THE KERNEL MUTUAL INFORMATION

Arthur Gretton

MPI for Biological Cybernetics
Spemannstr 38
D-72076 Tübingen - Germany
arthur@tuebingen.mpg.de

Ralf Herbrich

Microsoft Research
7 J J Thomson Avenue
Cambridge CB3 0FB, UK
rherb@microsoft.com

Alexander J. Smola

RSISE, MLG
Australian National University
Canberra, Australia
Alex.Smola@anu.edu.au

ABSTRACT

We introduce a new contrast function, the kernel mutual information (KMI), to measure the degree of independence of continuous random variables. This contrast function provides an approximate upper bound on the mutual information, as measured near independence, and is based on a kernel density estimate of the mutual information between a discretised approximation of the continuous random variables. We show that Bach and Jordan's kernel generalised variance (KGV) is also an upper bound on the same kernel density estimate, but is looser. Finally, we suggest that the addition of a regularising term in the KGV causes it to approach the KMI, which motivates the introduction of this regularisation.

1. INTRODUCTION

The problem of separating mixtures of signals, so as to recover the original signals prior to mixing, is a much studied challenge in signal processing. Methods of solution generally depend on the nature of the signals, and the manner in which they are mixed; in particular, a criterion known as the *contrast function* is required to determine when the demixing is successful. We assume here that the original signals are generated i.i.d. according to some unknown probability distributions, and are combined in a scalar mixing process: demixing is then achieved by ensuring that the recovered signals are statistically independent. This is the framework for *instantaneous ICA*[1], and has been used successfully in a wide variety of problems: for instance, the separation of linearly mixed audio signals, and the recovery of evoked potentials from EEG signals (see [10, 4], and references therein).

A measure of statistical independence between two random variables is the *mutual information* [5], which for random vectors $\vec{x}, \vec{y}$ is zero if and only if the random vectors are independent. This may also be interpreted as the KL divergence $D_{\mathrm{KL}}\left(\mathbf{f}_{\vec{x},\vec{y}} || \mathbf{f}_{\vec{x}}\mathbf{f}_{\vec{y}}\right)$ between the joint density $\mathbf{f}_{\vec{x},\vec{y}}$ and the product of the marginal densities $\mathbf{f}_{\vec{x}}\mathbf{f}_{\vec{y}}$; the latter quantity generalises readily to distributions of more than two random variables. We therefore propose two quantities, based on the mutual information, that may be used as contrast functions in ICA. The first, which we call the kernel covariance (KC), can be shown to be zero if and only if the random variables are independent. The second function, the kernel mutual information (KMI), is an upper bound on the Parzen window estimate of the mutual information, and is also zero if and only if the random variables are independent. Both functions bear a strong resemblance to the kernel canonical correlation (KCC) and kernel generalised variance (KGV) introduced by Bach and Jordan [3]: indeed, we demonstrate that the KGV can also be thought of as a (looser) upper bound on the same Parzen window estimate. An important advantage of the derivation described herein, however, is that it addresses the behaviour of the contrast functions for finite kernel sizes, rather than relying on a limiting argument in which the kernel size approaches zero, as in [3]. Our approach thus allows us to apply well established methods for selecting kernel size as a function of the number of observations; see for instance [14].

In Section 2, we introduce the ICA problem, and describe our terminology. We then introduce the KC and KCC in Section 3, and derive the KMI and KGV in Section 4. Finally, we show in Section 5 that the performance of the KMI, when used in ICA, is competitive with that of the KGV, and that both the KMI and KGV outperform many traditional ICA algorithms.

2. ICA: PROBLEM STATEMENT

We begin by introducing the ICA problem.The discussion draws on the numerous existing surveys of ICA and related methods; see for instance [10, 4]. Suppose we have a random vector $\vec{\mathsf{s}}$ of dimension N, with independent identically distributed (i.i.d.) components (we use the sans serif to indicate random variables);

$$\mathbf{f}_{\vec{\mathsf{s}}}(\vec{s}) = \prod_{i=1}^{N} \mathbf{f}_{\mathsf{s}_i}(s_i),$$

where $s_i \in \mathbb{R}$. We do not observe $\vec{\mathsf{s}}$, however: instead, we observe the random vector $\vec{\mathsf{t}}$, such that

