Newton-Like Methods for Nonparametric Independent Component Analysis

Hao Shen[1,3], Knut Hüper[1,3], and Alexander J. Smola[2,4]

[1] Systems Engineering and Complex Systems Research Program
[2] Statistical Machine Learning Research Program
National ICT Australia, Canberra ACT 2612 Australia
[3] Department of Information Engineering
[4] Computer Science Laboratory
Research School of Information Sciences and Engineering
The Australian National University, Canberra ACT 0200 Australia
Hao.Shen@rsise.anu.edu.au, {Knut.Hueper, Alex.Smola}@nicta.com.au

Abstract. The performance of ICA algorithms significantly depends on the choice of the contrast function and the optimisation algorithm used in obtaining the demixing matrix. In this paper we focus on the standard linear nonparametric ICA problem from an optimisation point of view. It is well known that after a pre-whitening process, the problem can be solved via an optimisation approach on a suitable manifold. We propose an approximate Newton's method on the unit sphere to solve the one-unit linear nonparametric ICA problem. The local convergence properties are discussed. The performance of the proposed algorithms is investigated by numerical experiments.

1 Introduction

Since the success of Independent Component Analysis (ICA) for solving the Blind Source Separation (BSS) problem, ICA has received considerable attention in numerous areas, such as signal processing, statistical modeling, and unsupervised learning. The applications of ICA include image processing, bioinformatics, brain computer interfaces, dimensionality reduction, etc.

The performance of ICA algorithms depends significantly on the choice of the contrast function measuring statistical independence of signals and on the appropriate optimisation technique. There exist numerous parametric and nonparametric approaches to designing ICA contrast functions. For the parametric approaches, contrast functions are selected according to certain hypothetical distributions (probability density functions) of the sources by a single fixed nonlinear function. In practical applications, however, the distribtions of the sources are unkown, and even can not be approximated by a single fixed function. It has been shown that the nonparametric counterparts are more robust in practical applications, regardless of the distributions of signals [1]. The focus of this paper is on the development of *optimisation algorithms* for nonparametric ICA contrast functions.

I. King et al. (Eds.): ICONIP 2006, Part I, LNCS 4232, pp. 1068–1077, 2006.

Since the influential paper of Comon [2], many efficient ICA algorithms have been developed by researchers from various communities. Recently, there has been an increasing interest in using geometric optimisation for ICA problems. The FastICA algorithm is a prominent parametric ICA algorithm proposed by the Finnish school, see [3]. It enjoys an interpretation of an approximate Newton's method in $\mathbb{R}^m$ with both good accuracy and fast speed of convergence. While its contrast function is not among the best [4], it proves highly attractive due to its efficient and simple implementation.

In this paper, we develop an approximate Newton's method on the unit sphere for optimising nonparametric one-unit ICA contrast functions. That is, we study algorithms which find solutions to the ICA problem one component at a time. The spirit of our approach originates in the idea developed already for FastICA. The local convergence properties of the proposed algorithm are also investigated subject to an ideal ICA model.

The paper is organised as follows. In Section 2, we review the linear ICA problem. Section 3 briefly discusses the fundamentals of Newton's method on manifolds. An approximate Newton's method on the unit sphere for one-unit nonparametric ICA is proposed in Section 4. Numerical experiments in Section 5 demonstrate the performance of the proposed algorithms compared with several existing linear ICA algorithms.

2 Linear ICA Model

2.1 Linear Mixtures

We consider the standard noiseless linear instantaneous ICA model

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

where $S \in \mathbb{R}^{m\times n}$ represents n samples of m sources with $m \ll n$ which are drawn independently identically distributed (i.i.d) from $\mathbb{R}^m$. The invertible matrix $A \in \mathbb{R}^{m\times m}$ is the mixing matrix and $M \in \mathbb{R}^{m\times n}$ contains the observed mixtures. We assume that the source signals S (while unknown) have zero mean and unit variance. Moreover we assume that they are statistically independent, and at most one being Gaussian. In the rest of this paper, we call the above model an *ideal ICA* model.

