Fast Kernel ICA using an Approximate Newton Method

Hao Shen
National ICT Australia and
Australian National University
Canberra, Australia

Stefanie Jegelka
MPI for Biological Cybernetics,
Spemannstr 38
72076 Tübingen, Germany

Arthur Gretton
MPI for Biological Cybernetics,
Spemannstr 38
72076 Tübingen, Germany

Abstract

Recent approaches to independent component analysis (ICA) have used kernel independence measures to obtain very good performance, particularly where classical methods experience difficulty (for instance, sources with near-zero kurtosis). We present fast kernel ICA (FastKICA), a novel optimisation technique for one such kernel independence measure, the Hilbert-Schmidt independence criterion (HSIC). Our search procedure uses an approximate Newton method on the special orthogonal group, where we estimate the Hessian locally about independence. We employ incomplete Cholesky decomposition to efficiently compute the gradient and approximate Hessian. FastKICA results in more accurate solutions at a given cost compared with gradient descent, and is relatively insensitive to local minima when initialised far from independence. These properties allow kernel approaches to be extended to problems with larger numbers of sources and observations. Our method is competitive with other modern and classical ICA approaches in both speed and accuracy.

1 Introduction

The problem of instantaneous independent component analysis involves the recovery of linearly mixed, independent sources, in the absence of information about the source distributions beyond their mutual independence (Hyvärinen et al., 2001). Classical approaches to this problem, which use as their independence criterion the sum of expectations of a fixed nonlinear function (or a small number of such functions) on each recovered source, scale well to large numbers of sources and samples. On the other hand, they only ensure local convergence in the vicinity of independence (Shen and Hüper, 2006, give one such analysis for FastICA), and do not guarantee independent sources are recovered at the *global* optimum of the independence criterion. Statistical tests of independence should then be applied (as by Ku and Fine, 2005) to verify independent sources are recovered.

A popular modern approach to ICA has been to directly optimise a criterion that measures the statistical independence of the sources: Stögbauer et al. (2004); Chen (2006); Learned-Miller and Fisher III (2003) minimise the mutual information between the sources, Eriksson and Koivunen (2003); Chen and Bickel (2005); Murata (2001) optimise a characteristic function-based independence measure, and Bach and Jordan (2002); Gretton et al. (2005a,b) employ kernel independence measures. While the above studies report excellent performance, efficient optimisation of these dependence measures for ICA remains an ongoing problem,[1] and a barrier to using the approaches when the number of sources, m, is large. ICA is generally decomposed into two sub-problems: signal decorrelation, which is straightforward and is not discussed further, and optimisation over the special orthogonal group $SO(m)$, for which the bulk of the computation is required. Bach and Jordan (2002); Gretton et al. (2005a,b); Chen and Bickel (2005) all do gradient descent on $SO(m)$ in accordance with Edelman et al. (1998), choosing the step width by a Golden search. This is inefficient on two counts: gradient descent can require a very large number of steps for convergence even on relatively benign cost functions, and the Golden search requires many costly evaluations of the independence measure. Although Jegelka and Gretton (2007) propose a cheaper local quadratic approximation to choose the step size, this does not address the question of better search direction choice.

[1] Most of the effort in increasing efficiency has gone into cheaply and accurately approximating the independence measures (Chen, 2006; Bach and Jordan, 2002; Jegelka and Gretton, 2007).

Recent work by Hüper and Trumpf (2004) on Newton-like methods for optimisation on manifolds applies in particular to $SO(m)$. This technique has successfully been used for one-unit ICA (Shen et al., 2006), and a similar fully parallelised method on $SO(m)$ applies to multi-unit ICA (Shen and Hüper, 2006), when independence measures from FastICA are used. The extension of these Newton-like geometric methods to multi-unit kernel ICA is therefore of interest. In this study, we introduce a novel approximate Newton method to optimise a kernel independence measure on the special orthogonal group, using a similar local Hessian approximation about independence to Shen and Hüper (2006): we call this algorithm fast kernel ICA (FastKICA). An additional advantage of this approximate Newton approach over gradient descent is that it is resistant to local minima, i.e. it converges more often to the correct solution even in the absence of a good initialisation. Previous kernel methods require either a large number of restarts or a good initial guess (provided for instance by another algorithm).

We begin our presentation in Section 2, where we introduce the problem of ICA, and describe independence measures based on covariance operators in RKHSs. We compute the gradient and approximate Hessian of our independence measure in Section 3, and describe a Newton-like method based on these quantities. We also show that incomplete Cholesky approximations of the Gram matrices can be used to speed the algorithm substantially. Finally, we present our experiments in Section 4.

2 Linear ICA and independence measures

We describe the goal of instantaneous independent component analysis (ICA), drawing mainly on (Hyvärinen et al., 2001; Cardoso, 1998), as well as the core properties of ICA explored by Comon (1994). We then describe the Hilbert-Schmidt independence criterion (HSIC), which is the independence measure we optimise when doing ICA. We are given an $m \times n$ matrix of mixtures C, where m is the number of sources and n the number of samples. We denote as c a particular column of the mixture matrix, which corresponds to a single sample of all the mixtures. Each sample is drawn independently and identically from the distribution $\mathbf{P}_c$. The matrix C is related to the matrix S of *sources* (also of dimension $m \times n$) by the linear mixing process

$$C = AS, \tag{1}$$

where A is an $m \times m$ matrix with full rank. We refer to our ICA problem as being *instantaneous* as a way of describing the dual assumptions that any observation vector c depends only on the source vector s at that instant, and that the samples s are drawn independently and identically.