$$\vec{\mathsf{t}} = \mathbf{A}\vec{\mathsf{s}}, \tag{2.1}$$

where $\mathbf{A}$ is an $N \times N$ matrix[2]. Clearly, the components of $\vec{\mathsf{t}}$ will not be independent unless $\mathbf{A} = \mathbf{PS}$, where $\mathbf{P}$ is a permutation matrix and $\mathbf{S}$ is a diagonal scaling matrix. Our goal is to find an approximation $\mathbf{V}$ to the *inverse* of the matrix[3] $\mathbf{A}$, given m i.i.d. samples from $\mathbf{f}_{\vec{\mathsf{t}}}$, and using *only* the model (2.1) and the fact that the unmixed components are independent. The determination of $\mathbf{A}$ can only be made within certain identifiability constraints, however; in particular, no more than one source can be Gaussian.

Assume we have m observations $\boldsymbol{t} := (\vec{t}_1, \ldots, \vec{t}_m)$. Our first step in computing $\mathbf{V}$ is to subtract the mean of $\boldsymbol{t}$ from each $\vec{t}_i$, and to whiten it, $\tilde{\vec{t}} = \mathbf{Q}\vec{t}$, such that the new observations $\tilde{\vec{t}}$ have

[1] We shall in future refer to this problem simply as ICA.

[2] This corresponds to the number of sources being equal to the number of sensors. In fact, it is possible to recover (2.1) when the number of sources is less than the number of sensors by a change of basis, although the presence of noise makes this more difficult.

[3] Up to permutation and scaling.

a unit covariance matrix. Our estimate of the demixing matrix then becomes $\mathbf{V} := \mathbf{WQ}$, where $\mathbf{W}$ is an orthogonal matrix; our estimate of $\vec{s}$ is $\vec{x} := \mathbf{W}\vec{t}$. Although the determination of $\mathbf{W}$ remains difficult, there are only $N(N-1)$ degrees of freedom involved in this problem, as opposed to the N^2 degrees of freedom present in the estimation of $\mathbf{V}$.

3. THE KERNEL COVARIANCE AND CORRELATION

We now describe the kernel covariance, which is proposed as a measure of statistical independence of the random vectors $\vec{x}$ and $\vec{y}$, defined on $\mathcal{X} := \mathbb{R}^{n_x}$ and $\mathcal{Y} := \mathbb{R}^{n_y}$. The generalisation to more than two vectors is addressed in [8]. We define the vectors $\mathbf{x}$ and $\mathbf{y}$ and the random vectors x and y in the feature spaces $\mathcal{F}_\mathcal{X}$ and $\mathcal{F}_\mathcal{Y}$, and the mappings $\boldsymbol{\phi}_x : \mathcal{X} \to \mathcal{F}_\mathcal{X}$ and $\boldsymbol{\phi}_y : \mathcal{Y} \to \mathcal{F}_\mathcal{Y}$ such that

$$\mathbf{x} := \boldsymbol{\phi}_x(\vec{x}) \quad \text{and} \quad \mathbf{y} := \boldsymbol{\phi}_y(\vec{y}).$$

The feature spaces may be the reproducing kernel Hilbert spaces (and subspaces of ℓ_2^∞) associated with particular kernels, which represent the inner products[4] on $\mathcal{F}_\mathcal{X}$ and $\mathcal{F}_\mathcal{Y}$. We define

$$\mathbf{C}_{xy} := \mathbf{E}_{\mathsf{x},\mathsf{y}}\left((\mathbf{x} - \mathbf{E}_\mathsf{x}(\mathbf{x}))(\mathbf{y} - \mathbf{E}_\mathsf{y}(\mathbf{y}))^\top\right), \quad (3.1)$$

$$\mathbf{C} := \begin{bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{xy}^\top & \mathbf{C}_{yy} \end{bmatrix}, \quad (3.2)$$

where $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ are given by analogy. We observe m i.i.d. samples of data; $\boldsymbol{z} = ((\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_m, \mathbf{y}_m))$, where $\mathbf{x}_i \in \mathcal{F}_\mathcal{X}$ and $\mathbf{y}_i \in \mathcal{F}_\mathcal{Y}$.