The task of ICA is to find a linear transformation $B \in \mathbb{R}^{m\times m}$ of the observations M to recover the sources by

$$Y = BM = BAS, \tag{2}$$

where $Y \in \mathbb{R}^{m\times n}$ is the estimation of sources. It is well known that the sources can be extracted up to an arbitrary order and certain scaling, i.e., B is the inverse of A up to an $m \times m$ permutation matrix P and an $m \times m$ diagonal(scaling) matrix D [2]

$$B = PDA^{-1}. \tag{3}$$

According to the assumption of independence between sources, the task of ICA can be naturally interpreted as to find B to estimate the sources as independent as possible. In order to fulfil this task, a certain contrast function is required to be minimised as a measure of statistical independence over estimated signals. Generally speaking, such measures can be evaluated via both parametric and nonparametric approaches, which are given later in this section.

To reduce the computational complexity of performing ICA, it has been shown that a pre-whitening process can be used without loss of statistical consistency [5]. Hence the whitened demixing ICA model can be formulated as

$$Z = X^\top W, \tag{4}$$

where $W = VM \in \mathbb{R}^{m\times n}$ is the whitened observation, the invertible matrix $V \in \mathbb{R}^{m\times m}$ is the whitening matrix, the orthogonal matrix $X \in \mathbb{R}^{m\times m}$, i.e., $X^\top X = I$, is a new parameterisation of the problem as the demixing matrix in the whitened model, and $Z \in \mathbb{R}^{m\times n}$ is the recovered signal.

In this paper, we only focus the problem of obtaining the columns of X sequentially (which we will refer to as the one-unit ICA problem). For a parallel approach, see [6,7].

2.2 Contrast Functions

A large class of contrast functions can be stated as follows: let $X = [x_1, \ldots, x_m]$ be an orthogonal matrix with column $x_i \in S^{m-1} := \{x \in \mathbb{R}^m | \|x\| = 1\}$ for all $i = 1, \ldots, m$ and w_i the i-th column of W, respectively. A generic parametric one-unit contrast function can be defined as

$$g : S^{m-1} \to \mathbb{R}, \quad x \mapsto \mathbb{E}_w^{emp}[G(x^\top w)], \tag{5}$$

where $G : \mathbb{R} \to \mathbb{R}$ is differentiable and $\mathbb{E}_w^{emp}[\cdot]$ denotes the empirical mean, i.e., $\mathbb{E}_w^{emp}[f(w)] = \frac{1}{n}\sum_{i=1}^{n} f(w_i)$. Different functions G can be used to estimate certain statistical properties over the sampled data. For a standard parametric approach, one might use Kurtosis or higher order moments, e.g., $G(u) = u^4$, see [3]. We refer to [8,9] for more details about the linear parametric ICA algorithms.

A typical nonparametric contrast function using a kernel density estimation technique can also be generalised in the similar form as g in (5), see [1],

$$f : S^{m-1} \to \mathbb{R}, \qquad x \mapsto \mathbb{E}_i^{emp}\left[\log\left(\tfrac{1}{h}\mathbb{E}_j^{emp}\left[\phi\left(\tfrac{x^\top(w_i - w_j)}{h}\right)\right]\right)\right], \tag{6}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is an appropriate kernel function and $h \in \mathbb{R}^+$ is the kernel bandwidth. It is essentially a smoothed entropy estimate of a recovered signal, which utilises an appropriate entropy estimate following a Parzen density estimate. It is similar to a popular parametric ICA approach, Infomax [3]. In this paper we specify ϕ as a Gaussian kernel $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$.

Note the performances of such nonparametric contrast function (6) are out of scope of this paper. We only focus on the problem from an optimisation point of view. All computations in this paper are performed using coordinate functions of $\mathbb{R}^m$ which is the embedding space of the unit sphere S^{m-1}.

3 Newton's Method on the Unit Sphere

In this section we give an introduction to Newton's Method for solving optimisation problems defined on manifolds, specifically the unit sphere.

