Cholesky decompositions (we cache the remainder); the assembly of the matrix in (33) then entails $2n-3$ matrix products, for an overall cost (Cholesky + matrix assembly for all the Jacobi rotations) of $\mathrm{O}(mn^3d^2)$. The eigenvalue computations used to obtain the gradient of COCO cost $\mathrm{O}(n^4d^2)$, and the determinants used in the KMI gradient cost $\mathrm{O}(n^5d^3)$.

5. Experimental results on ICA

In this section, we examine the performance of our independence functionals (COCO, KMI) as it compares to the KGV and KCC, when used to address the problem of linear instantaneous ICA. Since the objective is to find an estimate $\mathbf{W}$ of the *inverse* of the mixing matrix $\mathbf{B}$ (the reader is referred to Section 4.2 for a description of the ICA problem), we require a measure of distance between our approximation and the true inverse: this is given by the *Amari divergence*, which is introduced in Section 5.1. Next, in Section 5.2, we present results obtained when separating a range of artificial signals mixed using randomly generated matrices, including cases in which the observations are corrupted by noise. Finally, we describe our attempts at separating artificial mixtures of audio signals representing a number of musical genres. Results are compared with those obtained using standard methods (FastICA, Jade, Infomax) and recent state-of-the-art methods (RADICAL, CFICA), as well as the KCC and KGV.

5.1 Measurement of performance

We use the Amari divergence, defined by Amari et al. (1996), as an index of ICA algorithm performance: this is an adaptation and simplification of a criterion proposed earlier by Comon (1994). Note that the properties of this quantity in Lemma 26 were not described by Amari et al. (1996), but follow from the proof of Comon (1994).

Definition 25 (Amari divergence) *Let* $\mathbf{B}$ *and* $\mathbf{W}$ *be two* $n \times n$ *matrices, where* $\mathbf{B}$ *is the mixing matrix and* $\mathbf{W}$ *the estimated unmixing matrix (these need not be orthogonal here), and let* $\mathbf{D} = \mathbf{WB}$. *Then the Amari*

divergence between $\mathbf{B}$ *and* $\mathbf{W}$ *is*

$$D(\mathbf{WB}) = \frac{100}{2n(n-1)} \sum_{i=1}^{n} \left(\frac{\sum_{j=1}^{n} |d_{i,j}|}{\max_j |d_{i,j}|} - 1 \right) + \frac{1}{2n(n-1)} \sum_{j=1}^{n} \left(\frac{\sum_{i=1}^{n} |d_{i,j}|}{\max_i |d_{i,j}|} - 1 \right).$$

Although this measure is not, strictly speaking, a distance metric for general matrices $\mathbf{B}, \mathbf{W}$, it nonetheless possesses certain useful properties, as shown below.

Lemma 26 (Properties of the Amari divergence) *The Amari divergence* $D(\mathbf{WB})$ *between the* $n \times n$ *matrices* $\mathbf{B}, \mathbf{W}$ *has the following properties:*

- $0 \leq D(\mathbf{WB}) \leq 100$. *The factor of 100 is not part of the original definition of Amari et al. (1996), who defined the Amari divergence on* $[0,1]$. *In our experiments, however, the Amari divergence was generally small, and we scaled it by 100 to make the results tables more readable.*
- *Let* $\mathbf{P}$ *be an arbitrary permutation matrix (a matrix with a single* 1 *in each row and column, and with remaining entries* 0*), and* $\mathbf{S}$ *be a diagonal matrix of non-zero scaling factors. Then* $\mathbf{W} = \mathbf{B}^{-1}$ *if and only if* $D(\mathbf{WB}) = 0$*, or equivalently* $D(\mathbf{WBSP}) = 0$ *or* $D(\mathbf{SPWB}) = 0$.

The final property in the above Lemma is particularly useful in the context of ICA, since it causes our performance measure to be invariant to output ordering ambiguity once the sources have been demixed (see Theorem 24).

5.2 Experiments and performance assessment

Since our main purpose is to compare the performance with that reported by Bach and Jordan (2002a), we generated our test distributions independently following their descriptions. A list of the distributions used in our experiments, and their respective kurtoses, is given in Table 3. While these distributions represent a broad range of behaviours, we note that negative kurtoses predominate, which should be borne in mind when evaluating performance. We used the KGV and KCC Matlab implementations downloadable from (Bach and Jordan) (thus, we employ the KGV as originally defined by Bach and Jordan (2002a), and not the version described in Section 3.1.6). The precision of the incomplete Cholesky decomposition, used to approximate the Gram matrices for the kernel dependence functionals, was set at $\eta := \varepsilon n$; our choice of ε represents a tradeoff between accuracy and computation speed. Unless otherwise specified, the kernel algorithm results were refined in a "polishing step", in which the kernel size was halved upon convergence, and the gradient descent procedure recommenced with this smaller kernel. This polishing was carried out since the larger kernel size results in the kernel dependence measures being a smoother function of the estimated unmixing matrix, making it easier to find the global minimum; but making the location of this global minimum less precise than obtained with a smaller kernel. The polishing step usually caused a measurable improvement in our results.

As well as the kernel algorithms, we compare with three standard ICA methods: FastICA (Hyvärinen et al., 2001), Jade (Cardoso, 1998a), and Infomax (Bell and Sejnowski, 1995); and two more sophisticated methods, neither of them based on kernels: RADICAL (Miller and Fisher III, 2003), which uses order statistics to obtain entropy estimates; and characteristic function based ICA (CFICA) (Chen and Bickel, 2004). [24] 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 FastICA 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. In the case of FastICA, we used the nonlinearity most appropriate to the signal characteristics: this was generally the kurtosis based contrast, since the predominantly negative kurtoses in Table 3 made this a good choice (see Hyvärinen et al., 2001). In some experiments, however, the

24. We are aware that Chen and Bickel 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.

Table 3: Labels of distributions used, and their respective kurtoses. All distributions have zero mean and unit variance.

Label	Definition	Kurtosis
a	Student's t distribution, 3 DOF	∞
b	Double exponential	3.00
c	Uniform	-1.20
d	Students's *t* distribution, 5 DOF	6.00
e	Exponential	6.00
f	Mixture, 2 double exponentials	-1.70
g	Symmetric mixture 2 Gauss., multimodal	-1.85
h	Symmetric mixture 2 Gauss., transitional	-0.75
i	Symmetric mixture 2 Gauss., unimodal	-0.50
j	Asymm. mixture 2 Gauss., multimodal	-0.57
k	Asymm. mixture 2 Gauss., transitional	-0.29
l	Asymm. mixture 2 Gauss., unimodal	-0.20
m	Symmetric mixture 4 Gauss., multimodal	-0.91
n	Symmetric mixture 4 Gauss., transitional	-0.34
o	Symmetric mixture 4 Gauss., unimodal	-0.40
p	Asymm. mixture 4 Gauss., multimodal	-0.67
q	Asymm. mixture 4 Gauss., transitional	-0.59
r	Asymm. mixture 4 Gauss., unimodal	-0.82

kurtosis was unsuited to the source characteristics, in which case we signal our alternative choice of nonlinearity. The Infomax algorithm selects its contrast automatically based on the super- or sub-Gaussianity of the signal, and does not require manual contrast choice. Likewise Jade uses only a kurtosis-based contrast, and thus does not require the user to choose a demixing function.

We begin with a brief investigation into the form taken by the various kernel dependence functionals for a selection of the data in Table 3. Contours of the KGV, COCO, KMI, and Amari divergence are plotted in Figure 1, which describes the demixing of samples from three distributions, combined using a product of known Jacobi rotations. All kernel functionals in this demonstration were computed with a Gaussian RBF kernel,

$$k_G(x,x') = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}\|x-x'\|^2\right). \tag{38}$$

We observe that each of the functionals exhibits local minima at locations distant from independence, but that each possesses a "basin of attraction" in the vicinity of the correct answer. Moreover, we note that each of the functionals is smooth (given the choice of kernel size), and that the global minima are fairly symmetric. For these reasons, the gradient descent algorithm described in Section 4.3 should converge rapidly to the global optimum, given a reasonable initialisation point. Our solution method differs from that of Bach and Jordan (2002a), however, in that we generally use Jade (unless specified otherwise) to initialise the kernel functionals (COCO, KCC, KGV, KMI), whereas Bach and Jordan only do this when separating large numbers of signals (in most cases, they initialise using a one-unit kernel dependence functional with deflation, and with a less costly polynomial kernel). For more than two signals, this process is repeated several times, starting from different initialising matrices. While Jade is less computationally costly as an initialisation method, it might be less reliable in certain cases (where the sources are near-Gaussian, or when a large number of outliers exist due to noise, both of which can cause Jade to misconverge).

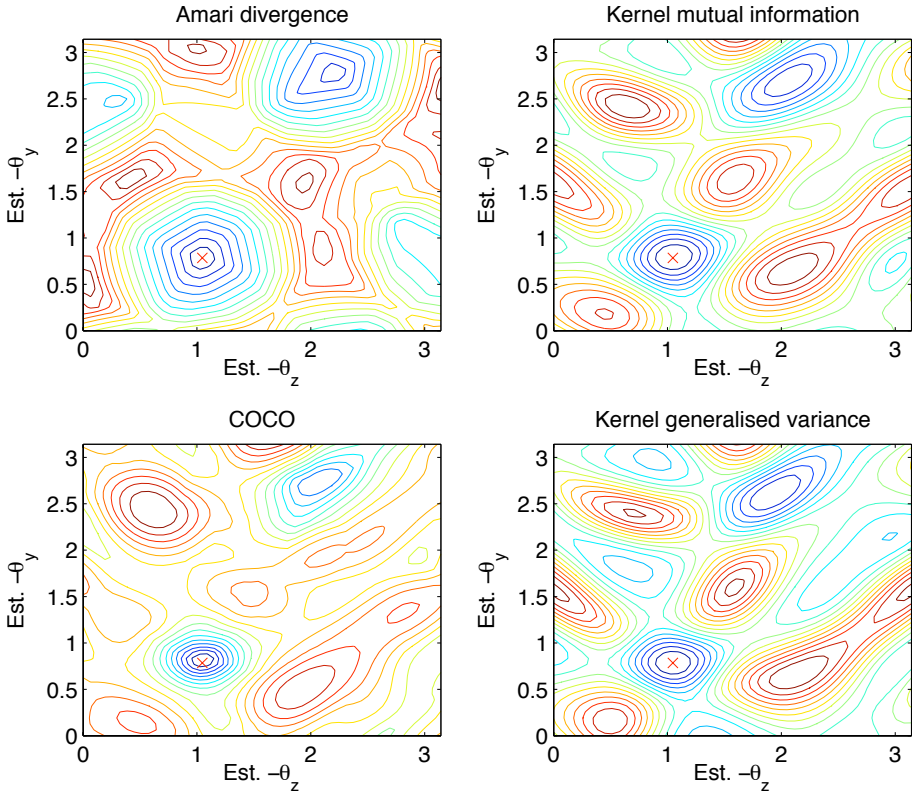

Figure 1: Contour plots of kernel independence functionals. Top left: Amari divergence. Top right: kernel mutual information. Bottom left: constrained covariance. Bottom right: kernel generalised variance. Three signals of length 1000 and with respective distributions g, k, and q (this choice was random) were combined using a 3×3 orthogonal rotation matrix. This matrix was expressed as a product of Jacobi rotations $\mathbf{B} = \mathbf{R}_z(\theta_z)\mathbf{R}_y(\theta_y)\mathbf{R}_x(\theta_x)$, where $\theta_x = -\pi/6$, $\theta_y = -\pi/4$, and $\theta_z = -\pi/3$; the subscript denotes the axis about which the rotation occurs. An estimate $\mathbf{W} = \mathbf{R}_x(-\theta_x)\mathbf{R}_y(\hat{\theta}_y)\mathbf{R}_z(\hat{\theta}_z)$ of $\mathbf{B}^{-1}$ was made, in which $\hat{\theta}_y$ and $\hat{\theta}_z$ took values in the range $[0,\pi]$. The red "x" in each plot is located at the coordinates $(-\theta_z,-\theta_y)$ corresponding to the optimal estimate of $\mathbf{B}$. A Gaussian kernel of size $\sigma^2 = 1$ was used in all cases, and $\kappa = 10^{-3}$ for the KGV.