We may also define the *Gram matrices* $\mathbf{K}_{mm}^{(x)}$, $\mathbf{K}_{mm}^{(y)}$ of inner products between the mapped observations above, in the case where $\mathcal{F}_\mathcal{X}$ and $\mathcal{F}_\mathcal{Y}$ are reproducing kernel Hilbert spaces (RKHSs) with associated kernels $k(\vec{x}_i, \vec{x}_j) := \mathbf{x}_i^\top \mathbf{x}_j = (\mathbf{K}_{mm}^{(x)})_{i,j}$ and $k(\vec{y}_i, \vec{y}_j) := \mathbf{y}_i^\top \mathbf{y}_j = (\mathbf{K}_{mm}^{(y)})_{i,j}$. According to [13], Gram matrices for the variables centred *in feature space* are $\tilde{\mathbf{K}}_{mm}^{(x)} := \mathbf{H}\mathbf{K}_{mm}^{(x)}\mathbf{H}$, $\tilde{\mathbf{K}}_{mm}^{(y)} := \mathbf{H}\mathbf{K}_{mm}^{(y)}\mathbf{H}$, where $\mathbf{H} = \mathbf{I}_m - m^{-1}\mathbf{1}_m\mathbf{1}_m^\top$, and $\mathbf{1}_m$ is an $m \times 1$ vector of ones.

We can now introduce the kernel covariance (KC). In the population case, the KC is

$$\mathcal{I} = \sup_{f \in \tilde{\mathcal{F}}_\mathcal{X},\, g \in \tilde{\mathcal{F}}_\mathcal{Y}} \left|\mathbf{E}_{\vec{x},\vec{y}}[f(\vec{x})g(\vec{y})] - \mathbf{E}_{\vec{x}}[f(\vec{x})]\,\mathbf{E}_{\vec{y}}[g(\vec{y})]\right|,$$

where $\tilde{\mathcal{F}}_\mathcal{X} := \{f \in \mathcal{F}_\mathcal{X} : \|f\|_{\mathcal{F}_\mathcal{X}} \leq 1\}$, and $\tilde{\mathcal{F}}_\mathcal{Y}$ is analogous. An empirical estimate $\mathcal{I}(\boldsymbol{z})$ may be obtained from the finite sample $\boldsymbol{z}$, using the representer theorem (Schölkopf *et al.* [12]) to replace

$$f(\vec{x}) = \sum_{l=1}^m c_l k(\vec{x}, \vec{x}_l) = \sum_{l=1}^m c_l \mathbf{x}^\top \mathbf{x}_l, \quad (3.3)$$

with a similar replacement for $g(\vec{y})$; it follows that $\mathcal{I}(\boldsymbol{z}) := \max_i \gamma_i$, where γ_i are the eigenvalues of

$$\begin{bmatrix} \tilde{\mathbf{K}}_{mm}^{(x)} & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{K}}_{mm}^{(y)} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{0} & \tilde{\mathbf{K}}_{mm}^{(x)}\tilde{\mathbf{K}}_{mm}^{(y)} \\ \tilde{\mathbf{K}}_{mm}^{(y)}\tilde{\mathbf{K}}_{mm}^{(x)} & \mathbf{0} \end{bmatrix}$$

We now describe the link between the kernel covariance and independence; details are given in [8].

[4] To be a kernel associated with a RKHS, $k(\vec{x}_i, \vec{x}_j)$ must satisfy the Mercer conditions [1]; these hold for Gaussian and Laplace kernels, among (many) others. Note also that the argument of the kernel specifies whether the kernel pertains to $\mathcal{F}_\mathcal{X}$ or $\mathcal{F}_\mathcal{Y}$, although these kernels are identical in the present study.

Theorem 1 (Kernel covariance and independence). $\mathcal{I} = 0$ *if and only if* x, y *are independent.*

Finally, we introduce the canonical correlation, as described in [7, 3, 11]; the final reference is a particularly insightful investigation of the canonical correlation in high dimensional spaces (such as RKHSs). We first define the canonical correlation in the general case, without reference to its interpretation when $\mathcal{F}_\mathcal{X}, \mathcal{F}_\mathcal{Y}$ are RKHSs. We would like to find vectors $\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i$ onto which $\mathbf{x}$ and $\mathbf{y}$ respectively project, such that the *correlation* ρ_i between these projections is a stationary point with respect to $\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i$. The canonical correlations, ρ_i, are thus given by

$$\rho_i := \frac{\boldsymbol{\alpha}_i^\top \mathbf{C}_{xy}\boldsymbol{\beta}_i}{\sqrt{\left(\boldsymbol{\alpha}_i^\top \mathbf{C}_{xx}\boldsymbol{\alpha}_i\right)\left(\boldsymbol{\beta}_i^\top \mathbf{C}_{yy}\boldsymbol{\beta}\right)}}. \quad (3.4)$$