Given a smooth cost function $f : \mathbb{R}^m \to \mathbb{R}$ and denote by $\nabla f(x)$ and $\mathcal{H}f(x)$ the gradient and Hessian with respect to the standard Euclidean inner product. Let $x^* \in \mathbb{R}^m$ be a nondegenerate critical point of f, i.e., $\nabla f(x^*) = 0$ and $\mathcal{H}f(x^*)$ is nondegenerate. Newton's method for optimising f is the iteration

$$x_0 \in \mathbb{R}^n, \qquad x_{k+1} = x_k - (\mathcal{H}f(x_k))^{-1}\nabla f(x_k). \tag{7}$$

The Newton iteration describes a step moves along a straight line, in the so-called Newton direction

$$\mathcal{N}f(x_k) = -(\mathcal{H}f(x_k))^{-1}\nabla f(x_k). \tag{8}$$

It enjoys the significant property of local quadratic convergence to a nondegenerate critical point of f.

We now consider an optimisation problem defined on an m-dimensional manifold $\mathcal{M}$. By abuse of notation, let $f : \mathcal{M} \to \mathbb{R}$ be a smooth function on $\mathcal{M}$. A typical approach is to endow the manifold with a Riemannian structure. Rather than moving along a straight line, the Riemannian Newton iteration moves along a geodesic γ_x of $\mathcal{M}$, i.e.,

$$x_{k+1} = \gamma_{x_k}(\varepsilon)\,, \tag{9}$$

where $\gamma_{x_k} : \mathbb{R} \to \mathcal{M}$ is a geodesic emanating from $x_k \in \mathcal{M}$, such that $\gamma_{x_k}(0) = x_k$ and $\dot{\gamma}_{x_k}(0) = \mathcal{N}f(x_k)$ and $\varepsilon \in \mathbb{R}$ is a step size for current iteration. For the standard Newton's method, the step size is $\varepsilon = 1$.

Newton's method for optimising a function defined on a manifold can be summarised as follows

Newton's method on a manifold $\mathcal{M}$
Step 1: Given an initial guess $x \in \mathcal{M}$; Step 2: Compute the Newton direction $\mathcal{N}f(x) = -(\mathcal{H}f(x))^{-1}\nabla f(x)$; Step 3: Move from x in direction $\mathcal{N}f(x)$ and update x by setting $x = \gamma_x(1)$; Step 4: Goto step 2.

Note, by replacing the true Hessian $\mathcal{H}f(x)$ with an approximation $\widetilde{\mathcal{H}}f(x)$, an approximate Newton's method on a manifold $\mathcal{M}$ can be easily defined.

Before moving on to the next section, we give the definitions of the tangent space T_xS^{m-1} and geodesic mapping γ_x of S^{m-1} at $x \in S^{m-1}$

$$T_xS^{m-1} = \left\{ \xi \in \mathbb{R}^m \middle| \, x^\top \xi = 0 \right\}, \tag{10}$$

and

$$\gamma_x : \mathbb{R} \to S^{m-1}, \qquad \varepsilon \mapsto \exp\left(\varepsilon(\xi x^\top - x\xi^\top)\right) x, \tag{11}$$

where $\xi \in T_xS^{m-1}$ and $\exp(\cdot)$ is the matrix exponential.

4 Approximate Newton's Methods for One-Unit ICA

In this section, we develop an approximate Newton's method to solve the one-unit ICA problem via a nonparametric approach (6) on the unit sphere. For a Newton's method, the Hessian is often expensive to evaluate at each step. Therefore, for the nonparametric ICA problem, we propose a scalar approximation of the Hessian instead, to reduce the burden of computations. Using such an approximation we will show that it actually gives the true Hessian at a non-degenerate critical point. Moreover we will prove local quadratical convergence to a nondegenerate critical point.

To simplify calculations in our analysis, it is useful to find the right coordinate system. Without loss of generality, in the sequel, we assume $A = I$, i.e. $W = S$.

