EFFICIENT GEOMETRIC METHODS FOR KERNEL DENSITY ESTIMATION BASED INDEPENDENT COMPONENT ANALYSIS

Hao Shen, Martin Kleinsteuber and Knut Hüper

National ICT Australia, Canberra Research Laboratory
Locked Bag 8001, Canberra ACT 2601, Australia
and
Department of Information Engineering, RSISE,
The Australian National University, Canberra ACT 0200, Australia
Hao.Shen@rsise.anu.edu.au, Kleinsteuber@mathematik.uni-wuerzburg.de, Knut.Hueper@nicta.com.au

ABSTRACT

The performance of Independent Component Analysis (ICA) methods significantly depends on the choice of the contrast function and the optimisation algorithm used in obtaining the demixing matrix. It has been shown that nonparametric ICA approaches are more robust than its parametric counterparts. One basic nonparametric ICA contrast was developed by approximating mutual information using kernel density estimations. In this work we study the kernel density estimation based linear ICA problem from an optimisation point of view.

Two geometric methods are proposed to optimise the kernel density estimation based linear ICA contrast function, a Jacobi-type method and an approximate Newton-like method. Rigorous analysis shows that both geometric methods converge locally quadratically fast to the correct demixing. The performance of the proposed algorithms is investigated by numerical experiments.

1. INTRODUCTION

In the past decade, Independent Component Analysis (ICA) has become a standard statistical tool for solving the Blind Source Separation (BSS) problem. It has received considerable attentions in various communities. Generally, the performance of an ICA method depends significantly on the choice of the contrast function measuring statistical independence of signals and on the appropriate optimisation technique as well.

Approaches to designing ICA contrasts can be classified into two categories: parametric and nonparametric. Parametric approaches construct the ICA criterion according to certain hypotheses of distributions of the source signals by some parameterised families of density functions. Since the distributions of sources are generally unknown, incorrect hypothesis on these distributions can lead to poor separation performance, and may even cause parametric ICA methods to fail completely in many real applications. For these reasons, there has been an increasing interest in developing nonparametric ICA methods. One of such nonparametric ICA approaches is to minimise the empirical mutual information between recovered signals by employing the Kernel Density Estimation (KDE) technique to deal with the unknown distributions of the sources, such as [14, 20, 2]. There also exist other sophisticated nonparametric ICA approaches, which do not work with the probability density, refer to [6, 1, 9] and references therein. In this work we only focus on the nonparametric linear ICA problem based on the kernel density estimation proposed in [2] from an optimisation point of view.

From an optimisation point of view, development of efficient geometric ICA methods has been very active ever since the early work of Comon [5]. Jacobi-type methods are one of the prominent methods, including Joint Approximate Diagonalisation of Eigenmatrices (JADE) [3], Extended Maximum Likelihood (EML) [21], MaxKurt [3], Jacobi optimised Empirical Characteristic Function ICA (JECFICA) [6], RADICAL [14], and so on. Although the efficiency of these Jacobi-type ICA methods has been verified by numerical experiments, the convergence properties of most Jacobi-type ICA algorithms are still theoretically unclear. Recently with a complete understanding of the FastICA algorithm [12, 18], the so-called approximate Newton-like method on manifolds, which is locally quadratically convergent, has been successfully applied to the problems of parallel parametric linear ICA [17] and one-unit KDE based linear ICA [19]. The focus of this work is on the development of efficient geometric methods, namely a Jacobi-type method and an approximate Newton-like method, for solving the KDE based linear ICA problem. Rigorous analysis shows that both methods are locally quadratically convergent to a correct demixing matrix under reasonable assumptions.

The paper is organised as follows. Section 2 introduces the KDE based linear ICA problem and gives a critical point analysis of the contrast function followed by a study of the Hessian. In Section 3, we present a Jacobi-type method and an approximate Newton-like method for optimising the KDE based linear ICA contrast. The local convergence properties of both methods are studied as well. Finally in Section 4, several numerical experiments are provided to investigate the performance of the proposed algorithms.