5.3 General mixtures of artificial data

We now describe the ICA experiments performed with the distributions in Table 3, where the Amari divergence is used to measure the closeness of the estimated mixing matrix to the true matrix. Kernels used include the Gaussian RBF kernel in (38), and the Laplace kernel,

$$k_L(x,x') = \frac{\lambda}{2}\exp(-\lambda\|x-x'\|).$$

We combined the independent sources using random mixing matrices, with condition numbers between 1 and 2, and then whitened the resulting observations before estimating the orthogonal de-mixing matrix. [25]

Our first experiment consisted in de-mixing data drawn independently from 2-16 sources chosen at random with replacement from Table 3. Results are given in Table 4. The KMI with Gaussian kernel matches or exceeds KGV performance in the final four experiments; and, with the Laplace kernel, in five of the seven experiments. Moreover, the KMI yields performance statistically indistinguishable from RADICAL in four of the seven experiments.[26] On the other hand, the KGV outperforms the KMI in the first and third case, where the number m of samples is small (although in the $n = 4, m = 1000$ case, the difference is not statistically significant). The superior performance of the Laplace kernel compared with the Gaussian may be due to its slower decaying spectrum, which allows dependence encoded at higher frequencies in the source density to induce a greater departure of COCO from zero (making this dependence easier to detect): see Gretton et al. (2005b, Section 4.2). The Laplace kernel has a greater computational cost, however, 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 to maintain good performance. Finally, the extended Infomax algorithm seems unable to separate the signals in 250 sample, 2 signal case: the Amari divergence was spread almost uniformly over the range $[0, 100]$.

5.4 Performance on difficult artificial problems

In our next experiment, we investigated the effect of outlier noise added to the observations. We selected two generating distributions from Table 3, randomly and with replacement. After combining these signals with a randomly generated matrix with 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. In the case of COCO, this can be explained by functions in the Laplace RKHS having less penalisation at high frequencies, causing the functions attaining the supremum in Definition 2 to adapt to (and be affected by) outliers to a greater degree than functions in the Gaussian RKHS (the KMI is also subject to this effect). Results are shown in the left hand plot in Figure 2. 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 by Bach and Jordan (2002a): this resulted in an improvement in performance (broadly speaking, an increase in κ causes the KGV to approach the KMI, and the KCC to approach COCO).[27] It is clear that the kernel methods substantially outperform both the standard and recent alternatives in outlier resistance (we omitted the remaining standard methods, since their performance was worse than FastICA).

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 the right hand plot of Figure 2 that too small a κ can cause severe underperformance for the KCC and KGV. On the other hand, κ is required to be

25. We did not use simple orthogonal matrices to mix our sources, since this would have lowered the variance in our estimate of $\mathbf{W}$, making the problem (slightly) easier than that of estimating a truly random mixing matrix (Cardoso, 1998a).
26. The mean performance of the various methods, both kernel and otherwise, is affected in some experiments by a small number of misconverged results with large Amari divergence (although misconvergence of the kernel methods does not always correspond to misconvergence of the Jade initialisation). These results may arise from diversion to local minima, causing an increase in the overall mean Amari divergence that does not reflect the typical behaviour of the kernel algorithms. Such outliers occur less often, or not at all, at larger sample sizes, as can be seen by the decreased variance in these cases.
27. The results presented here for the KCC and KGV also improve on those of Miller and Fisher III (2003); Bach and Jordan (2002a) since they include a polishing step for the KCC and KGV, which was not carried out in these earlier studies.

small for good performance at large sample sizes in Table 4. A major advantage of COCO and the KMI is that these do not require any additional tuning beyond the selection of a kernel.

Our third experiment addresses the effects of low kurtosis, since many ICA methods rely (sometimes implicitly, through their choice of nonlinearity) on the kurtosis as an index of signal independence. Two samples were drawn from a single distribution, consisting of a mixture of two Gaussians with means $+5$ and -5 and unit variance, with a selection of mixture weights chosen such that, following normalisation of the overall sample to zero mean and unit variance, the (empirical) kurtosis took on a range of positive, near-zero, and negative values. Results are given in Figure 3. All kernel based methods were unaffected by near-zero kurtosis, as were CFICA and RADICAL; the remaining ICA methods do less well (Infomax was omitted since it performed worse than Jade).

Table 4: Illustration of the demixing of n randomly chosen signals of length m, drawn independently with replacement from Table 3. For COCO and the KMI, we used a Gaussian kernel of size $\sigma = 1$ in the experiments labelled (g), and a Laplace kernel of size $\lambda = 3$ for those experiments labelled (l). In the case of the KCC and KGV, we used $\sigma = 1$ and $\kappa = 2 \times 10^{-2}$ for signals of length $m \leq 1000$, and $\sigma = 0.5$ and $\kappa = 2 \times 10^{-3}$ for the remaining signals. In all cases, we used $\varepsilon = 1 \times 10^{-5}$ for the Gaussian kernels, and $\varepsilon = 0.01$ for the Laplace kernels. We initialised the kernel methods with Jade in all cases but $n = 16$, for which we used FastICA (due to its more stable output). The performance figures are an average over *Rep.* independent runs. The best results are shown in boldface, as are those results statistically indistinguishable from the best according to a level 0.05 left-tailed paired difference t-test.

n	m	Rep.	Fica	Jade	Imax	CFICA	RAD	KCC	COCO(g)	COCO(l)	KGV	KMI(g)	KMI(l)
2	250	1000	10.5 ± 0.4	9.5 ± 0.4	44.4 ± 1	7.2 ± 0.3	$\mathbf{5.4 \pm 0.2}$	7.0 ± 0.3	7.8 ± 0.3	7.0 ± 0.3	$\mathbf{5.3 \pm 0.2}$	6.0 ± 0.2	5.7 ± 0.2
2	1000	1000	6.0 ± 0.3	5.1 ± 0.2	11.3 ± 0.6	3.2 ± 0.1	$\mathbf{2.4 \pm 0.1}$	3.3 ± 0.1	3.5 ± 0.1	2.9 ± 0.1	$\mathbf{2.3 \pm 0.1}$	2.6 ± 0.1	$\mathbf{2.3 \pm 0.1}$
4	1000	100	5.7 ± 0.4	5.6 ± 0.4	13.3 ± 1	3.3 ± 0.2	$\mathbf{2.5 \pm 0.1}$	4.5 ± 0.4	4.2 ± 0.3	4.6 ± 0.6	$\mathbf{3.1 \pm 0.6}$	4.0 ± 0.7	$\mathbf{3.5 \pm 0.7}$
4	4000	100	3.1 ± 0.2	2.3 ± 0.1	5.9 ± 0.7	1.5 ± 0.1	$\mathbf{1.3 \pm 0.1}$	2.4 ± 0.5	1.9 ± 0.1	1.6 ± 0.1	1.4 ± 0.1	1.4 ± 0.05	$\mathbf{1.2 \pm 0.05}$
8	2000	50	4.1 ± 0.2	3.6 ± 0.2	9.3 ± 0.9	2.4 ± 0.1	$\mathbf{1.8 \pm 0.1}$	4.8 ± 0.9	3.7 ± 0.9	5.2 ± 1.3	2.6 ± 0.3	2.1 ± 0.1	1.9 ± 0.1
8	4000	50	3.2 ± 0.2	2.7 ± 0.1	6.4 ± 0.9	1.6 ± 0.1	$\mathbf{1.3 \pm 0.05}$	2.1 ± 0.2	2.0 ± 0.1	1.9 ± 0.1	1.7 ± 0.2	1.5 ± 0.1	$\mathbf{1.3 \pm 0.05}$
16	5000	25	2.9 ± 0.1	3.1 ± 0.3	9.4 ± 1.1	1.7 ± 0.1	$\mathbf{1.2 \pm 0.05}$	3.7 ± 0.6	2.4 ± 0.1	2.6 ± 0.2	1.7 ± 0.1	1.5 ± 0.1	1.5 ± 0.1

Table 5: Illustration of the demixing of n music segments of length $m = 55272$, taken from the collection of 17 music samples at (Pearlmutter). The $n = 2$ case represents an average over 136 samples, and the $n = 4$ case is an average over 120 samples. Details of the KGV and KMI parameters may be found in Section 5.5. The best results are shown in boldface, as those results statistically indistinguishable from the best according to a level 0.05 left-tailed paired difference t-test.

n	Fica	Jade	Imax	CFICA	RADICAL	KGV	KMI
2	0.92 ± 0.07	0.99 ± 0.07	1.07 ± 0.10	$\mathbf{0.84 \pm 0.06}$	1.02 ± 0.07	$\mathbf{0.65 \pm 0.05}$	$\mathbf{0.51 \pm 0.13}$
4	0.93 ± 0.03	0.87 ± 0.03	1.09 ± 0.06	0.89 ± 0.03	0.91 ± 0.03	$\mathbf{0.62 \pm 0.02}$	0.68 ± 0.03

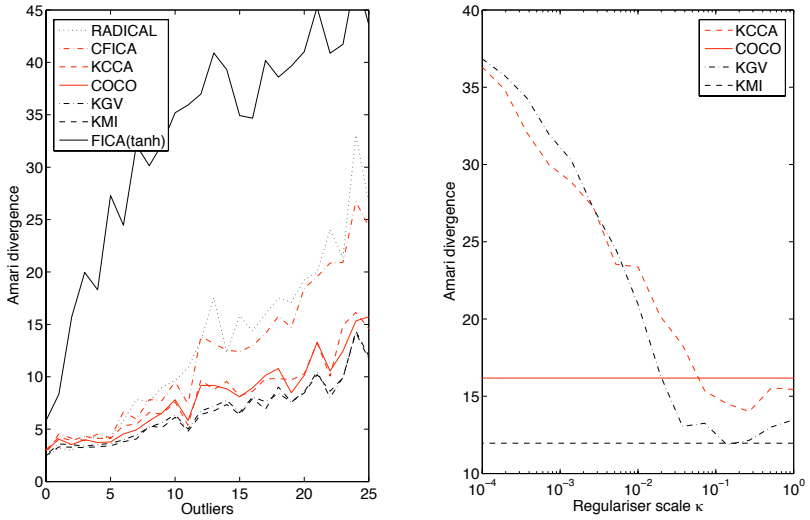

Figure 2: **Left:** Effect of outliers on the performance of the ICA algorithms, for two sources of length $m = 1000$, drawn independently with replacement from Table 3, and corrupted at random observations with outliers at ± 5 (where each sign has probability 0.5). Each point represents an average over 100 independent experiments. The number of corrupted observations in *both* signals is given on the horizontal axis. The kernel methods used $\sigma = 1$, $\varepsilon = 2 \times 10^{-5}$, and $\kappa = 0.11$ (KCC and KGV only). The tanh nonlinearity was used for the FastICA algorithm, since this is more resistant to outliers than the kurtosis (Hyvärinen, 1997). **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.