When $\mathcal{F}_\mathcal{X}$ and $\mathcal{F}_\mathcal{Y}$ are RKHSs, then care must be taken when finding empirical estimates of the canonical correlates, to ensure that these estimates are data dependent. This may be done by confining $\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i$ to subspaces of the space spanned by the sample in $\mathcal{F}_\mathcal{X}, \mathcal{F}_\mathcal{Y}$, as in Kuss [11], or by regularising, as in [3]; in the latter case, the largest kernel canonical correlation may be used as a contrast function for ICA.

4. UPPER BOUNDS ON MUTUAL INFORMATION

We now apply both these definitions to derive an approximation of the mutual information between random variables x and y, defined on the respective bounded intervals $\mathcal{X}$ and $\mathcal{Y}$ on $\mathbb{R}$. Full details of the proofs, and a generalisation to more than two random variables, may be found in [8]. We begin by introducing the Gaussian mutual information, and its relation with the canonical correlation. If $\vec{x}_\mathrm{G}, \vec{y}_\mathrm{G}$ are Gaussian random variables in $\mathbb{R}^{p_x}, \mathbb{R}^{p_y}$ respectively, then according to [3] the mutual information between them can be written

$$I(\vec{x}_\mathrm{G}; \vec{y}_\mathrm{G}) = -\frac{1}{2}\log\left(\prod_{i=1}^{\min(p_x,p_y)}\left(1 - \rho_i^2\right)\right), \quad (4.1)$$

where the ρ_i are given by the canonical correlations in (3.4).

Next, consider a grid of size $p_x \times p_y$ over $\mathcal{X}$ and $\mathcal{Y}$ respectively. Let the indices i, j denote the point $(q_i, r_j) \in \mathcal{X} \times \mathcal{Y}$ on this grid, and let $\boldsymbol{q} := (q_1, \ldots, q_{p_x})$, $\boldsymbol{r} := (r_1, \ldots, r_{p_y})$ be the grid coordinates. The spacing between points along the x and y axes is respectively Δ_x and Δ_y. 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 (we write the complete $p_x \times p_y$ matrix of such probabilities as $\mathbf{P}_{xy}$), where

$$\mathbf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j) := \int_{q_i}^{q_i+\Delta_x}\int_{r_j}^{r_j+\Delta_y} \mathbf{f}_{\mathsf{x},\mathsf{y}}(x,y)\,dx\,dy.$$

Thus $\mathbf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j)$ is a discretisation of $\mathbf{P}_{\mathsf{x},\mathsf{y}}$. 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. We may always write $\mathbf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j) = \mathbf{P}_{\hat{\mathsf{x}}}(i)\,\mathbf{P}_{\hat{\mathsf{y}}}(j)\,(1 + \epsilon_{i,j})$ for an appropriate choice of $\epsilon_{i,j}$. If $\epsilon_{i,j}$ is small, we approximate

$$I(\hat{\mathsf{x}}; \hat{\mathsf{y}}) \approx \frac{1}{2}\sum_{i=1}^{p_x}\sum_{j=1}^{p_y} \mathbf{P}_{\hat{\mathsf{x}}}(i)\,\mathbf{P}_{\hat{\mathsf{y}}}(j)\,\epsilon_{i,j}^2, \quad (4.2)$$

It is well known (see [5]) that $I(\mathsf{x}, \mathsf{y})$ represents the upper bound on $I(\hat{\mathsf{x}}; \hat{\mathsf{y}})$ as the discretisation becomes infinitely fine.

We next define an equivalent multidimensional representation $\vec{x}, \vec{y}$ of $\hat{x}, \hat{y}$ in the previous section, where $\vec{x} \in \mathbb{R}^{p_x}$ and $\vec{y} \in \mathbb{R}^{p_y}$, such that $\hat{x} = i$ is equivalent to $(\vec{x})_i = 1$ and $(\vec{x})_{j:j \neq i} = 0$. Using

$$\mathbf{E}_{\mathsf{x},\mathsf{y}}\left(\vec{\mathsf{x}}\vec{\mathsf{y}}^\top\right) = \mathbf{P}_{xy}, \quad \mathbf{E}_{\mathsf{x}}\left(\vec{\mathsf{x}}\right) = \mathbf{p}_x, \quad \mathbf{E}_{\mathsf{x}}\left(\vec{\mathsf{x}}\vec{\mathsf{x}}^\top\right) = \mathbf{D}_x,$$

where $\mathbf{D}_x = \operatorname{diag}(\mathbf{p}_x)$, it is possible to define covariances

$$\mathbf{C}_{xy} = \mathbf{P}_{xy} - \mathbf{p}_x\mathbf{p}_y^\top, \qquad \mathbf{C}_{xx} = \mathbf{D}_x - \mathbf{p}_x\mathbf{p}_x^\top. \tag{4.3}$$