4.1 Derivation of Algorithm

By choosing a suitable coordinate system, we first show that the standard basis vectors $\{e_1, \ldots, e_m\}$, which correspond to the correct separations, are critical points of f as in (6). For the convenience of derivation, we set the kernel bandwidth $h = 1$.

To further simplify our analysis we define the following for future use

$$\begin{cases} \rho_i(x) := \mathbb{E}_j^{emp}\left[\phi\left(x^\top(s_i - s_j)\right)\right] & \in \mathbb{R}, \\ \rho_i'(x) := \mathbb{E}_j^{emp}\left[\phi'\left(x^\top(s_i - s_j)\right)(s_i - s_j)\right] & \in \mathbb{R}^m, \\ \widetilde{\rho}_i'(x) := \mathbb{E}_j^{emp}\left[\phi'\left(x^\top(s_i - s_j)\right)\right] & \in \mathbb{R}, \\ \rho_i''(x) := \mathbb{E}_j^{emp}\left[\phi''\left(x^\top(s_i - s_j)\right)(s_i - s_j)(s_i - s_j)^\top\right] & \in \mathbb{R}^{m\times m}, \\ \widetilde{\rho}_i''(x) := \mathbb{E}_j^{emp}\left[\phi''\left(x^\top(s_i - s_j)\right)\right] & \in \mathbb{R}, \end{cases} \tag{12}$$

where ϕ' and ϕ'' are the first and second derivatives of the kernel function ϕ.

Lemma 4.1. *Let $X = [x_1, x_2, \ldots, x_m]$ be an orthogonal matrix with column $x_k \in S^{m-1}$. $x_k \in S^{m-1}$ is a critical point of the function f defined by* (6) *if and only if, for some $\kappa \in \mathbb{R}$*

$$\mathbb{E}_i^{emp}\left[\frac{\rho_i'(x_k)}{\rho_i(x_k)}\right] = \kappa x_k. \tag{13}$$

Proof. By the chain rule, the first derivative of f is computed along a geodesic γ_x with $\gamma_x(0) = x$ and $\dot{\gamma}_x(0) = \xi \in T_x S^{m-1}$

$$\mathrm{D}\, f(x)\xi = \tfrac{\mathrm{d}}{\mathrm{d}\,\varepsilon}(f \circ \gamma_x)(\varepsilon)\big|_{\varepsilon=0} = \xi^\top \mathbb{E}_i^{emp}\left[\frac{\rho_i'(x)}{\rho_i(x)}\right]. \tag{14}$$

Recall a point $x \in \mathcal{M}$ is called a critical point of a smooth function $h : \mathcal{M} \to \mathbb{R}$ if the first derivative $\mathrm{D}\, h(x)\xi$ vanishes with $\xi \in T_x\mathcal{M}$. Thus critical points of f are characterised as solutions of

$$\xi^\top \mathbb{E}_i^{emp}\left[\frac{\rho_i'(x)}{\rho_i(x)}\right] = 0, \tag{15}$$

for all $\xi \in T_x S^{m-1}$. By the tangent space of the unit sphere (10), the result follows. □

By the prior assumption of independence between sources, it is easily seen that, for all $k = 1, \ldots, m$, the term $\rho_i'(e_k)$ as defined in (12) is a multiple of e_k, i.e., standard basis vectors $\{e_1, \ldots, e_m\}$, which correspond to the exact separations,

fulfil the above condition (13). With respect to the Riemannian metric on S^{m-1} induced by the Euclidean metric in $\mathbb{R}^m$

$$\langle p_1, p_2 \rangle = p_1^\top p_2, \tag{16}$$

the Riemannian gradient of f at $x \in S^{m-1}$ can be computed as

$$\nabla f(x) = \left(I - xx^\top\right) \mathbb{E}_i^{emp} \left[\frac{\rho_i'(x)}{\rho_i(x)}\right], \tag{17}$$

where the term $(I - xx^\top)$ is an orthogonal projector onto the complement of the span of x. In the next lemma, we characterise the structure of the Hessian of f at the critical points $\{e_1, \ldots, e_m\}$.