5.5 Audio signal demixing

Our final experiment involved demixing brief extracts from various musical sources, which were combined using a randomly generated matrix (in the same manner as the artificial signals described in the previous section). A total of 17 different extracts were taken from the ICA benchmark set at (Pearlmutter). These consist of 5 second segments sampled at 11 kHz with a precision of 8 bits, and represent a wide variety of musical genres. While samples of a musical signal are certainly not generated independently and identically in time, many ICA algorithms have nonetheless been applied successfully to this problem, which is why we investigate this benchmark. Indeed, many practical applications of ICA are in a context where complete independence of the unmixed signals is *not* a goal, in theory or in practice: rather, the objective of the linear unmixing is to obtain signals that are relatively "more independent" than the original observations, in the hope that these will be physically interpretable in the light of the system generating the data.

A summary of our results is given in Table 5: the KMI, KGV, and CFICA are statistically indistinguishable for two extracts, and the KGV does best with four extracts, followed by the KMI. In the $n = 2$ case, every possible combination of two different extracts was investigated (for a total of 136 experiments), and the results averaged. We used $\kappa = 2 \times 10^{-3}$, $\sigma = 0.5$, $\varepsilon = 1 \times 10^{-5}$, and a Gaussian kernel for the KGV; and $\lambda = 3$, $\varepsilon = 1 \times 0.01$, and a Laplace kernel for the KMI. In both cases, a polishing step was applied to refine the result. For each experiment with $n = 4$, music segments were drawn randomly and without replacement from the 17 available extracts, and the results averaged over 120 repetitions. All kernel algorithm parameters were the same as in the $n = 2$ case besides the Laplace kernel size, which was increased to $\lambda = 4$. In addition,

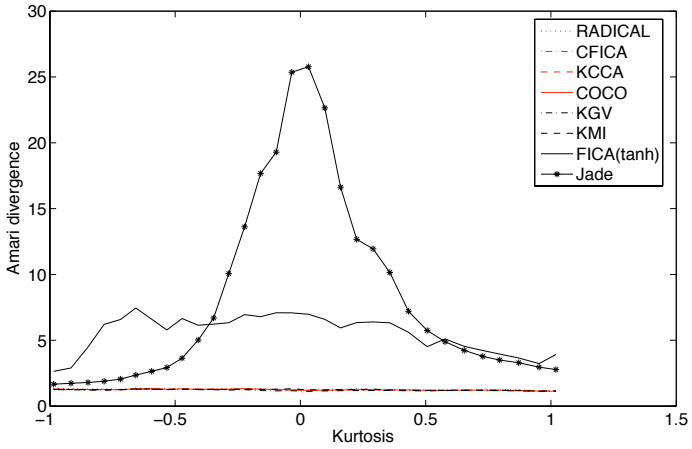

Figure 3: Effect of near-zero kurtosis on the performance of the algorithms, for two signals of length 1000 drawn from a range of mixtures of two Gaussians. Each point represents an average over 100 independent experiments. We used a Gaussian kernel with $\sigma = 1$ and precision $\varepsilon = 2 \times 10^{-5}$ for all kernel dependence functionals, and $\kappa = 2 \times 10^{-2}$ for the KCC and KGV.

no polishing step was applied to the KGV or KMI, since it caused a drop in performance in both cases. [28] Our use of the Laplace kernel in the KMI was motivated by music generally being super-Gaussian (Bell and Sejnowski, 1995). Random permutation of time indices was used to reduce the statistical dependence of adjacent samples in the music, since this was found to improve performance (note that this permutation was carried out on the mixed signals, and was the same for each of the observed mixtures). It is notable that RADICAL, which performs best in the case of noise-free artificial data, does not improve on standard methods in the case of musical sources.

Although the results in Table 5 are quite similar for the KGV and KMI, it is instructive to compare the distribution of the outcomes obtained in each experiment. Generally, the KGV results are more tightly grouped about their mean, whereas the KMI yields more results at smaller Amari divergences, but a larger number of outliers with greater error.

6. Conclusions and outlook

6.1 Conclusions

We have introduced two novel functionals to measure independence: the constrained covariance (COCO), which is the spectral norm of the covariance operator between reproducing kernel Hilbert spaces, and the kernel mutual information (KMI), which is a function of the entire spectrum of the empirical estimate of this covariance operator. The first quantity is analogous to the kernel canonical correlation (KCC), which is the spectral norm of the correlation operator; the second is analogous to the kernel generalised variance (KGV), which is a function of the empirical correlation operator spectrum (see Table 1 in the introduction). We prove two main results. First, we describe the class of *all* reproducing kernel Hilbert spaces for which these four functionals determine independence: the RKHSs must be universal. Second, we link the KMI and the KGV with the mutual information, proving the KMI is an upper bound near independence on the Parzen window estimate of the mutual information, and the KGV is a looser upper bound under certain conditions.

28. This is perhaps surprising, given that the polishing step caused a minor increase in performance in the $n = 2$ case. On the other hand, the larger dimension of the $n = 4$ problem makes the global minimum harder to find, and diversion to local minima more likely.

We emphasise that the KMI and KGV do not require the space partitioning or binning approximations usually associated with estimates of the mutual information (Paninski, 2003).

Our experiments demonstrate the effectiveness of kernel algorithms in ICA, as compared with both standard methods (Jade, Fast ICA, and Extended Infomax); and modern approaches (CFICA, RADICAL). We emphasise that kernel methods (the KMI and KGV in particular) are clearly superior to the alternatives when outlier noise is present in the observations, and are also best at unmixing real (musical) signals. In addition, all modern methods are unaffected by the sources having zero kurtosis, which is not true of earlier algorithms.

Our experiments also point to the superiority of the KMI and KGV over the KCC and COCO in measuring independence. Since independence of two random variables implies that the entire spectrum of the associated covariance (or correlation) operator is zero, it comes as no surprise that measures using the whole spectrum are more robust than those using only the largest singular value. This intuition remains to be formalised, however.

The choice between the KGV and KMI (or, alternatively, COCO and the KCC) is more complicated. The methods proposed by Bach and Jordan (2002a) appear to do well when there is little data available, as in the $n = 2$, $m = 250$ and $n = 4$,$m = 1000$ cases in Table 4, although the mechanism by which this is achieved remains unclear. On the other hand, the KCC and KGV do less well when the sample size/number of sources are large. The KGV and KCC can also be more susceptible to noise in the observations, which is particularly apparent when κ becomes small[29] (and the bound on mutual information provided by the KGV is looser). Indeed, in our outlier resistance experiments, the KMI and COCO achieve by default the optimal performance of the KCC and KGV with model selection over κ. The absence of a separate regularisation parameter in our kernel functionals therefore greatly simplifies model selection, especially if the observations are known to be corrupted by outliers.

6.2 Directions for future study

A number of extensions to this work are readily apparent. For instance, the behaviour of the KMI has not been studied in detail for more than two univariate random variables, besides the discussion in Section 3.2 which guarantees it to be zero when the empirical COCO is zero. In particular, it would be of interest to prove that (30) in Section 3.2 is an upper bound on the Gaussian mutual information, in the manner described in Section 3.1.5 for two random variables. This would incidentally require the link between the Gaussian mutual information and the discrete mutual information, described in Section 3.1 for the two variable case, to be extended to a greater number of random variables. The optimisation procedure we use for ICA might also be made faster, for instance by implementing Newton's method or conjugate gradient descent on the Stiefel manifold (as described by Edelman et al. (1998)), rather than simple gradient descent.

We also need to ensure that both the KMI and COCO approach their population expressions as the sample size increases. In the case of COCO, Gretton et al. (2005b, 2004) give probabilistic bounds for deviations from the expected value using standard tools from uniform convergence theory. The application of these results to the empirical KMI is less clear, however, since the KMI is a *product* of multiple COCO-type quantities, and we do not know what expression it approaches in the population limit. More generally, it is necessary to further investigate methods for model selection (i.e., for choosing the kernel size and type) in COCO and the KMI. It is not presently known whether performance is most effectively tuned by simple cross-validation, using bounds derived from concentration inequalities, or via the properties of Parzen window estimates described by Silverman (1986).

Many real life problems do not fit neatly into the linear ICA framework: we now outline ways in which our kernel dependence functionals might be used to improve performance in these more difficult signal separation problems. First, let us consider the separation of random processes, as opposed to random variables. It is rare in practice to encounter signals that do not depend on their previous outputs. Rather, most real signals exhibit statistical dependencies between the observations at different times (this is obviously true of music, for example). These random processes may be stationary, meaning that their statistical properties (for instance the mean and correlation) do not change over time; or they may be nonstationary. In both cases, however, the

29. κ is the regularisation scaling factor for these dependence functionals.

time dependence greatly assists in separating signals into independent components, the idea being that the independence of different random processes should hold not only between samples drawn at the same time, but also between samples drawn at *different* times. Approaches to this problem include that of Belouchrani et al. (1997), who separate the signals using decorrelation between the sources at any time shift, and the more general approach of Belouchrani and Amin (1998), who use Cohen's class time-frequency kernels to transform the signal and facilitate source separation. The former approach is limited since it breaks down when the sources have overlapping spectra, due to its using only a second order dependence measure. Thus, it would be interesting to generalise the approach of Belouchrani et al. (1997) using kernel measures of dependence, rather than correlation. This generalisation has been investigated, using the mutual information as a dependence measure, by Stögbauer et al. (2004).

Another generalisation of ICA is the separation of sources when mixing is nonlinear. This is considerably more difficult than linear ICA, due to the increased complexity of the mixing model. One simplification, which makes the problem more tractable, is the *post-nonlinear* model: the ith component of the observation vector $\mathbf{t}$ is

$$t_i = f_i(\mathbf{b}_i \mathbf{s}), \tag{39}$$

where f_i is the ith (unknown) nonlinearity, and $\mathbf{b}_i$ is the ith row of the mixing matrix $\mathbf{B}$. This situation corresponds for instance to the observations being distorted by the sensors. Approaches to this problem include the methods of Taleb and Jutten (1999); Achard et al. (2001, 2003)—a comparison of these techniques with COCO and the KMI would therefore be of interest (this would require an efficient optimisation algorithm for our dependence measures under the setting (39)).

Various efforts have also been made to solve the more general case

$$\mathbf{t} = f(\mathbf{s}).$$

This problem requires additional constraints on f, to avoid a trivial solution via the Darmois decomposition (Hyvärinen and Pajunen, 1999) (even then, it is generally the case that each source s_i can only be recovered up to a nonlinear distortion; this is the analogue of the scaling indeterminacy (Theorem 24) in the linear mixing case). It may also be necessary for the observations to arise from random processes, rather than being i.i.d. For instance, according to Hosseni and Jutten (2003), enforcing temporal decorrelation over a single time step is sufficient to test whether the recovered independent processes are simply the result of a Darmois decomposition. While this does not rule out other transforms that return independent signals unrelated to the sources, it suggests that time dependencies have a crucial role to play in general nonlinear mixing. In the scheme suggested by Harmeling et al. (2003), demixing is achieved by mapping the observations to a reproducing kernel Hilbert space, finding a low dimensional basis in the feature space which approximately spans the subspace formed by the observations, and enforcing the second order temporal decorrelation of projections onto this basis. The applicability of the KMI is less clear than in the case of post-nonlinear mixtures, although this might follow from a better understanding of the technique of Harmeling et al. (2003) and its relation to our work.

Finally, Bach and Jordan (2002b) propose using kernel dependence measures in representing probability distributions as tree structured graphical models. Fitting these models requires in particular that the mutual information between pairs of random variables be maximised: thus, Bach and Jordan compare the KGV to a Parzen window estimate of the mutual information in this context. Although the Parzen window approach generally performs better, the KGV is also very effective. We have shown, however, that the KGV is an upper bound (near independence) on the mutual information: thus the KGV performance is a possible indication of the tightness of this upper bound. Given that the KMI is in theory a tighter upper bound than the KGV, it would be interesting to compare its performance with the KGV in this setting.