We define the Gaussian random variables $\vec{\mathsf{x}}_{\mathsf{G}}, \vec{\mathsf{y}}_{\mathsf{G}}$ to have the same covariance structure as $\vec{\mathsf{x}}, \vec{\mathsf{y}}$, and with mutual information given by (4.1). The mutual information for this Gaussian case may then be approximated by (4.2) near independence; see [3, 8].

Given that we are not provided with the distribution $\mathbf{P}_{\hat{\mathsf{x}},\hat{\mathsf{y}}}(i,j)$, but rather a finite sample $\boldsymbol{z}$ of size m, we make use of a kernel density estimate of the mutual information for the discretised random variables. A detailed discussion of the properties and behaviour of such estimates may be found in [14], and previous work on their application to the computation of entropies in [9]. The kernel density (Parzen window) estimates of $\mathbf{f}_{\mathsf{x}}$ and $\mathbf{f}_{\mathsf{x},\mathsf{y}}$ are

$$\begin{aligned}\widehat{\mathbf{f}}_{\mathsf{x}}(x) &:= \frac{1}{m}\sum_{l=1}^{m} k(x_l, x), \\ \widehat{\mathbf{f}}_{\mathsf{x},\mathsf{y}}(x,y) &:= \frac{1}{m}\sum_{l=1}^{m} k(x_l, x)\, k(y_l, y).\end{aligned}$$

The kernels must be non-negative and continuous, with unit integral w.r.t. to its two arguments. We require approximations to the covariance matrices in the Gaussian mutual information, as described in (4.3). We therefore define the vectors $\hat{\mathbf{p}}_x, \hat{\mathbf{p}}_y$, and the matrix $\widehat{\mathbf{P}}_{xy}$, using the expectations computed with these kernel expressions;

$$\widehat{\mathbf{E}}_{\mathsf{x}\mathsf{y}}\left(\vec{\mathsf{x}}\vec{\mathsf{y}}^\top\right) = \widehat{\mathbf{P}}_{xy}, \quad \widehat{\mathbf{E}}_{\mathsf{x}}\left(\vec{\mathsf{x}}\right) = \hat{\mathbf{p}}_x, \quad \widehat{\mathbf{E}}_{\mathsf{x}}\left(\vec{\mathsf{x}}\vec{\mathsf{x}}^\top\right) = \widehat{\mathbf{D}}_x.$$

Let $\mathbf{K}_{pm}^{(x)}$ be the matrix of inner products in $\mathcal{F}_{\mathcal{X}}$ between the grid points and sample, with $\mathbf{K}_{pm}^{(y)}$ defined by analogy; we assume $p_x \gg m$ and $p_y \gg m$. The first subscript specifies whether the grid ($\boldsymbol{q}$ or $\boldsymbol{r}$) or the sample ($\boldsymbol{x}$ or $\boldsymbol{y}$) is used in the rows of this matrix, and the second subscript whether the grid or sample is used in the columns. By analogy, we may also define the matrices $\mathbf{K}_{pp}^{(x)}, \mathbf{K}_{mm}^{(x)}, \mathbf{K}_{pp}^{(y)}, \mathbf{K}_{mm}^{(y)}$. In the limit where Δ_x, Δ_y are small (and thus, by implication, $p_x \gg m$, $p_y \gg m$, $\sigma \gg \Delta_x$, and $\sigma \gg \Delta_y$, where σ defines the kernel size), we make the approximations

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

$$\widehat{\mathbf{D}}_x = \frac{\Delta_x}{m}\operatorname{diag}\left(\mathbf{K}_{pm}^{(x)}\mathbf{1}_m\right) = \frac{\Delta_x^2}{m}\operatorname{diag}\left(\mathbf{K}_{pm}^{(x)}\left(\mathbf{K}_{pm}^{(x)}\right)^\top\mathbf{1}_{p_x}\right).$$