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). Random variables are mutually independent if and only if their probability distribution factorises, $\mathbf{P}_s = \prod_{i=1}^{m} \mathbf{P}_{s_i}$. It follows easily that the random variables are *pairwise* independent if they are *mutually* independent, where pairwise independence is defined as $\mathbf{P}_{s_i}\mathbf{P}_{s_j} = \mathbf{P}_{s_i s_j}$ for all $i \neq j$. The reverse does not hold, however: pairwise independence does not imply mutual independence. That said, we are able to find a unique optimal unmixing matrix using only the *pairwise* independence between elements of the estimated sources Y, which is equivalent to recovering the *mutually* independent terms of S. This is due to Theorem 11 of Comon (1994).

The task of ICA is to recover the independent sources via an estimate B of the inverse of the matrix A, such that the recovered vector $Y = BAS$ has mutually independent components.[2] In practice, B is found in two steps: first, the mixtures are decorrelated using a matrix V to give the whitened signals $W = VC$. Next, all remaining dependence is removed using an orthogonal matrix[3] $X \in \mathbb{R}^{m \times m}$ (i.e. $X^\top X = I$), yielding $Y = X^\top W$.

Next, we describe our independence criterion. A version of this bivariate criterion was originally proposed by Feuerverger (1993). Gretton et al. (2005a) obtained Feuerverger's independence criterion in a more general setting, proving that it is the Hilbert-Schmidt (HS) norm of the covariance operator between mappings to RKHSs, and thus integrating it into the family of kernel independence measures (Bach and Jordan, 2002; Gretton et al., 2005b). Thus, we introduce the criterion from this perspective. Consider a Hilbert space $\mathcal{F}$ of functions from a compact subset $\mathcal{Y} \subset \mathbb{R}$ to $\mathbb{R}$. The Hilbert space $\mathcal{F}$ is an RKHS if at each $y_u \in \mathcal{Y}$, the point evaluation operator $\delta_{y_u} : \mathcal{F} \to \mathbb{R}$, which maps $f \in \mathcal{F}$ to $f(y_u) \in \mathbb{R}$, is a bounded linear functional. To each point $y_u \in \mathcal{Y}$, there corresponds an

[2] 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 s is swapped, or when components of s are scaled by different constant amounts. Thus, source recovery takes place up to these invariances, and $BA = PD$, where P is a permutation matrix and D a diagonal scaling matrix.

[3] Our notation differs from that in other ICA presentations since X is the variable for which we wish to solve; i.e. we adopt a standard convention used in optimisation.

element $\phi(y_u) \in \mathcal{F}$ (we call ϕ the *feature map*) such that $\langle \phi(y_u), \phi(y'_u) \rangle_{\mathcal{F}} = k(y_u, y'_u)$, where $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a unique positive definite kernel. We also define a second RKHS $\mathcal{G}$ with respect to $\mathcal{Y}$, with feature map ψ and kernel $\langle \psi(y_v), \psi(y'_v) \rangle_{\mathcal{G}} = l(y_v, y'_v)$.

Let $\mathbf{P}_{y_u,y_v}$ be a joint measure on $(\mathcal{Y} \times \mathcal{Y}, \Gamma \times \Lambda)$ (here Γ and Λ are Borel σ-algebras on $\mathcal{Y}$), with associated marginal measures $\mathbf{P}_{y_u}$ and $\mathbf{P}_{y_v}$. The covariance operator $C_{uv} : \mathcal{G} \to \mathcal{F}$ is defined by Fukumizu et al. (2004) such that for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$,

$$\langle f, C_{uv} g \rangle_{\mathcal{F}} = \mathbf{E}[f(y_u)g(y_v)] - \mathbf{E}[f(y_u)]\mathbf{E}[g(y_v)].$$

The HS norm of the covariance operator C_{uv}, which we denote the Hilbert-Schmidt independence criterion (HSIC), is written (Gretton et al., 2005a)

$$\begin{aligned} h(\mathbf{P}_{y_u,y_v}) := \|C_{uv}\|^2_{\mathrm{HS}} \;=\; & \mathbf{E}_{\dagger,\ddagger}\left[k\left(y_u^\dagger, y_u^\ddagger\right) l\left(y_v^\dagger, y_v^\ddagger\right)\right] \\ & + \mathbf{E}_{\dagger,\ddagger}\left[k\left(y_u^\dagger, y_u^\ddagger\right)\right] \mathbf{E}_{\dagger,\ddagger}\left[l\left(y_v^\dagger, y_v^\ddagger\right)\right] \\ & - 2\mathbf{E}_\dagger\left[\mathbf{E}_\ddagger\left[k\left(y_u^\dagger, y_u^\ddagger\right)\right] \mathbf{E}_\ddagger\left[l\left(y_v^\dagger, y_v^\ddagger\right)\right]\right], \end{aligned}$$

where $(y_u^\dagger, y_v^\dagger) \sim \mathbf{P}_{y_u,y_v}$, $\mathbf{E}_\dagger$ is the expectation over these random variables, $(y_u^\ddagger, y_v^\ddagger) \sim \mathbf{P}_{y_u,y_v}$ are independent copies[4] of $(y_u^\dagger, y_v^\dagger)$, and $\mathbf{E}_{\dagger,\ddagger}$ is the expectation over both independent copies. We require $\mathcal{F}$ and $\mathcal{G}$ to be universal in the sense of Steinwart (2002), since under these conditions $h_{u,v}(\mathbf{P}_{y_u,y_v}) = 0$ if and only if y_u and y_v are independent (see Gretton et al., 2005a). The Gaussian and Laplace kernels are both universal on compact domains. In this work, we confine ourselves to a Gaussian kernel,[5] and use an identical kernel for both $\mathcal{F}$ and $\mathcal{G}$,

$$k(a, b) = l(a, b) = \exp(-\sigma^{-2} \|a - b\|^2 / 2).$$

The first derivative of this kernel is $k'(a, b) = (b - a)\sigma^{-2}k(a, b)$, and the second derivative is $k''(a, b) = (a - b)^2\sigma^{-4}k(a, b) - \sigma^{-2}k(a, b)$.