Lemma 4.2. *By choosing a suitable coordinate system, the Hessian operator at a critical point e_k acts on tangent vectors simply by scalar multiplication, the scalar being equal to*

$$-\mathbb{E}_i^{emp} \left[\left(\frac{\widetilde{\rho}_i'(e_k)}{\rho_i(e_k)}\right)^2\right] + 2\,\mathbb{E}_i^{emp} \left[\frac{\widetilde{\rho}_i''(e_k)}{\rho_i(e_k)}\right] - e_k^\top \mathbb{E}_i^{emp} \left[\frac{\rho_i'(e_k)}{\rho_i(e_k)}\right]. \tag{18}$$

Proof. By the chain rule, one computes the second derivative of f along a geodesic γ_x with $\gamma_x(0) = x$ and $\dot{\gamma}_x(0) = \xi \in T_x S^{m-1}$

$$\begin{aligned} &\mathrm{D}^2 f(x)(\xi, \xi) = \tfrac{\mathrm{d}^2}{\mathrm{d}\,\varepsilon^2}(f \circ \gamma_x)(\varepsilon)\Big|_{\varepsilon=0} \\ &= \xi^\top \left(-\mathbb{E}_i^{emp} \left[\frac{\rho_i'(x)(\rho_i'(x))^\top}{(\rho_i(x))^2}\right] + \mathbb{E}_i^{emp} \left[\frac{\rho_i''(x)}{\rho_i(x)}\right] - x^\top \mathbb{E}_i^{emp} \left[\frac{\rho_i'(x)}{\rho_i(x)}\right] I\right) \xi. \end{aligned} \tag{19}$$

By evaluating (19) at a critical point e_k, we get

$$\begin{aligned} &\mathrm{D}^2 f(x)(\xi, \xi)\big|_{x=e_k} = \tfrac{\mathrm{d}^2}{\mathrm{d}\,\varepsilon^2}(f \circ \gamma_x)(\varepsilon)\Big|_{\substack{\varepsilon=0 \\ x=e_k}} \\ &= \left(-\mathbb{E}_i^{emp} \left[\left(\frac{\widetilde{\rho}_i'(e_k)}{\rho_i(e_k)}\right)^2\right] + 2\,\mathbb{E}_i^{emp} \left[\frac{\widetilde{\rho}_i''(e_k)}{\rho_i(e_k)}\right] - e_k^\top \mathbb{E}_i^{emp} \left[\frac{\rho_i'(e_k)}{\rho_i(e_k)}\right]\right) \xi^\top \xi. \end{aligned} \tag{20}$$

Hence, the Hessian of f at e_k is a scalar matrix. The result follows. □

To ensure the inverse of the Hessian always exists at critical points e_k, in the rest of the paper, we might assume the scalar (18) does not vanish.

Due to the simplicity of the Hessian of f at a critical point e_k, it seems to be natural to approximate the Hessian of f at an *arbitrary point* $x \in S^{m-1}$ by a scalar as well, i.e.,

$$\tau(x) = -\mathbb{E}_i^{emp} \left[\left(\frac{\widetilde{\rho}_i'(x)}{\rho_i(x)}\right)^2\right] + 2\,\mathbb{E}_i^{emp} \left[\frac{\widetilde{\rho}_i''(x)}{\rho_i(x)}\right] - x^\top \mathbb{E}_i^{emp} \left[\frac{\rho_i'(x)}{\rho_i(x)}\right]. \tag{21}$$

It is worthwhile to notice that when evaluated at a critical point e_k, the above approximation gives the true Hessian. Thus an approximate Newton direction for optimising (6) can be formulated as

$$\widetilde{\mathcal{N}} f(x) = -\frac{\left(I - xx^\top\right) \mathbb{E}_i^{emp} \left[\frac{\rho_i'(x)}{\rho_i(x)}\right]}{\tau(x)}. \tag{22}$$

It is well known that computing the Newton iteration along geodesics (11) requires matrix exponentiation which is computationally expensive. Hence instead of using geodesics, we specify the following curve on S^{m-1} through $x \in S^{m-1}$, which reduces the computational burdens,

$$\mu_x : (-t, t) \to S^{m-1}, \quad \varepsilon \mapsto \frac{x + \varepsilon\xi}{\|x + \varepsilon\xi\|}, \tag{23}$$

with $\xi \in T_x S^{m-1}$ arbitrary. Such Newton-type iterations along μ_x still preserve the local quadratic convergence property to a nondegenerate critical point [10].