The kernel density approximation to the discretised mutual information is then found by replacing the ρ_i in (4.1) with

$$\hat{\rho}_i := \frac{\hat{\mathbf{c}}_i^\top\left(\widehat{\mathbf{P}}_{xy} - \hat{\mathbf{p}}_x\hat{\mathbf{p}}_y^\top\right)\hat{\mathbf{d}}_i}{\sqrt{\hat{\mathbf{c}}_i^\top\widehat{\mathbf{D}}_x\hat{\mathbf{c}}_i\,\hat{\mathbf{d}}_i^\top\widehat{\mathbf{D}}_y\hat{\mathbf{d}}_i}}. \tag{4.4}$$

This cannot easily be computed, however, since it is computationally prohibitive to evaluate the Gram matrices on a sufficiently fine grid. Noting that

$$\breve{\nu}_{\boldsymbol{x}}\hat{\mathbf{c}}_i^\top\mathbf{K}_{pp}^{(x)}\hat{\mathbf{c}}_i \leq \hat{\mathbf{c}}_i^\top\operatorname{diag}\left(\mathbf{K}_{pm}^{(x)}\left(\mathbf{K}_{pm}^{(x)}\right)^\top\mathbf{1}_p\right)\hat{\mathbf{c}}_i, \tag{4.5}$$

where $\breve{\nu}_{\boldsymbol{x}} := \min_{j\in\{1\ldots p_x\}}\sum_{l=1}^{m} k(x_l, q_j)$, we replace the matrix term in $\widehat{\mathbf{D}}_x$ by the left hand expression in (4.5), yielding a new quantity $|\tilde{\gamma}_i| \geq |\hat{\rho}_i|$; it follows that replacing $\hat{\rho}_i$ with $\tilde{\gamma}_i$ yields an upper bound on (4.1). In fact, $\tilde{\gamma}_i$ is simply the kernel covariance, but with the additional requirement that the functions f, g be projected in their respective feature spaces onto the basis spanned by the columns of the grid $\boldsymbol{q}, \boldsymbol{r}$, as well as an added scaling factor $\breve{\nu}_{\boldsymbol{x}}$. We use this insight to replace $\tilde{\gamma}_i$ in (4.1) with an appropriately scaled γ_i, and $\breve{\nu}_{\boldsymbol{x}}$ with $\nu_{\boldsymbol{x}} := \min_{j\in\{1\ldots m\}}\sum_{l=1}^{m} k(x_l, x_j)$, to obtain the empirical *kernel mutual information* (KMI),

$$\mathcal{M}(\boldsymbol{z}) := -\frac{1}{2}\log\left(\left|\mathbf{I} - (\nu_{\boldsymbol{x}}\nu_{\boldsymbol{y}})\,\widetilde{\mathbf{K}}_{mm}^{(x)}\widetilde{\mathbf{K}}_{mm}^{(y)}\right|\right), \tag{4.6}$$

which is also an upper bound on (4.1). It follows from Theorem 1 that the random variables x, y are independent if and only if the *population* KMI satisfies $\mathcal{M} = 0$.

Bach and Jordan [3] propose a related quantity as a contrast function for ICA: the kernel generalised variance (KGV). In fact, the latter quantity may also be derived by finding an upper bound on (4.1): this is a different approach to the proof in [3], which uses a limit as the kernel becomes infinitely small. Using

$$\hat{\mathbf{c}}_i^\top\mathbf{K}_{pm}^{(x)}\left(\mathbf{K}_{pm}^{(x)}\right)^\top\hat{\mathbf{c}}_i \leq \hat{\mathbf{c}}_i^\top\operatorname{diag}\left(\mathbf{K}_{pm}^{(x)}\left(\mathbf{K}_{pm}^{(x)}\right)^\top\mathbf{1}_{p_x}\right)\hat{\mathbf{c}}_i,$$

we replace the right hand term in the above with the left hand term in the denominator of (4.4) to get a set of kernel canonical correlations $\bar{\rho}_i \geq \hat{\rho}_i$, restricted to the basis spanned by the grid. The mutual information computed using the unrestricted kernel canonical correlations ρ_i is therefore an upper bound on (4.1). The contrast function thus derived is never used in practice, since it is infinite; in other words, the approximation we made above is too loose. If we instead make the replacement