HSIC belongs to a family of kernel independence measures that differ in the way they summarise the covariance operator spectrum, and in the normalisation they use. The simplest alternative to HSIC is the spectral norm of the covariance operator, or COCO (Gretton et al., 2005b). The kernel canonical correlation (Bach and Jordan, 2002) is similar to COCO, but represents the regularised spectral norm of the functional *correlation* operator, rather than the covariance operator. The kernel generalised variance (Bach and Jordan, 2002) and kernel mutual information (Gretton et al., 2005b) were shown by Gretton et al. (2005b) to upper bound the mutual information near independence (and to be tight at independence). Experiments by Gretton et al. (2005a), however, indicate that these methods do not outperform HSIC in linear ICA for large sample sizes, at least on the benchmark data of Bach and Jordan (2002).

HSIC was used for ICA by Eriksson and Koivunen (2003); Gretton et al. (2005a), where the former solved for pairwise independence by searching over all Jacobi rotations that make up X, and the latter obtained pairwise independence via the same gradient descent method as Bach and Jordan (2002). Jegelka and Gretton (2007) further reduced the cost of the second approach using a series of local quadratic approximations, rather than a Golden search. A generalisation of HSIC to a measure of mutual independence for more than two variables was proposed by Kankainen (1995), and was applied to ICA by Chen and Bickel (2005), using gradient descent on $SO(m)$ with a Golden search.

For the purpose of multivariate ICA, however, it is sufficient to enforce pairwise independence, and we simply sum the pairwise criteria to get our multivariate criterion,

$$H(\mathbf{P}_y) := \sum_{u,v=1,u\neq v}^{m} h_{u,v}(\mathbf{P}_{y_u,y_v}). \tag{2}$$

The pairwise independence criterion can be restated as an explicit function of columns x_u of the unmixing matrix X,

$$\begin{aligned} & h_{uv} : SO(m) \to \mathbb{R} \\ h_{uv}(X) := \; & \mathbf{E}_{\dagger,\ddagger}[k(x_u^\top w^\dagger, x_u^\top w^\ddagger)k(x_v^\top w^\dagger, x_v^\top w^\ddagger)] \\ & + \mathbf{E}_{\dagger,\ddagger}[k(x_u^\top w^\dagger, x_u^\top w^\ddagger)]\mathbf{E}_{\dagger,\ddagger}[k(x_v^\top w^\dagger, x_v^\top w^\ddagger)] \\ & - 2\mathbf{E}_\dagger[\mathbf{E}_\ddagger[k(x_u^\top w^\dagger, x_u^\top w^\ddagger)]\mathbf{E}_\ddagger[k(x_v^\top w^\dagger, x_v^\top w^\ddagger)]], \end{aligned}$$

where $y_u := (Xe_u)^\top w = x_u^\top w$ is the uth estimated source, $w^\dagger$ and $w^\ddagger$ are independent copies of the random variable w, and $\mathbf{E}_\dagger, \mathbf{E}_\ddagger, \mathbf{E}_{\dagger,\ddagger}$ denote expectations over the first, second, or both copies, respectively. An important advantage of the pairwise criterion is its diagonal Hessian (with respect to the parameter space of X; see below) at independence, which we use in Section 3.2 to obtain an inexpensive approximate Newton method. Kankainen's multivariate generalisation does not have this property.

3 Derivation of Newton-ICA

We now present an approximate Newton-like algorithm for optimisation of HSIC on the special orthog-

[4] That is, random variables drawn independently according to the same law.

[5] Gretton et al. (2005a,b) also used the Laplace kernel for ICA: this gives slightly better performance, but requires a higher rank in the incomplete Choleksy approximations used when obtaining HSIC and its derivatives (see the end of Section 3.2), and hence is slower. Non-universal kernels, such as polynomial kernels, could also be used, if certain properties of the source distributions are assumed (Bach and Jordan, 2002).

onal group, following Hüper and Trumpf (2004). We first give definitions of our terms and variables, followed by expressions for the gradient and Hessian of HSIC in the parameter space of $SO(m)$. We demonstrate that HSIC has a non-degenerate critical point at independence, and that the Hessian at this point is diagonal. This leads us to an approximate Newton method. Finally, we show the gradient and approximate Hessian may be efficiently estimated using the incomplete Cholesky decomposition.

We begin with some definitions and terminology. Let $SO(m) := \{X \in \mathbb{R}^{m\times m} | X^\top X = I, \det(X) = 1\}$ be the special orthogonal group. A local parametrisation of $SO(m)$ around a point $X \in SO(m)$ is

$$\alpha_X : \mathfrak{so}(m) \to SO(m), \qquad \Omega \mapsto X \exp(\Omega),$$

where $\mathfrak{so}(m) := \{\Omega \in \mathbb{R}^{m\times m} | \Omega = -\Omega^\top\}$ is the set of skew-symmetric matrices. We denote by $z \in \mathbb{R}^{m(m-1)/2}$ the vector of unique entries in Ω, i.e. $\omega_{u,v}$ for $v > u$. Without any ambiguity, we might likewise use α_X to denote the mapping $\mathbb{R}^{m(m-1)/2} \to SO(m)$. The tangent space of $SO(m)$ at a point X is

$$T_X SO(m) := \{\Xi \in \mathbb{R}^{m\times m} | \Xi = X\Omega, \Omega \in \mathfrak{so}(m)\},$$

where $\Xi = [\xi_1, \ldots, \xi_m] \in T_X SO(m)$ with $\xi_u := \Xi e_u$. Here e_u is the u-th standard basis vector in $\mathbb{R}^{m\times m}$. Moreover, by means of the matrix exponential map, a geodesic emanating from a point $X \in SO(m)$ is defined as

$$\gamma_X : \mathbb{R} \to SO(m), \qquad \varepsilon \mapsto X \exp(\varepsilon\Omega),$$

such that $\gamma_X(0) = X$, and $\dot\gamma_X(0) = \Xi \in T_X SO(m)$.