Finally by substituting (22) into (23), i.e., $\xi = \widetilde{\mathcal{N}} f(x)$, an approximate Newton's method for the parametric one-unit ICA is stated as follows

The Approximate Newton Nonparametric ICA algorithm (ANNICA)
Step 1: Given an initial guess $x \in S^{m-1}$; Step 2: Compute the approximate Newton direction, $\xi = -\frac{\left(I - xx^\top\right) \mathbb{E}_i^{emp}\left[\frac{\rho_i'(x)}{\rho_i(x)}\right]}{\tau(x)}$; Step 3: Move from x in the direction ξ, compute $\widehat{x} = \mu_x(1)$. If $\|x - \widehat{x}\|$ is small enough, Stop. Otherwise, set $x = \widehat{x}$; Step 4: Goto step 2.

4.2 Local Convergence Analysis of ANNICA

We now show that ANNICA still enjoys local quadratic convergence. ANNICA can be restated as a selfmap on S^{m-1}, i.e.,

$$\psi : S^{m-1} \to S^{m-1}, \quad x \mapsto \tfrac{F(x)}{\|F(x)\|}, \tag{24}$$

where

$$F : S^{m-1} \to \mathbb{R}^m, \quad x \mapsto \tfrac{\zeta(x)}{\tau(x)}, \tag{25}$$

with $\tau(x)$ defined as in (21) and

$$\zeta(x) = \tau(x)x - \left(I - xx^\top\right) \mathbb{E}_i^{emp}\left[\tfrac{\rho_i'(x)}{\rho_i(x)}\right]. \tag{26}$$

Lemma 4.3. *Consider ψ as a selfmap on S^{m-1} as in* (24) *and choose a suitable coordinate system such that the standard basis vectors $\{e_1, \ldots, e_m\}$ are critical points of f defined in* (6)*. Then the standard basis vectors are fixed points of ψ.*

Proof. Recall the critical point condition of f (13) in Lemma 4.1, it is easily seen that at a critical point e_k, the second summand of ζ as defined in (26) vanishes, i.e., the expression $\zeta(e_k)$ is a scalar multiple of e_k. The result follows. □

The following theorem discusses the local convergence properties of ANNICA.

Theorem 4.1. *ANNICA considered as the map ψ as in* (24) *is locally quadratically convergent to a fixed point e_k.*

Proof. To investigate the local convergence property of ANNICA as the selfmap ψ, we will study the linear map

$$\mathrm{D}\,\psi(x) : T_x S^{m-1} \to T_{\psi(x)} S^{m-1} \tag{27}$$

at e_k, i.e. the linear map $\mathrm{D}\,\psi(e_k)$ assigns to an arbitrary tangent element $\xi \in T_{e_k} S^{m-1}$ the value $\mathrm{D}\,\psi(e_k)\xi$. Let $x = e_k$ and $\xi \in T_{e_k} S^{m-1}$, one computes

$$\begin{aligned} \mathrm{D}\,\psi(e_k)\xi &= \tfrac{\mathrm{d}}{\mathrm{d}\,\varepsilon}\psi(x+\varepsilon\xi)\big|_{\substack{\varepsilon=0 \\ x=e_k}} \\ &= \tfrac{1}{\|F(e_k)\|} \underbrace{\left(I - \tfrac{F(e_k)}{\|F(e_k)\|}\tfrac{F(e_k)^\top}{\|F(e_k)\|}\right)}_{=:P(e_k)} \mathrm{D}\,F(e_k)\xi. \end{aligned} \tag{28}$$

By Lemma 4.3, $P(e_k)$ is an orthogonal projection operator onto the complement of the span of e_k. Hence to make (28) vanish, we now show that the expression $\mathrm{D}\,F(e_k)\xi$ gives a scalar multiple of e_k.