Acknowledgments

The authors would like to thank to Jean-Yves Audibert and Matthias Seeger, who both discovered errors in our original reasoning for the KMI proof; Francis Bach and Michael Jordan, for providing the kernel ICA

code on the web, and for helpful comments; and Aiyou Chen and Erik Miller, for their assistance in our experimental comparison of the ICA algorithms. 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.

Appendix A. Proofs

A.1 COCO, kernel PCA, and kernel target alignment

In this appendix, we show that COCO is the quantity optimised when obtaining the first principal component in the kernel principal component analysis (kPCA) method of Schölkopf et al. (1998). This can be seen as follows: kPCA satisfies the eigenvalue problem

$$\max_{\|\mathbf{y}\|\leq 1} \mathbf{y}^\top \mathbf{K}\mathbf{y} = \lambda$$

(an inequality is used to keep the constraint set convex). This is rewritten

$$\begin{aligned}\max_{\|\mathbf{y}\|\leq 1} \mathbf{y}^\top \mathbf{K}\mathbf{y} &= \max_{\|\mathbf{y}\|\leq 1} \operatorname{tr}\left(\mathbf{K}\mathbf{y}\mathbf{y}^\top\right) \\ &= \max_{\|\mathbf{y}\|\leq 1} \left\|\mathbf{K}\mathbf{y}\mathbf{y}^\top\right\|_2,\end{aligned}$$

where the norm in the final line is the largest singular value. The final expression is just $\mathrm{COCO}^2_{\mathrm{emp}}$, with feature space $G := \mathbb{R}$ and inner product[30] $l(y_i, y_j) = y_i y_j$. The difference with respect to the dependence measurement framework described previously is that we now maximise over the members y_i of G, rather than being given them in advance. This last argument also shows that COCO is optimised in the spectral clustering/kernel target alignment framework of Cristianini et al. (2002).

30. Note that the linear kernel used here is *not* universal, and thus COCO is not a general dependence functional in this context: see Section 2.3.

A.2 Ratio of determinants

In this appendix, we prove Theorem 11. First, we note that both $\mathbf{A}$ and $\mathbf{C}$ must be positive definite, since they are submatrices of the positive definite matrix (8). Then

$$\begin{aligned}
\frac{\left|\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^\top & \mathbf{C}\end{bmatrix}\right|}{|\mathbf{A}|\,|\mathbf{C}|} &= \frac{\left|\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^\top & \mathbf{C}\end{bmatrix}\right|}{\left|\begin{bmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \mathbf{C}\end{bmatrix}\right|} \\
&\overset{(a)}{=} \frac{\left|\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^\top & \mathbf{C}\end{bmatrix}\right|}{\left|\begin{bmatrix}\mathbf{A}^{1/2} & \mathbf{0}\\ \mathbf{0} & \mathbf{C}^{1/2}\end{bmatrix}\begin{bmatrix}\mathbf{A}^{1/2} & \mathbf{0}\\ \mathbf{0} & \mathbf{C}^{1/2}\end{bmatrix}\right|} \\
&= \left|\begin{bmatrix}\mathbf{A}^{-1/2} & \mathbf{0}\\ \mathbf{0} & \mathbf{C}^{-1/2}\end{bmatrix}\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^\top & \mathbf{C}\end{bmatrix}\begin{bmatrix}\mathbf{A}^{-1/2} & \mathbf{0}\\ \mathbf{0} & \mathbf{C}^{-1/2}\end{bmatrix}\right| \\
&= \left|\begin{bmatrix}\mathbf{I} & \mathbf{A}^{-1/2}\mathbf{B}\mathbf{C}^{-1/2}\\ \mathbf{C}^{-1/2}\mathbf{B}^\top\mathbf{A}^{-1/2} & \mathbf{I}\end{bmatrix}\right|. \\
&\overset{(b)}{=} \left|\mathbf{I}-\mathbf{A}^{-1/2}\mathbf{B}\mathbf{C}^{-1}\mathbf{B}^\top\mathbf{A}^{-1/2}\right| \\
&= \left|\mathbf{I}-\mathbf{C}^{-1/2}\mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B}\mathbf{C}^{-1/2}\right| \\
&\overset{(c)}{=} \prod_i (1-\rho_i^2)
\end{aligned}$$

where (a) requires that $\mathbf{A}$ and $\mathbf{C}$ be positive definite,[31] (b) uses the relation between the determinant of a matrix and that of its Schur complement from Horn and Johnson (1985, p. 22), and (c) uses Theorem 7.3.7 of Horn and Johnson (1985) to determine that ρ_i are the singular values of $\mathbf{A}^{-1/2}\mathbf{B}\mathbf{C}^{-1/2}$. Note that since (8) has only positive eigenvalues, and the determinant of a symmetric matrix is the product of the eigenvalues, we are guaranteed

$$\prod_i (1-\rho_i^2) > 0.$$

From Horn and Johnson (1985, Theorem 7.3.7), we can write ρ_i as the positive solutions of the eigenvalue problem

$$\begin{bmatrix}\mathbf{0} & \mathbf{A}^{-1/2}\mathbf{B}\mathbf{C}^{-1/2}\\ \mathbf{C}^{-1/2}\mathbf{B}^\top\mathbf{A}^{-1/2} & \mathbf{0}\end{bmatrix}\mathbf{b}_i = \rho_i \mathbf{b}_i,$$

bearing in mind that these solutions come in pairs with equal magnitude and opposite sign. Rearranging and making an appropriate change of variables yields the generalised eigenvalue problem

$$\begin{bmatrix}\mathbf{0} & \mathbf{B}\\ \mathbf{B}^\top & \mathbf{0}\end{bmatrix}\mathbf{a}_i = \rho_i \begin{bmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \mathbf{C}\end{bmatrix}\mathbf{a}_i.$$

A.3 Determinant form of the Gaussian mutual information

In this section, we give a derivation of (18) in Section 3.1.3, which states that

$$I(\mathbf{x}_G;\mathbf{y}_G) = -\frac{1}{2}\log\left(\left|\mathbf{I}_{l_y} - \left(\mathbf{P}_{xy}-\mathbf{p}_x\mathbf{p}_y^\top\right)^\top \mathbf{D}_x^{-1}\left(\mathbf{P}_{xy}-\mathbf{p}_x\mathbf{p}_y^\top\right)\mathbf{D}_y^{-1}\right|\right). \tag{40}$$

31. A matrix has a square root if and only if it is positive definite.

This result was given without proof by Bach and Jordan (2002a, Appendix B). We begin with the mutual information between $\mathbf{x}_G$ and $\mathbf{y}_G$, which is written

$$I(\mathbf{x}_G;\mathbf{y}_G) = -\frac{1}{2}\log\left(\prod_i (1-\rho_i^2)\right), \tag{41}$$

where ρ_i are the positive solutions to the generalised eigenvalue problem

$$\begin{bmatrix} \mathbf{0} & \mathbf{P}_{xy}-\mathbf{p}_x\mathbf{p}_y^\top \\ \left(\mathbf{P}_{xy}-\mathbf{p}_x\mathbf{p}_y^\top\right)^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} = \rho_i \begin{bmatrix} \mathbf{D}_x-\mathbf{p}_x\mathbf{p}_x^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_y-\mathbf{p}_y\mathbf{p}_y^\top \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} \tag{42}$$

(this can be found by substituting the covariances (15)-(17) into (10)). Note that both $\mathbf{D}_x-\mathbf{p}_x\mathbf{p}_x^\top$ and $\mathbf{D}_y-\mathbf{p}_y\mathbf{p}_y^\top$ have rank l_x-1 and l_y-1 respectively, and are not invertible.[32] To see this, we make the expansions

$$\begin{aligned} \mathbf{D}_x-\mathbf{p}_x\mathbf{p}_x^\top &= \mathbf{D}_x\left(\mathbf{I}_{l_x}-\mathbf{1}_{l_x}\mathbf{p}_x^\top\right) = \mathbf{D}_x\mathbf{E}_x, \\ \mathbf{D}_y-\mathbf{p}_y\mathbf{p}_y^\top &= \mathbf{D}_y\left(\mathbf{I}_{l_y}-\mathbf{1}_{l_y}\mathbf{p}_y^\top\right) = \mathbf{D}_y\mathbf{E}_y, \end{aligned}$$

where $\mathbf{E}_x := \mathbf{I}_{l_x}-\mathbf{1}_{l_x}\mathbf{p}_x^\top$ and $\mathbf{E}_y := \mathbf{I}_{l_y}-\mathbf{1}_{l_y}\mathbf{p}_y^\top$ have zero eigenvalues corresponding to the eigenvectors $\frac{1}{\sqrt{l_x}}\mathbf{1}_{l_x}$ and $\frac{1}{\sqrt{l_y}}\mathbf{1}_{l_y}$, respectively. In addition, we note that

$$\begin{aligned} \left(\mathbf{P}_{xy}-\mathbf{p}_x\mathbf{p}_y^\top\right)\mathbf{E}_y &= \left(\mathbf{P}_{xy}-\mathbf{p}_x\mathbf{p}_y^\top\right)\left(\mathbf{I}_{l_y}-\mathbf{1}_{l_y}\mathbf{p}_y^\top\right) \\ &= \mathbf{P}_{xy}-\mathbf{p}_x\mathbf{p}_y^\top-\mathbf{P}_{xy}\mathbf{1}_{l_y}\mathbf{p}_y^\top+\mathbf{p}_x\mathbf{p}_y^\top\mathbf{1}_{l_y}\mathbf{p}_y^\top \\ &= \mathbf{P}_{xy}-\mathbf{p}_x\mathbf{p}_y^\top-\mathbf{p}_x\mathbf{p}_y^\top+\mathbf{p}_x\mathbf{p}_y^\top \\ &= \mathbf{P}_{xy}-\mathbf{p}_x\mathbf{p}_y^\top, \end{aligned}$$

with an analogous result for $\left(\mathbf{P}_{xy}-\mathbf{p}_x\mathbf{p}_y^\top\right)^\top\mathbf{E}_x$. We may therefore write (42) as

$$\begin{bmatrix} \mathbf{0} & \left(\mathbf{P}_{xy}-\mathbf{p}_x\mathbf{p}_y^\top\right)\mathbf{E}_y \\ \left(\mathbf{P}_{xy}-\mathbf{p}_x\mathbf{p}_y^\top\right)^\top\mathbf{E}_x & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} = \rho_i \begin{bmatrix} \mathbf{D}_x\mathbf{E}_x & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_y\mathbf{E}_y \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix},$$

from which we obtain a generalised eigenvalue problem with identical eigenvalues ρ_i,

$$\begin{bmatrix} \mathbf{0} & \mathbf{P}_{xy}-\mathbf{p}_x\mathbf{p}_y^\top \\ \left(\mathbf{P}_{xy}-\mathbf{p}_x\mathbf{p}_y^\top\right)^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{e}_i \\ \mathbf{f}_i \end{bmatrix} = \rho_i \begin{bmatrix} \mathbf{D}_x & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_y \end{bmatrix} \begin{bmatrix} \mathbf{e}_i \\ \mathbf{f}_i \end{bmatrix}.$$

Since $\mathbf{D}_x$ and $\mathbf{D}_y$ have full rank, we may now apply Theorem 11 to obtain (40).