$$\frac{m}{\Delta_x^2}\hat{\mathbf{c}}_i^\top\widehat{\mathbf{D}}_x\hat{\mathbf{c}}_i \Rightarrow \hat{\mathbf{c}}_i^\top\left(\theta_1\mathbf{K}_{pm}^{(x)}\left(\mathbf{K}_{pm}^{(x)}\right)^\top + \theta_2\nu_{\boldsymbol{x}}\mathbf{K}_{pp}^{(x)}\right)\hat{\mathbf{c}}_i,$$

where $\theta_1 \geq 0, \theta_2 \geq 0$, and $\theta_1 + \theta_2 \leq 1$, we recover an expression which, for correct choice of θ_1, θ_2, yields the *regularised* KGV proposed in [3][5]. We therefore expect the performance of both the KGV and KMI to be very similar when used for ICA: this is indeed the case in our experimental results.

We now briefly address the generalisation of the kernel covariance $\mathcal{I}$ to the case of N random variables x_j on bounded subsets $\mathcal{X}_j \subset \mathbb{R}$, by analogy with derivation of [3]; this can be used to measure the pairwise independence of our estimate $\vec{x}$ of the independent components $\vec{s}$. The KC is the largest eigenvalue λ_i of

$$\left(\widetilde{\mathbf{K}}\widetilde{\mathbf{K}}^\top - \begin{bmatrix}\widetilde{\mathbf{K}}_1\widetilde{\mathbf{K}}_1 & & \\ & \ddots & \\ & & \widetilde{\mathbf{K}}_N\widetilde{\mathbf{K}}_N\end{bmatrix}\right)\mathbf{c}_i = \lambda_i\left(\operatorname{diag}(\widetilde{\mathbf{K}})\mathbf{c}_i\right), \tag{4.7}$$

where $\mathbf{c}_i = (\mathbf{c}_{i,1}, \ldots, \mathbf{c}_{i,N})^\top$ and $\widetilde{\mathbf{K}} = [\widetilde{\mathbf{K}}_1, \widetilde{\mathbf{K}}_2, \ldots \widetilde{\mathbf{K}}_N]^\top$. To reduce computational cost, we use a reduced rank approximation of $\widetilde{\mathbf{K}}_j$, via an incomplete Cholesky factorization with appropriate pivoting [6] (that is, $\widetilde{\mathbf{K}}_j \approx \mathbf{Z}_j\mathbf{Z}_j^\top$ with $\mathbf{Z}_j \in \mathbb{R}^{m\times d}$ and $d \ll m$). We set $\mathbf{d}_i = [\mathbf{c}_{i,1}^\top\mathbf{Z}_1, \ldots, \mathbf{c}_{i,N}^\top\mathbf{Z}_N]^\top$, and rewrite (4.7) as

$$\left(\mathbf{Z}\mathbf{Z}^\top - \begin{bmatrix}\mathbf{Z}_1^\top\mathbf{Z}_1 & & \\ & \ddots & \\ & & \mathbf{Z}_N^\top\mathbf{Z}_N\end{bmatrix}\right)\mathbf{d}_i = \lambda_i\mathbf{d}_i, \tag{4.8}$$

[5] Specifically, the parameter denoting the amount of regularisation in [3] can be written $\kappa = \theta_2\nu_{\boldsymbol{x}}/\theta_1$, although we must be careful in our choice of θ_1, θ_2 to ensure we still have an upper bound; see [8] for details.

where $\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2, \ldots \mathbf{Z}_N]^\top$. This transformation takes care of the nullspace inherent in $\mathbf{Z}_j \mathbf{Z}_j^\top$ and reduces the eigenproblem to dN dimensions[6]. Finally, the KMI for more than two variables is

$$\mathcal{M}(z) := -\frac{1}{2} \log \left(\prod_{i=1}^{N} \nu_z \lambda_i \right) \tag{4.9}$$

where $\nu_z = \min_j \nu_{x_j}$ (in our experiments, we simply set $\nu_z = 1/m$; the performance remained satisfactory).