We now outline the Newton-like method for optimisation of HSIC on $SO(m)$. Let $\nabla(H \circ \alpha_X)(0)$ and $\mathcal{H}(H \circ \alpha_X)(0)$ be the gradient and Hessian of the smooth composition $H \circ \alpha_X$ at $0 \in \mathfrak{so}(m)$, respectively (these are simply computed with respect to the standard Euclidean inner product in the parameter space $\mathbb{R}^{m(m-1)/2}$). A Newton-like method for optimisation of HSIC on $SO(m)$ can then be summarised as follows:

Newton-like method for optimising HSIC on $SO(m)$
Step 1: Given an initial guess $X \in SO(m)$;
Step 2: Compute the Euclidean direction $z := -\mathcal{H}^{-1}(H \circ \alpha_X)(0)\nabla(H \circ \alpha_X)(0)$
Step 3: Update X by setting $X = \alpha_X(z)$;
Step 4: Go to step 2 until convergence.

In the following sections, we give explicit expressions for the derivative and Hessian, as well as an approximation to the Hessian that is exact at independence.

3.1 Gradient and critical point analysis of HSIC on $SO(m)$

We now obtain the gradient of the multivariate independence criterion in (2), and show that this quantity has a critical point when the sources are correctly unmixed. Since the first derivative of H is a sum of the pairwise derivatives, we begin by obtaining the first derivative of a single $h_{u,v}$. By the chain rule, the first derivative of the composition $h_{u,v} \circ \gamma_X$ of HSIC is

$$\begin{aligned}
&\tfrac{\mathrm{d}}{\mathrm{d}\varepsilon}(h_{u,v} \circ \gamma_X)(\varepsilon)\big|_{\varepsilon=0} \\
&= \mathbf{E}_{\dagger,\ddagger}\left[k'\left(y_u^\dagger, y_u^\ddagger\right)\xi_u^\top(w^\dagger - w^\ddagger)k\left(y_v^\dagger, y_v^\ddagger\right)\right] \\
&\quad + \mathbf{E}_{\dagger,\ddagger}\left[k'\left(y_u^\dagger, y_u^\ddagger\right)\xi_u^\top(w^\dagger - w^\ddagger)\right]\mathbf{E}_{\dagger,\ddagger}\left[k\left(y_v^\dagger, y_v^\ddagger\right)\right] \\
&\quad - 2\mathbf{E}_\dagger\left[\mathbf{E}_\ddagger\left[k'\left(y_u^\dagger, y_u^\ddagger\right)\xi_u^\top(w^\dagger - w^\ddagger)\right]\mathbf{E}_\ddagger\left[k\left(y_v^\dagger, y_v^\ddagger\right)\right]\right] \\
&\quad + \cdots,
\end{aligned}$$

where for brevity we omit the terms with respect to $\xi_v = \Xi e_v$ (which are obtained by interchanging u and v in the above). We next use this quantity to obtain the derivative of (2) in terms of the parameter space representation Ω at X. This is

$$\tfrac{\mathrm{d}}{\mathrm{d}\varepsilon}(H \circ \alpha_X)(\varepsilon\Omega)\big|_{\varepsilon=0} = \sum_{u,v=1,u<v}^{m} \omega_{uv} \cdot (\phi_{uv} - \phi_{vu}), \quad (3)$$

where

$$\begin{aligned}
\phi_{uv} =& \mathbf{E}_{\dagger,\ddagger}\left[k'\left(y_u^\dagger, y_u^\ddagger\right)(y_u^\dagger - y_u^\ddagger)k\left(y_v^\dagger, y_v^\ddagger\right)\right] \\
&+ \mathbf{E}_{\dagger,\ddagger}\left[k'\left(y_u^\dagger, y_u^\ddagger\right)(y_u^\dagger - y_u^\ddagger)\right]\mathbf{E}_{\dagger,\ddagger}\left[k\left(y_v^\dagger, y_v^\ddagger\right)\right] \\
&- 2\mathbf{E}_\dagger\left[\mathbf{E}_\ddagger\left[k'\left(y_u^\dagger, y_u^\ddagger\right)(y_u^\dagger - y_u^\ddagger)\right]\mathbf{E}_\ddagger\left[k\left(y_v^\dagger, y_v^\ddagger\right)\right]\right].
\end{aligned}$$

To arrive at this form, we replace $\Xi = X\Omega$, and simplify further using that the diagonal entries of Ω are zero, and that many of the remaining terms vanish when we take expectations over the independent copies.