A.4 Details of Definition 13

In this section, we derive the Parzen window estimate of the Gaussian mutual information provided in Definition 13. The kernel density (Parzen window) estimates for $\mathbf{p}_{\mathsf{x},\mathsf{y}}$ and its marginals, on the basis of the sample z, are

$$\begin{aligned} \widehat{\mathbf{p}}_{\mathsf{x}}(x) &= \frac{1}{m}\sum_{l=1}^{m}\kappa(x_l-x), \qquad \widehat{\mathbf{p}}_{\mathsf{y}}(y) = \frac{1}{m}\sum_{l=1}^{m}\kappa(y_l-y), \\ \widehat{\mathbf{p}}_{\mathsf{x},\mathsf{y}}(x,y) &= \frac{1}{m}\sum_{l=1}^{m}\kappa(x_l-x)\,\kappa(y_l-y), \end{aligned}$$

32. This is why we use (41) as our expression for the mutual information, rather than the ratio of determinants (7) (which would be undefined here).

where the kernel argument indicates which kernel is used, to simplify notation. We require approximations to the terms in the Gaussian mutual information, as described in (18). We therefore define the vectors $\hat{\mathbf{p}}_x, \hat{\mathbf{p}}_y$, and the matrix $\widehat{\mathbf{P}}_{xy}$, using the expectations in (12)-(14) computed with these kernel expressions;

$$\widehat{\mathbf{E}}_{\mathsf{x},\mathsf{y}}\left(\check{\mathbf{x}}\check{\mathbf{y}}^\top\right) = \widehat{\mathbf{P}}_{xy}, \tag{43}$$

$$\widehat{\mathbf{E}}_{\mathsf{x}}\left(\check{\mathbf{x}}\right) = \hat{\mathbf{p}}_x, \tag{44}$$

$$\widehat{\mathbf{E}}_{\mathsf{x}}\left(\check{\mathbf{x}}\check{\mathbf{x}}^\top\right) = \widehat{\mathbf{D}}_x. \tag{45}$$

In the limit where Δ_x, Δ_y are small (and thus, by implication, $l_x \gg m$, $l_y \gg m$, $\sigma_x \gg \Delta_x$, and $\sigma_y \gg \Delta_y$, where σ_x and σ_y define the kernel sizes), we make the approximations

$$\widehat{\mathbf{E}}_{\mathsf{x}}\left((\check{\mathbf{x}})_i\right) = \widehat{\mathbf{P}}_{\hat{\mathsf{x}}}(i) = \frac{1}{m}\int_{q_i}^{q_i+\Delta_x}\sum_{l=1}^{m}\kappa(x_l - x)\,dx \approx \frac{\Delta_x}{m}\sum_{l=1}^{m}\kappa(x_l - q_i),$$

$$\widehat{\mathbf{E}}_{\mathsf{x}}\left(\left(\check{\mathbf{x}}\check{\mathbf{x}}^\top\right)_{i,j}\right) \approx \begin{cases} \frac{\Delta_x}{m}\sum_{l=1}^{m}\kappa(x_l - q_i) & i = j \\ 0 & \text{otherwise} \end{cases},$$

and

$$\begin{aligned}\widehat{\mathbf{E}}_{\mathsf{x},\mathsf{y}}\left(\left(\check{\mathbf{x}}\check{\mathbf{y}}^\top\right)_{i,j}\right) = \widehat{\mathbf{P}}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j) &= \frac{1}{m}\int_{q_i}^{q_i+\Delta_x}\int_{r_j}^{r_j+\Delta_y}\sum_{l=1}^{m}\kappa(x_l - x)\,\kappa(y_l - y)\,dxdy \\ &\approx \frac{\Delta_x\Delta_y}{m}\sum_{l=1}^{m}\kappa(x_l - q_i)\,\kappa(y_l - r_j).\end{aligned}$$

Before proceeding further, we define two matrices of kernel inner products to simplify our notation. Namely,

$$\mathbf{K}_l := \begin{bmatrix} \kappa(q_1 - x_1) & \dots & \kappa(q_1 - x_m) \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ \kappa(q_{l_x} - x_1) & \dots & \kappa(q_{l_x} - x_m) \end{bmatrix}, \quad \mathbf{L}_l := \begin{bmatrix} \kappa(r_1 - y_1) & \dots & \kappa(r_1 - y_m) \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ \kappa\left(r_{l_y} - y_1\right) & \dots & \kappa\left(r_{l_y} - y_m\right) \end{bmatrix}, \tag{46}$$

where we write the above in such a manner as to indicate $l_x \gg m$ and $l_y \gg m$. We now use the above results to re-write (43)-(45) as respectively

$$\widehat{\mathbf{P}}_{xy} - \hat{\mathbf{p}}_x\hat{\mathbf{p}}_y^\top \approx \frac{\Delta_x\Delta_y}{m}\left(\mathbf{K}_l\mathbf{L}_l^\top - \frac{1}{m}\mathbf{K}_l\mathbf{1}_m\mathbf{1}_m^\top\mathbf{L}_l^\top\right) = \frac{\Delta_x\Delta_y}{m}\mathbf{K}_l\mathbf{H}\mathbf{L}_l^\top,$$

$$\widehat{\mathbf{D}}_x \approx \frac{\Delta_x}{m}\mathrm{diag}\left(\mathbf{K}_l\mathbf{1}_m\right) =: \frac{\Delta_x^2}{m}\mathbf{D}_l^{(x)},$$

and

$$\widehat{\mathbf{D}}_y \approx \frac{\Delta_y}{m}\mathrm{diag}\left(\mathbf{L}_l\mathbf{1}_m\right) =: \frac{\Delta_y^2}{m}\mathbf{D}_l^{(y)},$$

where we introduce the terms

$$\mathbf{D}_l^{(x)} = \frac{1}{\Delta_x}\begin{bmatrix} \sum_{l=1}^{m}\kappa(q_1 - x_l) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sum_{l=1}^{m}\kappa(q_{l_x} - x_l) \end{bmatrix} \tag{47}$$

and

$$\mathbf{D}_l^{(y)} = \frac{1}{\Delta_y}\begin{bmatrix} \sum_{l=1}^{m}\kappa(r_1 - y_l) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sum_{l=1}^{m}\kappa(r_{l_y} - y_l) \end{bmatrix}. \tag{48}$$

With these substitutions, we can rewrite

$$\left(\widehat{\mathbf{D}}_x\right)^{-1/2}\left(\widehat{\mathbf{P}}_{xy}-\hat{\mathbf{p}}_x\hat{\mathbf{p}}_y^\top\right)\left(\widehat{\mathbf{D}}_y\right)^{-1/2}\approx\left(\mathbf{D}_l^{(x)}\right)^{-1/2}\left(\mathbf{K}_l\mathbf{H}\left(\mathbf{L}_l\right)^\top\right)\left(\mathbf{D}_l^{(y)}\right)^{-1/2},$$

which results in our definition.

A.5 Proof of Theorem 16

Our proof of Theorem 16 requires the following lemma.

Lemma 27 (Singular values of a matrix product) *Let* $\mathbf{A}$, $\mathbf{B}$ *be* $m\times n$ *matrices,* $q:=\min(m,n)$*, and* $\mathbf{A}$ *have singular values* $\sigma_1(\mathbf{A}),\ldots,\sigma_q(\mathbf{A})$ *(ordered from largest to smallest). Then* $\sigma_1(\mathbf{A}\mathbf{B}^\top)\leq\sigma_1(\mathbf{A})\sigma_1(\mathbf{B})$ *and*

$$\sigma_q(\mathbf{A}\mathbf{B}^\top)\quad\leq\quad\min\left\{\sigma_q(\mathbf{A})\sigma_1(\mathbf{B}),\sigma_1(\mathbf{A})\sigma_q(\mathbf{B})\right\}.$$

.

This is a special case of a result of Horn and Johnson (1985, p. 423). We now proceed with the proof. The principle we will follow is straightforward: we want to upper bound the Gaussian mutual information in (20) by upper bounding *each* of the $\hat{\rho}_i$ that define it. Indeed, if we can find a matrix to replace (21) with singular values $\alpha_i\geq\hat{\rho}_i$ for all i, it follows that $-\frac{1}{2}\log\left(\prod_i\left(1-\alpha_i^2\right)\right)\geq-\frac{1}{2}\log\left(\prod_i\left(1-\hat{\rho}_i^2\right)\right)$. First, we note that $\pm\hat{\rho}_i$ are the eigenvalues of the matrix

$$\underbrace{\begin{bmatrix}\left(\mathbf{D}_l^{(x)}\right)^{-1}&\mathbf{0}\\\mathbf{0}&\left(\mathbf{D}_l^{(y)}\right)^{-1}\end{bmatrix}}_{\mathbf{D}^{-1}}\underbrace{\begin{bmatrix}\mathbf{0}&\mathbf{K}_l\mathbf{H}\left(\mathbf{L}_l\right)^\top\\\mathbf{L}_l\mathbf{H}\left(\mathbf{K}_l\right)^\top&\mathbf{0}\end{bmatrix}}_{\mathbf{E}}.$$

According to (22), $\mathbf{D}_l^{(x)}$ is a diagonal matrix with jth entry $\frac{1}{\Delta_x}\sum_{i=1}^m\kappa(x_i-q_j)$, which is an unnormalised Parzen window estimate of $\mathbf{p}_\times$ at grid point q_j (an analogous result holds for $\mathbf{D}_l^{(y)}$). It follows that $\mathbf{D}$ is diagonal, and we denote its ith largest value as d_i (*i.e.*, d_1 is the overall maximum); we also define σ_i to be the ith singular value of $\mathbf{E}$. We may obtain a new matrix with singular values $\alpha_i\geq\hat{\rho}_i$ by replacing the diagonal entries of $\mathbf{D}$ with their smallest value,[33]

$$\begin{aligned}\mathbf{D}\quad&\rightarrow\quad\min_i(d_i)\mathbf{I}\\&=\quad\frac{\nu_z}{\Delta}\mathbf{I},\end{aligned}\tag{49}$$

where $\nu_z=\min\{\nu_x,\nu_y\}$ and

$$\nu_x:=\min_{j\in\{1\ldots l_x\}}\sum_{i=1}^m\kappa(x_i-q_j),\qquad\nu_y:=\min_{j\in\{1\ldots l_y\}}\sum_{i=1}^m\kappa(y_i-r_j).\tag{50}$$

The singular values α_i of $(\frac{\nu}{\Delta}\mathbf{I})^{-1}\mathbf{E}$ satisfy[34]

$$\begin{aligned}\hat{\rho}_i\quad&\leq\quad\min\left\{d_{l_x+l_y}^{-1}\sigma_i,d_{l_x+l_y-i+1}^{-1}\sigma_1\right\}\\&\leq\quad d_{l_x+l_y}^{-1}\sigma_i\\&=\quad\frac{\Delta}{\nu_z}\sigma_i=\alpha_i\end{aligned}$$

33. We assume without loss of generality that $\Delta_x=\Delta_y=\Delta$, since this simplifies notation.
34. Bear in mind that due to the ordering of the singular values, $\max_j d_j^{-1}=d_{l_x+l_y}^{-1}$; and the d_j^{-1} are sorted in reverse order to the d_j.

for all i, where the first inequality derives from Lemma 27. Rather than computing the minima in (50) over the grid, however, we may simply use

$$\nu_{\boldsymbol{x}} := \min_{j \in \{1 \ldots m\}} \sum_{i=1}^{m} \kappa(x_i - x_j), \qquad \nu_{\mathbf{y}} := \min_{j \in \{1 \ldots m\}} \sum_{i=1}^{m} \kappa(y_i - y_j),$$

which are respectively the smallest (unnormalised) Parzen window estimates of $\mathbf{p}_{\mathsf{x}}$ and $\mathbf{p}_{\mathsf{y}}$ at any *sample point*: these approach the smallest values of $\mathbf{p}_{\mathsf{x}}$ on X, and of $\mathbf{p}_{\mathsf{y}}$ on Y, as the sample size increases (the densities are bounded away from zero by assumption).