5. EXPERIMENTAL RESULTS

We now apply the KGV and KMI to the problem of ICA. Since the main purpose is to compare the performance with that reported in [3], we use identical settings and data. The mixing matrix $\mathbf{A}$ was chosen randomly, with condition number between 1 and 2. We used the Gaussian RBF kernel, $k(x, x') = \exp(-\frac{1}{2\sigma^2}\|x - x'\|^2)$, with $\sigma^2 = \frac{1}{4}$ and $\kappa = 2 \times 10^{-3}$ for the KGV, except in the case of the 250 point sample, where $\sigma^2 = 1$ and $\kappa = 2 \times 10^{-2}$. We only used $\sigma^2 = 1$ for the KMI. The orthogonal component $\mathbf{W}$ of the demixing matrix was found using gradient descent on the manifold of orthogonal matrices; see [3]. In order to measure the distance between the true ($\mathbf{A}^{-1}$) and approximate ($\mathbf{WQ}$) demixing matrices, we used the *Amari divergence* [4]. This metric is in the interval $[0, 100]$, is equal to zero if and only if $\mathbf{A}^{-1}, \mathbf{WQ}$ are piecewise identical, and is invariant to permutation and scaling of $\mathbf{A}^{-1}, \mathbf{WQ}$.

Our experiment consisted in de-mixing data drawn independently from 2 – 16 distributions, chosen at random with replacement from 18 possible options; these include signals with both positive and negative kurtosis, and are described in detail in [3, 8]. Table 1 summarises our results; the KMI seems somewhat better in the case of larger m and N, although further refinement of the parameter choices in both methods might be possible. Further experiments are described in [8], most notably addressing the problem of recovering signals in the presence of noise, and in the case of low kurtosis. In these cases, the KMI and KGV again yield the best observed performance.

Table 1. Illustration of the demixing of N randomly chosen signals; comparison with fast ICA, Jade, and extended Infomax. The best result is in boldface.

N	m	Rep.	fICA	Jade	Imax	KGV	KMI
2	250	1000	11.6± 0.4	10.6± 0.4	46.7± 0.9	**5.4± 0.2**	6.2 ± 0.2
2	1000	1000	6.2 ± 0.2	4.8 ± 0.2	10.9± 0.6	**2.5± 0.1**	2.8 ± 0.1
4	1000	100	6.0 ± 0.4	5.5 ± 0.4	10.7± 0.9	**3.5± 0.4**	3.7 ± 0.7
4	4000	100	3.3 ± 0.2	2.7 ± 0.1	6.2 ± 0.7	**1.4± 0.1**	**1.4± 0.05**
8	2000	50	4.0 ± 0.2	3.9 ± 0.3	8.2 ± 0.8	3.7 ± 0.9	**2.9± 0.4**
8	4000	50	3.0 ± 0.3	2.5 ± 0.1	5.6 ± 0.7	1.5 ± 0.1	**1.3± 0.04**
16	4000	28	3.1 ± 0.2	3.3 ± 0.2	11.1± 1.1	3.1 ± 0.9	**2.2± 0.3**

6. CONCLUSIONS

We have presented a novel derivation of several kernel based contrast functions for ICA (the KMI, KGV, and related), which yields useful insight both into the problem of model selection, and the function of the regularising term in these contrasts. The KMI and KGV are comparable in performance, and substantially outperform several alternative ICA approaches. Further work will focus on the application of kernel based contrasts to convolutive mixing, and to the recovery of random processes that are not i.i.d.; an application to graphical model estimation is given in [2].

7. REFERENCES

[1] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337 – 404, 1950.

[2] F. Bach and M. Jordan. Learning graphical models with mercer kernels. In *NIPS 2002 (to appear)*.

[3] F. Bach and M. Jordan. Kernel independent component analysis. *JMLR*, (3):1–48, 2002.

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

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

[6] S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representation. Technical report, IBM Watson Research Center, New York, 2000.

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

[8] A. Gretton, R. Herbrich, and A. Smola. *The Kernel Mutual Information*. Max Planck Institute for Biological Cybernetics, 2002. Forthcoming (draft available on request).

[9] L. Györfi and E. van der Meulen. An entropy estimate based on a kernel density estimation. *Colloquia Mathematica Societatis János Bolyai*, 57: Limit Theorems in Probability and Statistics:229–240.

[10] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, 2001.

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

[12] B. Schölkopf, R. Herbrich, and A. Smola. A generalised representer theorem. In *Proceedings of the Annual Conference on Computational Learning Theory*, 2001.

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

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

[6] Note that there is no reason why all $\mathbf{Z}_j$ should have the same dimensionality: it may in fact be more computationally efficient in some circumstances to use different decompositions for different x_j.