Let us now consider critical points of H, at which the derivative is zero. While we do not attempt to obtain all the critical points (which are a function of the source distributions), we need to ensure a global optimum X^* is a critical point of H (in the next section, we see this point is nondegenerate). At a global optimum X^*, we know[6] $y_u^\dagger = s_u^\dagger$, $y_v^\dagger = s_v^\dagger$, and $s_u^\dagger, s_v^\dagger$ are independent. We are left with

$$\begin{aligned}
&\tfrac{\mathrm{d}}{\mathrm{d}\varepsilon}(h_{u,v} \circ \gamma_{X^*})(\varepsilon)\big|_{\varepsilon=0} \\
&= \omega_{uv}\mathbf{E}_{\dagger,\ddagger}\left[k'\left(s_u^\dagger, s_u^\ddagger\right)\right]\mathbf{E}_{\dagger,\ddagger}\left[(s_v^\dagger - s_v^\ddagger)k\left(s_v^\dagger, s_v^\ddagger\right)\right] \\
&\quad - 2\omega_{uv}\mathbf{E}_{\dagger,\ddagger}\left[k'\left(s_u^\dagger, s_u^\ddagger\right)\right]\mathbf{E}_\dagger\left[(s_v^\dagger - s_v^\ddagger)\mathbf{E}_\ddagger\left[k\left(s_v^\dagger, s_v^\ddagger\right)\right]\right] \\
&\quad + \cdots,
\end{aligned}$$

again omitting the symmetric terms in ξ_v. Using the symmetry of the kernel, we have $\mathbf{E}_{\dagger,\ddagger}\left[k'\left(s_u^\dagger, s_u^\ddagger\right)\right] = 0$, and thus H has a critical point at X^*.

[6] We ignore the issue of permutation and scaling of the unmixed sources, which does not change the analysis.

3.2 The Hessian and its approximation near independence

In this section, we show that, at a critical point X^* corresponding to correctly recovered sources, the Hessian of the composition $H \circ \alpha_X$ at $0 \in \mathfrak{so}(m)$ is diagonal. In other words, an unmixing matrix X^* is a nondegenerate critical point of H. We approximate the Hessian by retaining only those terms that do not vanish at independence, which, due to the smoothness of HSIC, yields an accurate and easily invertible approximation of the Hessian near independence. Consequently, we observe in our experiments that a Newton-like approach using the approximate Hessian converges rapidly once the solution is close to independence.

By computing the second derivative of the pairwise term $h_{u,v}$ at $X \in SO(m)$, we get

$$
\begin{aligned}
&\tfrac{\mathrm{d}^2}{\mathrm{d}\,\varepsilon^2}(h_{u,v} \circ \gamma_X)(\varepsilon)\Big|_{\varepsilon=0} \\
&= \mathbf{E}_{\dagger,\ddagger}[k''(y_u^\dagger, y_u^\ddagger)\xi_u^\top (w^\dagger - w^\ddagger)(w^\dagger - w^\ddagger)^\top \xi_u k(y_v^\dagger, y_v^\ddagger)] \\
&- \mathbf{E}_{\dagger,\ddagger}[k'(y_u^\dagger, y_u^\ddagger)\xi_u^\top \Xi X^\top (w^\dagger - w^\ddagger) k(y_v^\dagger, y_v^\ddagger)] \\
&+ \mathbf{E}_{\dagger,\ddagger}[k'(y_u^\dagger, y_u^\ddagger)\xi_u^\top (w^\dagger - w^\ddagger)(w^\dagger - w^\ddagger)^\top \xi_v k'(y_v^\dagger, y_v^\ddagger)] \\
&+ \mathbf{E}_{\dagger,\ddagger}[k''(y_u^\dagger, y_u^\ddagger)\xi_u^\top (w^\dagger - w^\ddagger)(w^\dagger - w^\ddagger)^\top \xi_u] \\
&\qquad \mathbf{E}_{\dagger,\ddagger}[k(y_v^\dagger, y_v^\ddagger)] \\
&- \mathbf{E}_{\dagger,\ddagger}[k'(y_u^\dagger, y_u^\ddagger)\xi_u^\top \Xi X^\top (w^\dagger - w^\ddagger)]\mathbf{E}_{\dagger,\ddagger}[k(y_v^\dagger, y_v^\ddagger)] \\
&+ \mathbf{E}_{\dagger,\ddagger}[k'(y_u^\dagger, y_u^\ddagger)\xi_u^\top (w^\dagger - w^\ddagger)]\mathbf{E}_{\dagger,\ddagger}[k'(y_v^\dagger, y_v^\ddagger)\xi_v^\top (w^\dagger - w^\ddagger)] \\
&- 2\mathbf{E}_\dagger\left[\mathbf{E}_\ddagger[k''(y_u^\dagger, y_u^\ddagger)\xi_u^\top (w^\dagger - w^\ddagger)(w^\dagger - w^\ddagger)^\top \xi_u]\right. \\
&\qquad \left.\mathbf{E}_\ddagger[k(y_v^\dagger, y_v^\ddagger)]\right] \\
&+ 2\mathbf{E}_\dagger\left[\mathbf{E}_\ddagger[k'(y_u^\dagger, y_u^\ddagger)\xi_u^\top \Xi X^\top (w^\dagger - w^\ddagger)]\mathbf{E}_\ddagger[k(y_v^\dagger, y_v^\ddagger)]\right] \\
&- 2\mathbf{E}_\dagger\left[\mathbf{E}_\ddagger[k'(y_u^\dagger, y_u^\ddagger)\xi_u^\top (w^\dagger - w^\ddagger)]\right. \\
&\qquad \left.\mathbf{E}_\ddagger[k'(y_v^\dagger, y_v^\ddagger)\xi_v^\top (w^\dagger - w^\ddagger)]\right] + \cdots,
\end{aligned}
$$

omitting the terms for ξ_v as above. Let $X = X^*$ and recall the structure of $T_X SO(m)$. Following a tremendous amount of fairly straightforward algebra, many of the terms vanish. By substituting the appropriate derivatives for the Gaussian kernel, the second derivative of $H \circ \alpha_{X^*}$ in terms of the parameter space representation $\Omega \in \mathfrak{so}(m)$ at 0 is

$$
\tfrac{\mathrm{d}^2}{\mathrm{d}\,\varepsilon^2}(H \circ \alpha_{X^*})(\varepsilon\Omega)\Big|_{\varepsilon=0} = \sum_{u,v=1,u<v}^{m} \omega_{uv}^2 \psi_{uv}, \qquad (4)
$$