Having made the replacement in (49), it is straightforward to take a limit in which the grid becomes infinitely fine. We begin by rearranging the Lemma 13 definition as

$$\begin{aligned} \widehat{I}(\hat{\mathsf{x}};\hat{\mathsf{y}}) &\leq -\frac{1}{2}\log\left|\mathbf{I} - \left(\frac{\Delta}{\nu_{\boldsymbol{z}}}\right)^2 \left(\mathbf{K}_l \mathbf{H} (\mathbf{L}_l)^\top\right) \left(\mathbf{K}_l \mathbf{H} (\mathbf{L}_l)^\top\right)^\top\right| \\ &= -\frac{1}{2}\log\left|\mathbf{I} - \left(\frac{\Delta}{\nu_{\boldsymbol{z}}}\right)^2 \left(\mathbf{H}\mathbf{K}_l^\top \mathbf{K}_l \mathbf{H}\right) \left(\mathbf{H}\mathbf{L}_l^\top \mathbf{L}_l \mathbf{H}\right)\right|. \end{aligned}$$

We then have the limiting result

$$\begin{aligned} \lim_{l_x \to \infty} \left(\frac{\Delta_x}{\nu_{\boldsymbol{z}}} \mathbf{K}_l^\top \mathbf{K}_l\right)_{i.j} &= \nu_{\boldsymbol{z}}^{-1} \lim_{l_x \to \infty} \Delta_x \sum_{p=1}^{l_x} \kappa(x_i - q_p)\kappa(x_j - q_p) \\ &= \nu_{\boldsymbol{z}}^{-1} \int_X \kappa(x_i - q)\kappa(x_j - q) dq \\ &= \nu_{\boldsymbol{z}}^{-1} k(x_i, x_j), \end{aligned}$$

where we recover our RKHS kernel as the convolution of the kernel density functions at each pair of data points.

A.6 Proof of Theorem 18

In this section, we prove that the KGV upper bounds the KMI when conditions (28) hold. We recall the definition of the *unregularised* KGV,[35] which occurs at $\theta = 1$. It follows from Lemma 9 that

$$\mathrm{KGV}(\boldsymbol{z}; F, G, 1) = \infty,$$

since the associated eigenvalues ρ_i in (27) are all either 1, -1, or 0 (given we use universal kernels, there will be at least one pair of non-zero eigenvalues). Conversely, when $\theta = 0$, we recover the KMI. It remains to show that increasing θ from 0 to 1 causes the KGV to increase monotonically.

We may rearrange the eigenvalue problem in (27) as

$$\begin{bmatrix} \mathbf{I} & \left(\theta\widetilde{\mathbf{K}} + (1-\theta)\nu I\right)^{-1} \widetilde{\mathbf{L}} \\ \left(\theta\widetilde{\mathbf{L}} + (1-\theta)\nu I\right)^{-1} \widetilde{\mathbf{K}} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} = (1+\rho_i) \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix}.$$

Then

$$\begin{aligned} \mathrm{KGV}(\boldsymbol{z}; F, G, \theta) &= -\log\left|\begin{bmatrix} \mathbf{I} & \left(\theta\widetilde{\mathbf{K}} + (1-\theta)\nu_{\boldsymbol{z}}\mathbf{I}\right)^{-1} \widetilde{\mathbf{L}} \\ \left(\theta\widetilde{\mathbf{L}} + (1-\theta)\nu_{\boldsymbol{z}}\mathbf{I}\right)^{-1} \widetilde{\mathbf{K}} & \mathbf{I} \end{bmatrix}\right| \\ &= -\log\left|\mathbf{I} - \left(\theta\widetilde{\mathbf{K}} + (1-\theta)\nu_{\boldsymbol{z}}\mathbf{I}\right)^{-1} \widetilde{\mathbf{L}} \left(\theta\widetilde{\mathbf{L}} + (1-\theta)\nu_{\boldsymbol{z}}\mathbf{I}\right)^{-1} \widetilde{\mathbf{K}}\right|. \end{aligned}$$

35. We emphasise that only the regularised KGV is used in practice.

We now use the result that if $\mathbf{A}' \succ \mathbf{A} \succ \mathbf{0}$ and $\mathbf{B}' \succ \mathbf{B} \succ \mathbf{0}$, then $\mathbf{A}'\mathbf{B}' \succ \mathbf{A}\mathbf{B}$ (this is a straightforward corollary to Theorem 7.7.3 of Horn and Johnson, 1985). The desired result then holds as long as

$$\theta'\widetilde{\mathbf{K}} + (1-\theta')\nu_z\mathbf{I} \prec \theta\widetilde{\mathbf{K}} + (1-\theta)\nu_z\mathbf{I}$$

when $\theta' > \theta$ (as well as the analogous result for $\theta\widetilde{\mathbf{L}} + (1-\theta)\nu_z\mathbf{I}$), which means

$$(\theta-\theta')\widetilde{\mathbf{K}} + (\theta'-\theta)\nu_z\mathbf{I} \succ 0 \qquad \text{and} \qquad (\theta-\theta')\widetilde{\mathbf{L}} + (\theta'-\theta)\nu_z\mathbf{I} \succ 0,$$

or

$$\nu_z\mathbf{I} - \widetilde{\mathbf{K}} \succ 0 \qquad \text{and} \qquad \nu_z\mathbf{I} - \widetilde{\mathbf{L}} \succ 0. \tag{51}$$

A.7 Proof of Lemma 22

In this section, we show that the multivariate KMI is zero if and only if the empirical COCO between each pair of random variables is zero. This may be shown via a minor adaptation of the corresponding proof of Bach and Jordan (2002a, Appendix A.2). First, we may rewrite each factor $\breve{\lambda}_j + 1$ in (30) as the solution to

$$\begin{bmatrix} \mathbf{I} & \nu_z^{-1}\widetilde{\mathbf{K}}_1^{1/2}\widetilde{\mathbf{K}}_2^{1/2} & \dots & \nu_z^{-1}\widetilde{\mathbf{K}}_1^{1/2}\widetilde{\mathbf{K}}_n^{1/2} \\ \nu_z^{-1}\widetilde{\mathbf{K}}_2^{1/2}\widetilde{\mathbf{K}}_1^{1/2} & \mathbf{I} & \dots & \nu_z^{-1}\widetilde{\mathbf{K}}_2^{1/2}\widetilde{\mathbf{K}}_n^{1/2} \\ \vdots & \vdots & \ddots & \vdots \\ \nu_z^{-1}\widetilde{\mathbf{K}}_n^{1/2}\widetilde{\mathbf{K}}_1^{1/2} & \nu_z^{-1}\widetilde{\mathbf{K}}_n^{1/2}\widetilde{\mathbf{K}}_2^{1/2} & \dots & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{d}_{1,j} \\ \mathbf{d}_{2,j} \\ \vdots \\ \mathbf{d}_{n,j} \end{bmatrix} = \left(\breve{\lambda}_j + 1\right) \begin{bmatrix} \mathbf{d}_{1,j} \\ \mathbf{d}_{2,j} \\ \vdots \\ \mathbf{d}_{n,j} \end{bmatrix},$$

where $\widetilde{\mathbf{K}}_i^{1/2}\mathbf{c}_{i,j} = \mathbf{d}_{i,j}$, bearing in mind that the determinant of the left hand matrix is the product of these eigenvalues. Since the left hand matrix is symmetric, the trace is equal to the sum of the eigenvalues, and

$$\sum_{j=1}^{mn}\left(\breve{\lambda}_j + 1\right) = mn. \tag{52}$$

Assuming without loss of generality that the the mnth eigenvalue corresponds to $\breve{\lambda}_{\max} := \lambda_{\max}/\nu_z$, we rewrite (30) as

$$\begin{aligned} -\frac{1}{2}\log\prod_{j=1}^{mn}\left(1+\breve{\lambda}_j\right) &= -\frac{1}{2}\log(1+\breve{\lambda}_{\max}) - \frac{1}{2}\log\prod_{j=1}^{mn-1}\left(1+\breve{\lambda}_j\right) \\ &= -\frac{1}{2}\log(1+\breve{\lambda}_{\max}) - \frac{mn-1}{2}\sum_{j=1}^{mn-1}\frac{1}{mn-1}\log\left(1+\breve{\lambda}_j\right) \\ &\geq -\frac{1}{2}\log(1+\breve{\lambda}_{\max}) - \frac{mn-1}{2}\log\left(\frac{1}{mn-1}\sum_{j=1}^{mn-1}\left(1+\breve{\lambda}_j\right)\right) \\ &= -\frac{1}{2}\log(1+\breve{\lambda}_{\max}) - \frac{mn-1}{2}\log\left(\frac{mn-\breve{\lambda}_{\max}-1}{mn-1}\right), \end{aligned}$$

where the penultimate line uses Jensen's inequality, and we substitute (52) in the final line. The resulting expression is strictly convex with respect to $\breve{\lambda}_{\max}$ (its second derivative is everywhere positive), and has a global minimum at $\breve{\lambda}_{\max} = 0$. It follows that (30) is likewise minimised at $\mathrm{KMI}(\boldsymbol{z}; F_{X_1}, \dots, F_{X_n}) = 0$ (at which point $\breve{\lambda}_j = 0$ for all j), and that this corresponds to the point at which all pairs of empirical constrained covariances are zero, using Definition 19 and Lemma 20.

Appendix B. Discussion of Bach and Jordan's derivation of the KGV

B.1 Computation of the unregularised kernel canonical correlations

In this section, we prove Lemma 9, which is used to show a regularised empirical estimate for the kernel canonical correlates is needed when the associated RKHSs have high dimension. We begin with (5), which we restate below for reference;

$$\begin{bmatrix} \mathbf{0} & \widetilde{\mathbf{K}}\widetilde{\mathbf{L}} \\ \widetilde{\mathbf{L}}\widetilde{\mathbf{K}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} = \rho_i \begin{bmatrix} \left(\widetilde{\mathbf{K}}\right)^2 & \mathbf{0} \\ \mathbf{0} & \left(\widetilde{\mathbf{L}}\right)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix}.$$

This is equivalent to

$$\begin{bmatrix} \mathbf{0} & \left(\widetilde{\mathbf{K}}^-\right)^2 \widetilde{\mathbf{K}}\widetilde{\mathbf{L}} \\ \left(\widetilde{\mathbf{L}}^-\right)^2 \widetilde{\mathbf{L}}\widetilde{\mathbf{K}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} = \rho_i \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix},$$

where we use the pseudoinverses since the Gram matrices do not have full rank. If we recall that $\mathbf{H}$ is the centring matrix, then the solutions ρ_i correspond to the solutions of

$$\begin{aligned} 0 &= \begin{vmatrix} -\rho\mathbf{I} & \left(\widetilde{\mathbf{K}}^-\right)^2 \widetilde{\mathbf{K}}\widetilde{\mathbf{L}} \\ \left(\widetilde{\mathbf{L}}^-\right)^2 \widetilde{\mathbf{L}}\widetilde{\mathbf{K}} & -\rho\mathbf{I} \end{vmatrix} \\ &= |\rho\mathbf{I}| \left| \rho\mathbf{I} - \frac{1}{\rho}\left(\widetilde{\mathbf{L}}^-\right)^2 \widetilde{\mathbf{L}}\widetilde{\mathbf{K}} \left(\widetilde{\mathbf{K}}^-\right)^2 \widetilde{\mathbf{K}}\widetilde{\mathbf{L}} \right| \\ &= |\rho\mathbf{I}| \left| \rho\mathbf{I} - \frac{1}{\rho}\mathbf{H} \right| \\ &= \rho^m \frac{\left(\rho^2 - 1\right)^{m-1}}{\rho^{m-2}}, \end{aligned}$$

which has $m-1$ roots $+1$, $m-1$ roots -1, and 2 roots 0. To avoid this problem, a regularised empirical estimate is used, as shown by Bach and Jordan (2002a); Fukumizu et al. (2005); Leurgans et al. (1993).