Direct computation shows

$$\mathrm{D}\,F(e_k)\xi = -\tfrac{\mathrm{D}\,\tau(e_k)\xi}{(\tau(e_k))^2}\zeta(x) + \tfrac{\mathrm{D}\,\zeta(e_k)\xi}{\tau(e_k)}. \tag{29}$$

By the fact that the expression $-\frac{\mathrm{D}\,\tau(e_k)\xi}{(\tau(e_k))^2}$ gives a real number, the first summand on the right hand side of (29) is equal to a scalar multiple of e_k. Further computation shows that the term $\mathrm{D}\,\zeta(e_k)\xi$ is also equal to a scalar multiple of e_k. Hence the first derivative of ψ evaluated at e_k is equal to zero.

Using $x_{t+1} = \psi(x_t)$ and the fixed point condition $\psi(e_k) = e_k$, the result follows by the Taylor-type argument, as the first derivative of ψ at e_k vanishes:

$$\|\psi(x_t) - e_k\| \leq \sup_{y \in \overline{\chi(e_k)}} \|\,\mathrm{D}^2\,\psi(y)\| \cdot \|x_t - e_k\|^2$$

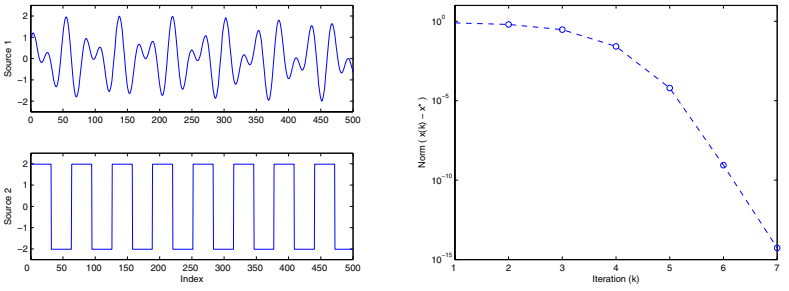
with $\overline{\chi(e_k)}$ being the closure of a sufficiently small open neighborhood of e_k. □

The performance of ANNICA will be studied below.

5 Numerical Experiments

Experiment 1. In this experiment, we construct an ideal example to show the local convergence properties of ANNICA, i.e., the approximation (21) gives the true Hessian at the critical points corresponding to a correct separation. We specify the source signal as shown in Fig. 1(a). The convergence properties of ANNICA are illustrated in Fig. 1(b), measured by the distance of the accumulation point x^* to the current iterate x_k, i.e., by $\|x_k - x^*\|$, with x^* being a column of the computed demixing matrix. The numerical results evidently verify the local quadratic convergence properties of ANNICA.

Experiment 2. To illustrate the separation performance of ANNICA, we consider an audio signal separation dataset provided by the Brain Science Institute, RIKEN, see http://www.bsp.brain.riken.jp/data. The dataset consists of 200



(a) Source signals (b) Convergence of ANNICA

Fig. 1. Local convergence properties of ANNICA

natural speech signals in Polish language sampled at 4 kHz with 20,000 samples per signal.

To extract multiple components, we parallelise ANNICA by Algorithm 2.2 in [6] (refferred here as PANNICA). PANLICA is compared with four existing linear ICA algorithms, FastICA [3], JADE [11], PCFICA [5] and RADICAL [12]. The separation performance/quality is measured by the Amari error [13], i.e.,

$$d(U,V) := \frac{1}{m}\left(\sum_{i=1}^{m}\frac{\sum_{j=1}^{m}|p_{ij}|}{\max_j |p_{ij}|} + \sum_{j=1}^{m}\frac{\sum_{i=1}^{m}|p_{ij}|}{\max_i |p_{ij}|}\right) - 2, \tag{30}$$

where $P = (p_{ij})_{i,j=1}^{m} = UV^{-1}$. Generally, the smaller the Amari error, the better the separation.