where

$$
\begin{aligned}
\psi_{uv} =& \frac{2}{\sigma^2} m_1(u) m_2(v) + \frac{2}{\sigma^2} m_2(u) m_1(v) \\
&+ \frac{4}{\sigma^4} m_2(u) m_2(v) - \frac{4}{\sigma^4} m_3(u) m_3(v)
\end{aligned}
$$

and

$$
\begin{aligned}
m_1(u) &:= \mathbf{E}_{\dagger\ddagger}[k(s_u^\dagger, s_u^\ddagger)] \\
m_2(u) &:= \mathbf{E}_{\dagger\ddagger}[k(s_u^\dagger, s_u^\ddagger) s_u^\dagger s_u^\ddagger] \\
m_3(u) &:= \mathbf{E}_{\dagger\ddagger}[k(s_u^\dagger, s_u^\ddagger)(s_u^\dagger)^2].
\end{aligned}
$$

Thus, the Hessian of $H \circ \alpha_{X^*}$ at 0 is diagonal in terms of ω_{uv}. Since HSIC has global minima at independence (Gretton et al., 2005b), and by the smoothness of HSIC, we can easily show that every entry ψ_{uv} of the Hessian must be positive. The simple structure of the Hessian of $H \circ \alpha_{X^*}$ at $0 \in \mathfrak{so}(m)$ leads us to use this expression to approximate the Hessian $\mathcal{H}(H \circ \alpha_X)(0)$ at all X, by replacing the true recovered sources $s_u^\dagger, s_u^\ddagger, s_v^\dagger, s_v^\ddagger$ with their current estimates $y_u^\dagger, y_u^\ddagger, y_v^\dagger, y_v^\ddagger$. Since the approximate Hessian is diagonal, the inversion required for the Newton step can then be done very cheaply. If we initialise close to independence, this approximation is accurate, since the missing terms are negligible. Remarkably, when we initialise far from the true solution, this approximation appears to make us *less* likely to be diverted to local minima than using gradient descent, since these minima will generally not have the Hessian structure specific to independence. These behaviours are illustrated in our experimental results in the next section.

We now outline how we use the incomplete Cholesky decomposition (Fine and Scheinberg, 2001) to estimate the approximate Hessian efficiently. Let K be the Gram matrix corresponding to the uth recovered source. An incomplete Cholesky G of size $n \times d$, where $d \ll n$, can be computed in time $O(nd^2)$, yielding an approximation $K = GG^\top$ that greedily minimises $\mathrm{tr}(K - GG^\top)$. Greater values of d result in a more accurate reconstruction of K, however, as Bach and Jordan (2002, Appendix C) point out, the spectrum of a Gram matrix based on the Gaussian kernel generally decays rapidly, and a small d yields a very good approximation. Using this estimate, the three terms in the Hessian are written

$$
\begin{aligned}
m_1(u) &= \frac{1}{n^2}(\mathbf{1}^\top G)(\mathbf{1}^\top G)^\top \\
m_2(u) &= \frac{1}{n^2}(s_u^\top G)(s_u^\top G)^\top \\
m_3(u) &= \frac{1}{n^2}((s_u \odot s_u)^\top G)(G^\top \mathbf{1}),
\end{aligned}
$$

where $\odot$ is the entry-wise product of the vectors of samples. Similar reasoning is used for the easier problem of cheaply approximating the gradient using the incomplete Cholesky decomposition.

We end this section with a brief note on the overall computational cost of FastKICA. As discussed by Jegelka and Gretton (2007, Section 1.5), the gradient and Hessian are computable in $O(nm^3d^2)$. A more detailed breakdown of how we arrive at this cost may be

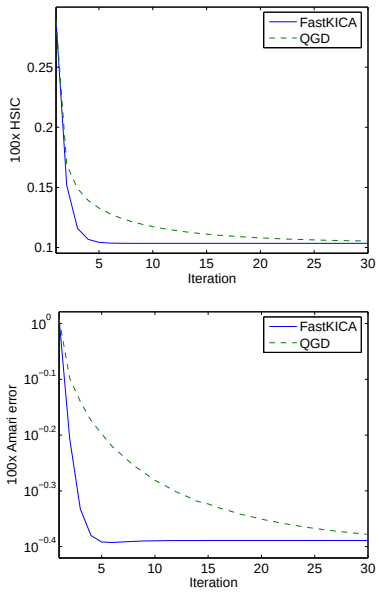


Figure 1: Convergence of HSIC (top) and the Amari error (bottom) for FastKICA and QGD. The plots show averages for 21 runs, $n = 40{,}000$, $m = 16$, with an initialisation by FastICA. FastKICA converges faster.

found in (Jegelka and Gretton, 2007), bearing in mind that the Hessian has the same cost as the gradient thanks to our diagonal approximation.

4 Experiments

In our experiments, we demonstrate three main points: First, if no alternative algorithm is used to provide an initial estimate of X, FastKICA demonstrates resistance to local minima, and often converges to the correct solution. This is by contrast with gradient descent, which is more often sidetracked to local minima. In particular, if we choose sources incompatible with the initialising algorithm (so that it fails completely), our method can nonetheless find a good solution.[7] Second, when a good initial point is given, the Newton-like algorithm converges faster than gradient descent. Third, our approach is much faster than RADICAL and MILCA, and runs sufficiently quickly on large-scale problems to be used either as a standalone method (when a good initialisation is impossible or unlikely), or to fine tune the solution obtained by another method. Demixing performance of FastKICA was superior to the other methods we tested.

Our artificial data are generated in accordance with

[7]Note the criterion optimised by FastKICA is also the statistic of an independence test (Feuerverger, 1993). This test can be applied directly to the values of HSIC between pairs of unmixed sources, to verify the recovered signals are truly independent; no separate hypothesis test is required.