B.2 Discussion of the KGV proof of Bach and Jordan (2002a)

In this section, we describe a possible problem in the derivation by Bach and Jordan (2002a, Appendix B) of the kernel generalised variance (KGV). We begin with a quick summary of the steps from Section 3 needed to get us to the point where the proof begins.[36] Assume that X and Y are both bounded intervals on $\mathbb{R}$. In Section 3.1.2, we recall the standard result from Cover and Thomas (1991) that the mutual information $I(\mathsf{x},\mathsf{y})$ between two real-valued, univariate random variables $\mathsf{x} \in X$ and $\mathsf{y} \in Y$ can be approximated by imposing a uniform grid of size $l_x \times l_y$ over $X \times Y$, and defining a multinomial distribution over the discrete valued random variables $\hat{\mathsf{x}} \in \{1,\ldots,l_x\}$ and $\hat{\mathsf{y}} \in \{1,\ldots,l_y\}$ using the probability mass in the resulting bins (this multinomial distribution is described by the matrix $\mathbf{P}_{xy}$ of joint probabilities, with marginal distribution vectors $\mathbf{p}_x$ and $\mathbf{p}_y$).[37] We denote the resulting discrete mutual information as $I(\hat{\mathsf{x}};\hat{\mathsf{y}})$. In Section 3.1.3, we approximate $I(\hat{\mathsf{x}};\hat{\mathsf{y}})$ using the *Gaussian* mutual information $I(\mathbf{x}_G;\mathbf{y}_G)$ between vectors $\mathbf{x}_G;\mathbf{y}_G$, defined to have the same covariance as $\check{\mathbf{x}}$ and $\check{\mathbf{y}}$, where $\hat{\mathsf{x}} = i$ is equivalent to $(\check{\mathbf{x}})_i = 1$ and $(\check{\mathbf{x}})_{j:j\neq i} = 0$ (likewise for $\hat{\mathsf{y}}$). Bach and Jordan (2002a, Appendix B.1) show this approximation holds when the random variables are close to independence,

36. The reader is strongly advised to consult Sections 3.1.1-3.1.3 before proceeding, since the following discussion might not otherwise make much sense.
37. The approximation becomes exact in the limit of an infinitely fine grid.

in which case

$$I(\hat{\mathrm{x}};\hat{\mathrm{y}}) \approx I(\mathbf{x}_G;\mathbf{y}_G) = -\frac{1}{2}\log\left(\prod_i (1-\rho_i^2)\right),$$

where ρ_i are the positive solutions to the generalised eigenvalue problem

$$\begin{bmatrix} \mathbf{0} & \mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top \\ \left(\mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top\right)^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} = \rho_i \begin{bmatrix} \mathbf{D}_x - \mathbf{p}_x\mathbf{p}_x^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_y - \mathbf{p}_y\mathbf{p}_y^\top \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix},$$

and $\mathbf{D}_x = \mathrm{diag}(\mathbf{p}_x)$, $\mathbf{D}_y = \mathrm{diag}(\mathbf{p}_y)$ (see (41) in Appendix A.3).

We are now at the point where we can describe the reasoning of Bach and Jordan (2002a, Appendix B.3) in establishing a link between $I(\hat{\mathrm{x}};\hat{\mathrm{y}})$ and the KGV. Rather than replacing $\hat{\mathrm{x}}$ and $\hat{\mathrm{y}}$ by $\mathbf{x}_G$ and $\mathbf{y}_G$, we may instead replace them with the *smoothed approximations*

$$\mathbf{k}_l = \Delta_x \begin{bmatrix} k(\mathsf{x},q_1) & \cdots & k(\mathsf{x},q_{l_x}) \end{bmatrix}^\top \quad \text{and} \quad \mathbf{l}_l = \Delta_y \begin{bmatrix} l(\mathsf{y},r_1) & \ldots & l(\mathsf{y},r_{l_y}) \end{bmatrix}^\top \tag{53}$$

to $\mathbf{x}_G$ and $\mathbf{y}_G$, respectively, where $k(\cdot,\cdot)$ and $l(\cdot,\cdot)$ are the RKHS kernels for F and G, and the grid coordinates $\boldsymbol{q} := (q_1,\ldots,q_{l_x})$ and $\boldsymbol{r} := (r_1,\ldots,r_{l_y})$ are defined in Section 3.1.2.[38] We can of course specify the Gaussian mutual information $I(\mathbf{k}_l;\mathbf{l}_l)$ between these smoothed vectors, using the appropriate log ratio of determinants. Two questions then arise. First, does this smoothed approximation $I(\mathbf{k}_l;\mathbf{l}_l)$ approach the Gaussian mutual information $I(\mathbf{x}_G;\mathbf{y}_G)$ as the kernel size drops? Second, under what conditions does the empirical estimate of $I(\mathbf{k}_l;\mathbf{l}_l)$ correspond to the KGV? We now describe the approach of Bach and Jordan (2002a) to solving the first question, and postpone discussion of the second question to the end of the section.

The link between the Gaussian approximation to the discrete mutual information and the KGV could be shown by demonstrating

$$\mathbf{P}_{xy} \overset{?}{\approx} \Delta_x\Delta_y\mathbf{E}_{\mathsf{x},\mathsf{y}}\left(\mathbf{k}_l\mathbf{l}_l^\top\right), \quad \mathbf{D}_x \overset{?}{\approx} \Delta_x^2\mathbf{E}_{\mathsf{x}}\left(\mathbf{k}_l\mathbf{k}_l^\top\right), \quad \mathbf{p}_x \overset{?}{\approx} \Delta_x\mathbf{E}_{\mathsf{x}}(\mathbf{k}_l) \tag{54}$$

under appropriate conditions, with similar results for the terms in y. We consider the case where both kernels are Gaussian; that is,

$$\begin{aligned} k(x-q_i) &= \frac{1}{\sqrt{2\pi\sigma_x^2}}\exp\left(-\frac{(x-q_i)^2}{2\sigma_x^2}\right), \\ l(y-r_j) &= \frac{1}{\sqrt{2\pi\sigma_y^2}}\exp\left(-\frac{(y-r_j)^2}{2\sigma_y^2}\right), \end{aligned}$$

bearing in mind that the impulse function is a limiting case (Bracewell, 1986);

$$\delta_{q_i}(x) = \lim_{\sigma_x\to 0}\frac{1}{\sqrt{2\pi\sigma_x^2}}\exp\left(-\frac{(x-q_i)^2}{2\sigma_x^2}\right) := \lim_{\sigma_x\to 0} k(x-q_i). \tag{55}$$

To compute the covariance structure of the vectors in (53), we require expressions for the expectations

$$\mathbf{E}_{\mathsf{x},\mathsf{y}}\left(\mathbf{k}_l\mathbf{l}_l^\top\right), \quad \mathbf{E}_{\mathsf{x}}(\mathbf{k}_l), \quad \mathbf{E}_{\mathsf{x}}\left(\mathbf{k}_l\mathbf{k}_l^\top\right), \quad \mathbf{E}_{\mathsf{y}}\left(\mathbf{l}_l\mathbf{l}_l^\top\right), \quad \mathbf{E}_{\mathsf{y}}(\mathbf{l}_l).$$

The expectation of individual entries in the matrix $\mathbf{k}_l\mathbf{l}_l^\top$ is

$$\begin{aligned} \mathbf{E}_{\mathsf{x},\mathsf{y}}[k(q_i,\mathsf{x})l(r_j,\mathsf{y})] &= \int_X\int_Y k(x-q_i)l(y-r_j)\mathbf{p}_{\mathsf{x},\mathsf{y}}(x,y)dxdy \\ &= \left[k(x)l(y)\star\mathbf{p}_{\mathsf{x},\mathsf{y}}(x,y)\right](q_i,r_j), \end{aligned}$$

38. We use a sans-serif font to define $\mathbf{k}_l$ and $\mathbf{l}_l$, to indicate that these are random vectors. In addition, Bach and Jordan (2002a) define these quantities without multiplying by Δ_x and Δ_y, but we believe these scalings to be necessary: see below.

which is the convolution of the product of kernels with the underlying (unknown) density $\mathbf{p}_{\mathsf{x},\mathsf{y}}(x,y)$ of the random variables x,y, evaluated at q_i, r_j. Since the kernels are normalised, the above expectation is also a probability density, smoothed by $k(x)l(y)$. Similarly,

$$\begin{aligned}\mathbf{E}_{\mathsf{x}}\left[k(q_i,\mathsf{x})k(q_j,\mathsf{x})\right] &= \int_X k(x-q_i)k(x-q_j)\mathbf{p}_{\mathsf{x}}(x)dx \\ &\approx \begin{cases} \left[k^2(x)\star\mathbf{p}_{\mathsf{x}}(x)\right](q_i) & i=j \\ 0 & \text{otherwise} \end{cases},\end{aligned}$$

where the above assumes $\sigma_x \ll \Delta_x \ll 1$. Note, however, that

$$\begin{aligned}k^2(x-q_i) &= \frac{1}{2\pi\sigma_x^2}\exp\left(-\frac{(x-q_i)^2}{\sigma_x^2}\right) &(56)\\ &= \frac{1}{2\sigma_x\sqrt{\pi}}\times\frac{1}{\sqrt{\pi\sigma_x^2}}\exp\left(-\frac{(x-q_i)^2}{\sigma_x^2}\right), &(57)\end{aligned}$$

and thus $k^2(x)$ is *not* a probability density (the integral over $\mathbb{R}$ is equal to $\frac{1}{2\sigma_x\sqrt{\pi}}$). Finally,

$$\begin{aligned}\mathbf{E}_{\mathsf{x}}\left[k(q_{i,},\mathsf{x})\right] &= \int_{\mathbb{R}} k(x-q_i)\mathbf{p}_{\mathsf{x}}(x)dx \\ &= \left[k(x)\star\mathbf{p}_{\mathsf{x}}(x)\right](q_i).\end{aligned}$$

In the light of these observations, it might seem that the relations in (54) ought to hold in the limit as $\Delta_x,\Delta_y\to 0$ and $\sigma_x,\sigma_y\to 0$, so long as $\sigma_x \ll \Delta_x$ and $\sigma_y \ll \Delta_y$: the grid size must be small to allow us to make the approximations

$$\mathbf{P}_{\hat{\mathsf{x}}}(i) = \int_{q_i}^{q_i+\Delta_x}\mathbf{p}_{\mathsf{x}}(x)\,dx \approx \Delta_x\mathbf{p}_{\mathsf{x}}(q_i)$$

and

$$\mathbf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j) = \int_{q_i}^{q_i+\Delta_x}\int_{r_j}^{r_j+\Delta_y}\mathbf{p}_{\mathsf{x},\mathsf{y}}(xy)dxdy \approx \Delta_x\Delta_y\mathbf{p}_{\mathsf{x},\mathsf{y}}(q_i,r_j),$$

and the kernel size is made small so that the kernel functions approach delta functions (although the squared kernel functions do not do so). In other words, the limit in the kernel size is taken *before* the limit in the grid size. We can then write population expression for the kernel generalised variance, in the limit of small kernel size, as