For each experiment, we separate $m = 9$ signals which are randomly chosen out of 200 sources, with a fixed sample size $n = 1,000$. By replicating the experiment 100 times, the boxplot of Amari errors is drawn in Fig. 2. It shows that without fine tuning of the kernel bandwidth h, in terms of separation quality, PANNICA outperforms three other ICA algorithms, excepct for RADICAL.

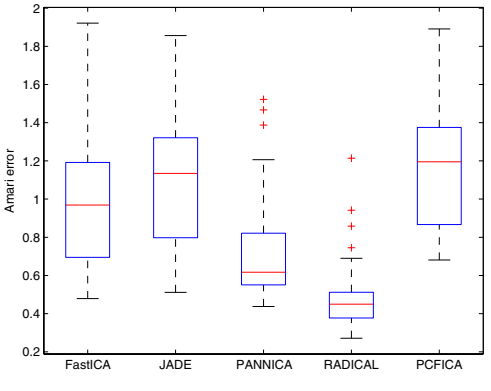


Fig. 2. Performance of PANNICA compared with different ICA algorithms

However, RADICAL is shown to be more sensitive to outlier effects, which are not considered here, than nonparametric methods. We refer to [4] for the details.

Acknowledgment

National ICT Australia is funded by the Australian Government's Department of Communications, Information Technology and the Arts and the Australian Research Council through *Backing Australia's Ability* and the ICT Research Centre of Excellence programs.

References

1. Boscolo, R., Pan, H., Roychowdhury, V.P.: Independent component analysis based on nonparametric density estimation. IEEE Transactions on Neural Networks **15(1)** (2004) 55–65
2. Comon, P.: Independent component analysis, a new concept? Signal Processing **36(3)** (1994) 287–314
3. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley, New York (2001)
4. Gretton, A., Bousquet, O., Smola, A.J., Schölkopf, B.: Measuring statistical dependence with Hilbert-Schmidt norms. In: Proceedings of the Sixteenth International Conference on Algorithmic Learning Theory (ALT 2005), Berlin/Heidelberg, Springer-Verlag (2005) 63–77
5. Chen, A., Bickel, P.J.: Consistent independent component analysis and prewhitening. IEEE Transactions on Signal Processing **53(10)** (2005) 3625–3632
6. Hüper, K., Shen, H., Seghouane, A.-K.: Local convergence properties of FastICA and some generalisations. In: Proceedings of the 31st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006) V1009–V1012
7. Shen, H., Hüper, K.: Newton-like methods for parallel independent component analysis. To appear in: MLSP 2006, Maynooth, Ireland, September 6-8, 2006.
8. Shen, H., Hüper, K.: Local convergence analysis of FastICA. In: Proceedings of the Sixth International Conference on Independent Component Analysis and Blind Source Separation (ICA 2006), Berlin/Heidelberg, Springer-Verlag (2006) 893–900
9. Shen, H., Hüper, K., Seghouane, A.-K.: Geometric optimisation and FastICA algorithms. To appear in: MTNS 2006, Kyoto, Japan, July 24-28, 2006.
10. Smith, S.: Optimization techniques on riemannian manifolds. Hamiltonian and gradientflows, algorithms and control, Fields Institute Communications **3** (1994) 113–136
11. Cardoso, J.F.: Blind source separation: statistical principles. In: Proceedings of the IEEE (90). (1998) 2099–2026
12. Miller, E.G., III, J.W.F.: ICA using spacings estimates of entropy. The Journal of Machine Learning Research **4(7-8)** (2004) 1271–1295
13. Amari, S., Cichocki, A., Yang, H.H.: A new learning algorithm for blind signal separation. In: Advances in Neural Information Processing Systems. Volume 8. (1996) 757–763