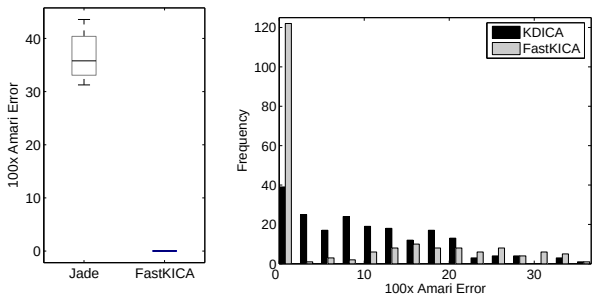


Figure 2: Comparison for arbitrary initialisations ($n = 40{,}000$, $m = 8$). Left: Data with near-zero kurtosis. Jade fails, but FastKICA succeeds in demixing the data. For FastKICA, the result with the lowest HSIC was chosen for each run (5 runs, 10 initialisations each). Right: Amari error histogram of FastKICA vs. KDICA for mixed artificial sources (10 data sets, 20 initialisations each). FastKICA reaches a global minimum far more often than KDICA.

Gretton et al. (2005b, Table 3), which is similar to the artificial benchmark data of Bach and Jordan (2002). Each source is chosen randomly with replacement from 18 different distributions having a wide variety of statistical properties and kurtoses. Sources are mixed using a random matrix with condition number between one and two. We use the Amari divergence, defined by Amari et al. (1996), as an index of ICA algorithm performance (we multiply this quantity by 100 to make the performance figures more readable). In all experiments, the precision of the incomplete Cholesky decomposition is $10^{-6}n$. Convergence is measured by the difference in HSIC values over consecutive iterations.

4.1 Comparison with gradient descent

We first compare the convergence of FastKICA with a simple gradient descent method (Jegelka and Gretton, 2007). In order to find a suitable step width along the gradient mapped to $SO(m)$, the latter uses a quadratic interpolation of HSIC along the geodesic. To this end HSIC needs to be evaluated at two additional points. Both FastKICA and quadratic gradient descent (QGD) use the same gradient and independence measure. Figure 1 compares the convergence of HSIC and the Amari error for both methods on the same data. The results are averages over 21 runs. In each run, 40,000 observations from 16 artificial, randomly drawn sources were generated and mixed. We initialise both methods with FastICA (Hyvärinen et al., 2001), and use a kernel width of $\sigma = 0.5$. As illustrated by the plots, FastKICA approaches the solution much faster than QGD. The numerical results in Table 1 confirm that FastKICA has converged after five iterations. We

also observe that the the number of iterations to convergence decreases when the sample size grows. For arbitrary initialisations, FastKICA is still applicable with multiple restarts, although a larger kernel width is more appropriate (local fluctuations in FastKICA far from independence are smoothed out, although the bias in the location of the global minimum increases). We set $\sigma = 1$ and a convergence threshold of 10^{-8} for both FastKICA and QGD. For 40,000 samples from 8 artificial sources, FastKICA converged on average for 37% of the random restarts with an average error ($\times$ 100) of 0.54 ± 0.01, whereas the QGD did not yield any useful results at all (mean error $\times$ 100: 74.14 ± 1.39). Here, averages are over 10 runs with 20 random initialisations each. The solution obtained with FastKICA can be refined further by shrinking the kernel width after initial convergence, to reduce the bias.

4.2 Poor initialising matrix and near-zero kurtosis

We generate sources with near-zero kurtosis using a mixture of Gaussians (see Gretton et al., 2005b, Section 5.4). These data cannot be separated by Jade (Cardoso, 1998), which uses the sum of the estimated source kurtoses as its independence measure (FastICA has also been shown by Gretton et al., 2005b, to perform less well than recent nonparametric methods). FastKICA, however, recovers the sources even with arbitrary initialisations (of course, a good initial guess should be used if available, since the optimum is then reached faster). On average over 5 runs with 10 restarts ($n = 40{,}000$, $m = 8$), FastKICA converged with an Amari error of 0.20 ± 0.01 (convergence threshold: 10^{-7}, $\sigma = 1.0$), whereas the error for Jade averages to 36.73. Figure 2 illustrates these findings. For FastKICA, averages were only taken over runs that converged. The example of near-zero kurtosis underlines the advantage of kernel methods, where the dependence measure is provably zero if and only if the signals are independent, as opposed to criteria based on assumed statistical properties of the sources.

4.3 Performance and cost vs other approaches

We conclude by comparing the performance and computational cost of FastKICA with Jade, KDICA (Chen, 2006), MILCA (Stögbauer et al., 2004), RADICAL (Learned-Miller and Fisher III, 2003), and quadratic gradient descent (QGD). All timing experiments were performed on 64 bit Opteron CPUs, running Debian/GNU Linux 3.1. We use 8 sources and 40,000 observations of the artificial data. The run times include the initialisation by Jade for FastKICA, QGD, and KDICA. QGD was run for 10 iterations, and the convergence threshold for FastKICA was 10^{-5}. Figure 3 displays the error and time for 25 data sets. Both FastKICA and QGD have lower mean and median errors than the other methods. The hypothesis that FastKICA does not have a lower mean error than MILCA, RADICAL, or KDICA can be rejected at the 5% level, using a left-tailed t-test. In addition, both QGD and FastKICA are faster than RADICAL and MILCA. The additional evaluations of HSIC for the quadratic approximation make QGD slower per iteration than FastKICA. As shown above, FastKICA also converges in fewer iterations than QGD, requiring 4.32 iterations on average. Apart from Jade, only KDICA is faster than FastKICA, although its performance (mean and median) is a little worse, and displays higher variance. We also compared KDICA and FastKICA when random initialisations are used. We see in Figure 2 that FastKICA solutions have a clear bimodal distribution, with a large number of initialisations reaching an identical global minimum: indeed, the correct solution is clearly distinguishable from local optima on the basis of its HSIC value. By contrast, KDICA appears to halt at a much wider variety of local minima, as evidenced by the broad range of Amari errors in the estimated unmixing matrices. Thus, in the absence of a good initialising estimate (where classical methods fail), FastKICA is to be preferred. Finally, KDICA can use only the Laplace kernel, whereas FastKICA is applicable with a range of kernels, so we can select kernels appropriate to the source distributions, while still benefiting from our optimisation and approximation techniques.