$$\begin{aligned}&\lim_{\sigma_x,\sigma_y\to 0} I(\mathbf{k}_l;\mathbf{l}_l) \\ =\ &\lim_{\sigma_x,\sigma_y\to 0} -\frac{1}{2}\log\left(\left|\mathbf{I}-\left(\mathbf{E}_{\mathsf{x},\mathsf{y}}\left(\mathbf{k}_l\mathbf{l}_l^\top\right)-\mathbf{E}_{\mathsf{x}}\left(\mathbf{k}_l\right)\mathbf{E}_{\mathsf{y}}\left(\mathbf{l}_l^\top\right)\right)^\top\left(\mathbf{E}_{\mathsf{x}}\left(\mathbf{k}_l\mathbf{k}_l^\top\right)-\mathbf{E}_{\mathsf{x}}\left(\mathbf{k}_l\right)\mathbf{E}_{\mathsf{x}}\left(\mathbf{k}_l^\top\right)\right)^{-1}\right.\right. \\ &\qquad\left.\left.\times\left(\mathbf{E}_{\mathsf{x},\mathsf{y}}\left(\mathbf{k}_l\mathbf{l}_l^\top\right)-\mathbf{E}_{\mathsf{x}}\left(\mathbf{k}_l\right)\mathbf{E}_{\mathsf{y}}\left(\mathbf{l}_l^\top\right)\right)\left(\mathbf{E}_{\mathsf{y}}\left(\mathbf{l}_l\mathbf{l}_l^\top\right)-\mathbf{E}_{\mathsf{y}}\left(\mathbf{l}_l\right)\mathbf{E}_{\mathsf{y}}\left(\mathbf{l}_l^\top\right)\right)^{-1}\right|\right) \\ \approx\ &\lim_{\sigma_x,\sigma_y\to 0} -\frac{1}{2}\log\left(\left|\mathbf{I}-\left(\mathbf{P}_{xy}-\mathbf{p}_x\mathbf{p}_y^\top\right)^\top\left(\frac{\Delta_x}{2\sigma_x\sqrt{\pi}}\mathbf{D}_x-\mathbf{p}_x\mathbf{p}_x^\top\right)^{-1}\right.\right. \\ &\qquad\left.\left.\times\left(\mathbf{P}_{xy}-\mathbf{p}_x\mathbf{p}_y^\top\right)\left(\frac{\Delta_y}{2\sigma_y\sqrt{\pi}}\mathbf{D}_y-\mathbf{p}_y\mathbf{p}_y^\top\right)^{-1}\right|\right) \\ =\ &0,\end{aligned}$$

where we use the expression for the squared kernel in (57). In other words, $I(\mathbf{k}_l;\mathbf{l}_l)$ does *not* approach $I(\hat{\mathrm{x}};\hat{\mathrm{y}})$ as the kernel size decreases. This problem reveals the need to enforce the opposite assumption to that made above, namely $\sigma_x \gg \Delta_x$ and $\sigma_y \gg \Delta_y$ (see Section 3.1.4).[39]

We conclude this section with a brief discussion of the link between the empirical estimate of $I(\mathbf{k}_l;\mathbf{l}_l)$ and the KGV. As described by Bach and Jordan (2002a) and by Gretton (2003, Section 9.2.3, Appendix D.5.2), an empirical estimate of $I(\mathbf{k}_l;\mathbf{l}_l)$ is obtained via the usual expression (9), where ρ_i are the solutions to the generalised eigenvalue problem

$$\begin{bmatrix} \mathbf{0} & \mathbf{K}_l\mathbf{H}(\mathbf{L}_l)^\top \\ \mathbf{L}_l\mathbf{H}(\mathbf{K}_l)^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} = \rho_i \begin{bmatrix} \mathbf{K}_l\mathbf{H}(\mathbf{K}_l)^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_l\mathbf{H}(\mathbf{L}_l)^\top \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix}, \tag{58}$$

and $\mathbf{K}_l$ and $\mathbf{L}_l$ are defined in Section 3.1.4 (replacing the Parzen windows with the appropriate RKHS kernels). This is simply the kernel CCA problem, but with the solutions expressed in terms of linear combinations of the grid points $\boldsymbol{q}$ and $\boldsymbol{r}$ mapped into F and G, respectively. As the grid becomes infinitely fine, and assuming $k(\cdot,\cdot)$ and $l(\cdot,\cdot)$ to be continuous, we recover the standard kernel CCA formulation.[40]

39. Also bear in mind that the expression for the KGV used in practice is defined in the limit of infinitely small grid size, but with finite kernel size, rather than vice versa. That said, the ratios $\frac{\Delta_x}{\sigma_x}$ and $\frac{\Delta_y}{\sigma_y}$ suggest a possible resolution of this convergence problem might be to decrease the kernel size and the grid spacing at the same time, as the number of samples rises.

40. This is not a proof - we would need to formally establish both convergence of the kernel CCA solutions in the limit of an infinitely fine grid size, and to demonstrate that the converged solutions lie in the span of the mapped data. These details fall outside the scope of the present study.

References

S. Achard, D.-T. Pham, and C. Jutten. Blind source separation in post-nonlinear mixtures. In *3rd International Conference on ICA and BSS*, 2001.

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

S. Akaho. A kernel method for canonical correlation analysis. In *Proceedings of the International Meeting of the Psychometric Society (IMPS2001)*, 2001.

S.-I. Amari, A. Cichoki, and Y. H. A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems*, volume 8, pages 757–763. MIT Press, 1996.

F. Bach and M. Jordan. Kernel independent component analysis - (matlab code, version 1.1). `http://www.cs.berkeley.edu/~fbach/kernel-ica/index.htm`

F. Bach and M. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3: 1–48, 2002a.

F. Bach and M. Jordan. Tree-dependent component analysis. In *Uncertainty in Artificial Intelligence*, volume 18, 2002b.

C. R. Baker. Mutual information for gaussian processes. *SIAM Journal on Applied Mathematics*, 19(2): 451–458, 1970.

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

G. Bakır, A. Gretton, M. Franz, and B. Schölkopf. Multivariate regression with stiefel constraints. Technical Report 101, Max Planck Institute for Biological Cybernetics, 2004.

A. Bell and T. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–444, 1997.

A. Belouchrani and M. G. Amin. Blind source separation based on time-frequency signal representations. *IEEE Transactions on Signal Processing*, 46(11):2888–2897, 1998.

R. N. Bracewell. *The Fourier Transform and its Applications*. McGraw Hill, New York, 1986.

L. Breiman and J. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80:580–598, 1985.

J.-F. Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 90(8):2009–2026, 1998a.

J.-F. Cardoso. Multidimensional independent component analysis. In *ICASSP*, 1998b.

A. Chen and P. Bickel. Consistent independent component analysis and prewhitening. Technical report, Berkeley, 2004.

A. Cichocki and S.-I. Amari. *Adaptive Blind Signal and Image Processing*. John Wiley and Sons, New York, 2002.

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

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.

N. Cristianini, J. Shawe-Taylor, and J. Kandola. Spectral kernel methods for clustering. In *NIPS*, volume 14, Cambridge, MA, 2002. MIT Press.

J. Dauxois and G. M. Nkiet. Nonlinear canonical analysis and independence tests. *Annals of Statistics*, 26 (4):1254–1278, 1998.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, New York, second edition, 2001.

A. Edelman, T. Arias, and S. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2(Dec):243–264, 2001.

K. Fukumizu, F. Bach, and A. Gretton. Consistency of kernel canonical correlation analysis. Technical Report 942, Institute of Statistical Mathematics, 2005.

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

M. Greenacre. *Theory and Applications of Correspondence Analysis*. Academic Press, London, 1984.

A. Gretton. *Kernel Methods for Classification and Signal Separation*. PhD thesis, Cambridge University Engineering Department, 2003.

A. Gretton, O. Bousquet, A. Smola, and B. Schoelkopf. Measuring statistical dependence with hilbert-schmidt norms. Technical Report 140, MPI for Biological Cybernetics, 2005a.

A. Gretton, R. Herbrich, and A. Smola. The kernel mutual information. Technical report, Cambridge University Engineering Department and Max Planck Institute for Biological Cybernetics, 2003a.

A. Gretton, R. Herbrich, and A. Smola. The kernel mutual information. In *ICASSP*, volume 4, pages 880–883, 2003b.

A. Gretton, A. Smola, O. Bousquet, and R. Herbrich. Behaviour and convergence of the constrained covariance. Technical Report 130, MPI for Biological Cybernetics, 2004.

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. In *AISTATS*, volume 10, 2005b.

D. Hardoon, J. Shawe-Taylor, and O. Friman. KCCA for fMRI analysis. In *Proceedings of Medical Image Understanding and Analysis*, London, 2004.

S. Harmeling, A. Ziehe, M. Kawanabe, and K.-R. Müller. Kernel-based nonlinear blind source separation. *Neural Computation*, 15(5):1089–1124, 2003.

S. Haykin. *Neural Networks : A Comprehensive Foundation*. Macmillan, New York, 2nd edition, 1998.

M. Hein and O. Bousquet. Kernels, associated structures, and generalizations. Technical Report 127, Max Planck Institute for Biological Cybernetics, 2004.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

S. Hosseni and C. Jutten. On the separability of nonlinear mixtures of temporally correlated sources. *IEEE Signal Processing Letters*, 10(2):43–46, 2003.

A. Hyvärinen. One-unit contrast functions for independent component analysis: A statistical analysis. In *Proc. IEEE Neural Networks for Signal Processing Workshop*, pages 388–397, 1997.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, New York, 2001.

A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

A. Hyvärinen and M. Plumbley. Optimization with orthogonality constraints: a modified gradient method. Unpublished note, 2002.

Y. I. Ingster. Asymptotically minimax testing of the hypothesis of independence. *Zap. Nauchn. Seminar. LOMI, 153 (1986) pp. 60-72, Translation in J. Soviet. Math.*, 44:466–476, 1989.

J. Jacod and P. Protter. *Probability Essentials*. Springer, New York, 2000.

M. Kuss. Kernel multivariate analysis. Master's thesis, Technical University of Berlin, 2001.

P. Lai and C. Fyfe. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10(5):365–377, 2000.

T.-W. Lee, M. Girolami, A. Bell, and T. Sejnowski. A unifying framework for independent component analysis. *Computers and Mathematics with Applications*, 39:1–21, 2000.

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(3):725–740, 1993.

T. Melzer, M. Reiter, and H. Bischof. Kernel canonical correlation analysis. Technical Report PRIP-TR-65, Pattern Recognition and Image Processing Group, TU Wien, 2001.

E. Miller and J. Fisher III. ICA using spacings estimates of entropy. *JMLR*, 4:1271–1295, 2003.

E. Mourier. Éléments aléatoires dans un éspace de Banach. *Ann. Inst. H. Poincaré Sect B.*, 161:161–244, 1953.

L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15:1191–1253, 2003.

A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw Hill, New York, 1991.

B. Pearlmutter. Music samples to illustrate the context-sensitive generalisation of ICA. `http://www.cs.unm.edu/~bap/demos.html`

D.-T. Pham. Fast algorithms for mutual information based independent component analysis. *IEEE Transactions on Signal Processing*, 2002. Submitted.

D.-T. Pham and P. Garat. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Transactions on Signal Processing*, 45(7):1712–1725, 1997.

A. Rényi. On measures of dependence. *Acta Math. Acad. Sci. Hungar.*, 10:441–451, 1959.

R. Rosipal and L. Trejo. Kernel partial least squares regression in reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 1(2):97–123, 2001.

A. Samarov and A. Tsybakov. Nonparametric independent component analysis. *Bernoulli*, 10:565–582, 2004.

B. Schölkopf and A. Smola. *Learning with Kernels*. MIT press, Cambridge, MA, 2002.

B. Schölkopf, A. J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK, 2004.

B. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York, 1986.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *JMLR*, 2, 2001.

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

A. Taleb and C. Jutten. Source separation in post-nonlinear mixtures. *IEEE Transactions on Signal Processing*, 47(10):2807–2820, 1999.

F. Theis. Blind signal separation into groups of dependent signals using joint block diagonalisation. In *ISCAS*, pages 5878–5881, 2005.

T. van Gestel, J. Suykens, J. de Brabanter, B. de Moor, and J. Vanderwalle. Kernel canonical correlation analysis and least squares support vector machines. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*. Springer Verlag, 2001.

X.-L. Zhu and X.-D. Zhang. Adaptive RLS algorithm for blind source separation using a natural gradient. *IEEE Signal Processing Letters*, 9(12):432–435, 2002.

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