2. THE KDE BASED LINEAR ICA PROBLEM

2.1 Linear ICA Model and the KDE Based Contrast

The standard instantaneous noiseless linear ICA model is stated as the following relation, see [11] for details,

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

where $S = [s_1, \ldots, s_n] \in \mathbb{R}^{m\times n}$ is a data matrix of n samples of m sources ($m \ll n$), which are mutually statistically independent and drawn independently and identically, $A \in \mathbb{R}^{m\times m}$ is the mixing matrix of full rank, and $Z = [z_1, \ldots, z_n] \in \mathbb{R}^{m\times n}$

represents the observed mixtures. Note that, the mutual statistical independence can only be ensured if the sample size n tends to infinity. Nevertheless, for our theoretical analysis in Section 2.2, we assume that the complete independence holds even if the sample size is finite. The task of the linear ICA problem is to recover the source signals S by estimating the mixing matrix A based only on the observed mixtures Z. After the so-called whitening process of the mixtures, the corresponding demixing model of (1) can be stated as follows

$$Y = X^\top W, \tag{2}$$

where $W = [w_1, \ldots, w_n] = VZ \in \mathbb{R}^{m\times n}$ is the whitened observation ($V \in \mathbb{R}^{m\times m}$ is the whitening matrix), $X \in \mathbb{R}^{m\times m}$ is an orthogonal matrix being the demixing matrix, and $Y = [y_1, \ldots, y_n] \in \mathbb{R}^{m\times n}$ contains the reconstructed signals. It has been shown in [4], that the whitening process is statistically $\sqrt{n}$-consistent. Let $SO(m) := \left\{X \in \mathbb{R}^{m\times m} | X^\top X = 1, \det(X) = 1\right\}$ denote the special orthogonal group of order m. In this work, we study the linear ICA problem of finding an $X = [x_1, \ldots, x_m] \in SO(m)$ which reconstructs the sources via the model (2) based only on the whitened observations.

Minimisation of the mutual information between the recovered signal is a common criterion for ICA. In absence of knowing the true distribution of the sources, an empirical mutual information between the extracted signals can be computed by employing the kernel density estimation technique, refer to [2] for more details,

$$f\colon SO(m) \to \mathbb{R}, \quad X \mapsto \sum_{k=1}^{m} \mathbb{E}_i\left[\log\left(\frac{1}{h}\mathbb{E}_j\left[\phi\left(\frac{x_k^\top \overline{w}_{ij}}{h}\right)\right]\right)\right], \tag{3}$$

where $\overline{w}_{ij} := w_i - w_j \in \mathbb{R}^m$ represents the difference between the i-th and j-th sample, $\phi\colon \mathbb{R} \to \mathbb{R}$ is an appropriate kernel function, and $h \in \mathbb{R}^+$ is the kernel bandwidth. In this work we specify ϕ as the Gaussian kernel $\phi(t) = \exp(-t^2/2)$.

2.2 Analysis of the KDE based linear ICA contrast

Before analysing the KDE based linear ICA contrast defined in (3), we recall the tangent space of $SO(m)$ at X as

$$T_X SO(m) := \left\{\, \Xi \in \mathbb{R}^{m\times m} \,\middle|\, \Xi = X\Omega,\ \Omega \in \mathfrak{so}(m)\right\}, \tag{4}$$

where $\mathfrak{so}(m) := \{\Omega \in \mathbb{R}^{m\times m} | \Omega = -\Omega^\top\}$ denote the set of all $m \times m$ skew-symmetric matrices.

Let $\Omega = [\omega_1, \ldots, \omega_m] = (\omega_{kl})_{k,l=1}^m \in \mathfrak{so}(m)$ and define $\overline{y}_{ij}(x_k) := (x_k^\top \overline{w}_{ij})/h$. By the chain rule, the first derivative of the contrast f in direction $X\Omega \in T_X SO(m)$ is calculated as

$$\mathrm{D}f(X)(X\Omega) = \sum_{1\le k<l\le m} \omega_{kl}\left(u_{kl}(X) - u_{lk}(X)\right), \tag{5}$$

where

$$u_{kl}(X) := \mathbb{E}_i\left[\frac{\mathbb{E}_j[\phi'(\overline{y}_{ij}(x_k))\overline{y}_{ij}(x_l)]}{\mathbb{E}_j[\phi(\overline{y}_{ij}(x_k))]}\right] \in \mathbb{R}. \tag{6}$$

It can be shown that if $X^* \in SO(m)$ is a correct demixing matrix, by the whitening properties of the sources, the term $u_{kl}(X^*)$ is equal to zero for all $k \neq l$. Thus it follows that the first derivative of f vanishes at X^*, i.e., $\mathrm{D}f(X^*)(X^*\Omega) = 0$. Therefore a correct demixing matrix X^* is indeed a critical point of f.