5 Conclusions

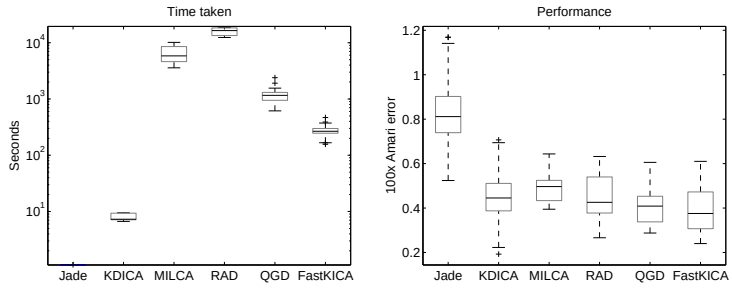
We demonstrate that an approximate Newton-like method, FastKICA, can improve the speed and performance of kernel ICA methods. We emphasise that FastKICA is applicable even if no good initialisation is at hand. With a modest number of restarts and a kernel width that shrinks near independence (on our data, from $\sigma = 1.0$ to $\sigma = 0.5$), the correct global optimum is consistently found. A good initialisation results in more rapid convergence, and we do not need to adapt the kernel size. It is certainly possible that the method of Chen (2006) would likewise benefit from an approximate Newton approach, although we would need to demonstrate that the Hessian behaves well at independence. This is an area of current research.

Acknowledgements

We would like to thank Aiyou Chen for providing us with his KDICA code. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

		5 iterations		10 iterations		20 iterations	
	init. AE	HSIC	AE	HSIC	AE	HSIC	AE
FastKICA	0.90	0.11 ± 0.002	0.39 ± 0.01	0.11 ± 0.002	0.39 ± 0.01	0.11 ± 0.002	0.39 ± 0.01
QGD	0.90	0.13 ± 0.004	0.58 ± 0.03	0.12 ± 0.002	0.50 ± 0.02	0.11 ± 0.002	0.42 ± 0.01

Table 1: Average Amari error (AE) and HSIC after 5, 10 and 20 iterations of FastKICA or the gradient descent with quadratic approximation. Note that both the Amari error and HSIC are multiplied by 100. The data are the same as in Figure 1.



	Error $\times 100$	Time [s]
Jade	0.82 ± 0.03	1.13±0.0005
KDICA	0.45 ± 0.03	8.1±0.2
MILCA	0.49 ± 0.01	6500±400
RAD	0.44 ± 0.02	15900±500
QGD	0.41 ± 0.02	1210±70
FastKICA	0.39 ± 0.02	270±10

Figure 3: Comparison of run times (left) and performance (middle) for various ICA algorithms. FastKICA is faster than MILCA, RADICAL, and gradient descent with quadratic approximation, and its results compare favorably to the other methods. KDICA is even faster, but performs less well than FastKICA. The table displays mean values over 25 data sets.

References

S.-I. Amari, A. Cichoki, and Yang H. A new learning algorithm for blind signal separation. In *NIPS*, volume 8, pages 757–763. MIT Press, 1996.

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *JMLR*, 3:1–48, 2002.

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

A. Chen. Fast kernel density independent component analysis. In *ICA*, volume 6, pages 24 – 31. Springer, 2006.

A. Chen and P. Bickel. Consistent independent component analysis and prewhitening. *IEEE Transactions on Signal Processing*, 53(10):3625–3632, 2005.

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

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.

J. Eriksson and V. Koivunen. Characteristic-function-based independent component analysis. *Signal Processing*, 83(10):2195 – 2208, 2003.

Andrey Feuerverger. A consistent test for bivariate dependence. *Int. Stat. Rev.*, 61(3):419–433, 1993.

S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *JMLR*, 2:243–264, 2001.

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

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT*, pages 63–78, 2005a.

A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *JMLR*, 6:2075–2129, 2005b.

K. Hüper and J. Trumpf. Newton-like methods for numerical optimisation on manifolds. In *Proceedings of Thirty-eighth Asilomar Conference on Signals, Systems and Computers*, pages 136–139, 2004.

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

S. Jegelka and A. Gretton. Brisk kernel ICA. In L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, editors, *Large Scale Kernel Machines*. MIT Press, 2007. To appear.

A. Kankainen. *Consistent Testing of Total Independence Based on the Empirical Characteristic Function*. PhD thesis, University of Jyväskylä, 1995.

C.-J. Ku and T. Fine. Testing for stochastic independence: application to blind source separation. *IEEE Transactions on Signal Processing*, 53(5):1815 – 1826, 2005.

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

N. Murata. Properties of the empirical characteristic function and its application to testing for independence. In *ICA*, volume 3, pages 19–24, 2001.

H. Shen and K. Hüper. Local convergence analysis of FastICA. In *ICA*, volume 6, pages 893–900. Springer, 2006.

H. Shen, K. Hüper, and A. Smola. Newton-like methods for nonparametric independent component analysis. In *International Conference on Neural Information Processing*, 2006. to appear.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *JMLR*, 2:67–93, 2002.

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