Now computing the second derivative of f, one gets the following quadratic form

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}t^2} f(Xe^{t\Omega})\Big|_{t=0} &= \sum_{k=1}^{m} \omega_k^\top H^{(k)}(X)\omega_k \\ &\quad - \sum_{1\le k<l\le m} \omega_{kl}^2\left(u_{kk}(X) + u_{ll}(X)\right),\end{aligned} \tag{7}$$

where $H^{(k)}(X) = (h_{pq}^{(k)}(X))_{p,q=1}^m \in \mathbb{R}^{m\times m}$ with the pq-th entry

$$\begin{aligned}h_{pq}^{(k)}(X) = {}& \mathbb{E}_i\left[\frac{\mathbb{E}_j[\phi''(\overline{y}_{ij}(x_k))\overline{y}_{ij}(x_p)\overline{y}_{ij}(x_q)]}{\mathbb{E}_j[\phi(\overline{y}_{ij}(x_k))]}\right] \\ & - \mathbb{E}_i\left[\frac{\mathbb{E}_j[\phi'(\overline{y}_{ij}(x_k))\overline{y}_{ij}(x_p)]\mathbb{E}_j[\phi'(\overline{y}_{ij}(x_k))\overline{y}_{ij}(x_q)]}{(\mathbb{E}_j[\phi(\overline{y}_{ij}(x_k))])^2}\right].\end{aligned} \tag{8}$$

Let $X = X^*$. Using the properties of the whitened signals, tedious but straightforward computations lead to

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}t^2} f(X^*e^{t\Omega})\Big|_{t=0} &= \mathrm{D}^2 f(X^*)(X^*\Omega, X^*\Omega) \\ &= \sum_{1\le k<l\le m} \omega_{kl}^2\left(\widehat{u}_{kk}(X^*) + \widehat{u}_{ll}(X^*)\right),\end{aligned} \tag{9}$$

where

$$\begin{aligned}\widehat{u}_{kk}(X) := {}& \tfrac{2}{h^2}\mathbb{E}_i\left[\tfrac{\mathbb{E}_j[\phi''(\overline{y}_{ij}(x_k))]}{\mathbb{E}_j[\phi(\overline{y}_{ij}(x_k))]}\right] - \tfrac{1}{h^2}\mathbb{E}_i\left[\left(\tfrac{\mathbb{E}_j[\phi'(\overline{y}_{ij}(x_k))]}{\mathbb{E}_j[\phi(\overline{y}_{ij}(x_k))]}\right)^2\right] \\ & - \mathbb{E}_i\left[\tfrac{\mathbb{E}_j[\phi'(\overline{y}_{ij}(x_k))\overline{y}_{ij}(x_k)]}{\mathbb{E}_j[\phi(\overline{y}_{ij}(x_k))]}\right].\end{aligned} \tag{10}$$

Consequently, the Hessian of f at X^* is indeed diagonal. It is important to notice that the properties in (9) hold true only if the complete statistical independence can be ensured for the sources.

It is known already that the global minimum of the actual mutual information between reconstructed signals gives the correct separation. However there is still no proof showing that the global minimum of the nonparametric empirical mutual information defined in (3) yields the best estimation of the sources. Nevertheless, in this work, we assume that the best estimation of the sources is attained at the global minimum of the KDE based contrast defined in (3). In other words, we assume that any correct demixing matrix X^* is a nondegenerate critical point of the contrast f, i.e., for a given set of sources, the term $\widehat{u}_{kk}(X^*)$ defined in (10) is nonzero.

3. GEOMETRIC METHODS FOR THE KDE BASED LINEAR ICA PROBLEM

3.1 Jacobi-type ICA Method

Jacobi-type methods are well-known tools for solving matrix eigenvalue problems or singular value problems [7]. They can also be considered as optimisation procedures. The local convergence properties of Jacobi-type methods have been recently discussed in [10, 13]. In this section, we develop a Jacobi-type method to optimise the ICA constrast (3).

A Jacobi-type method consists of iterating so-called sweeps. Within a sweep, the Jacobi-type method uses predetermined directions, which reduce the computational burden. Once fixing a predetermined direction, a step size is chosen to yield either a local or a global optimum of the restricted

cost function. However, this is often unfeasible for general cost functions. In this work, we follow a different approach that is based on a one dimensional Newton optimisation step. Similar approximations of the optimal step size have already been used in [15, 8].

Now let us denote a basis matrix of $\mathfrak{so}(m)$ by $\Omega_{kl} \in \mathfrak{so}(m)$ with $\omega_{kl} = -\omega_{lk} = 1$ and zeros anywhere else. We define a set of directions in $\mathfrak{so}(m)$ as $\Delta := \{\Omega_{kl} \in \mathfrak{so}(m) | 1 \leq k < l \leq m\}$. Recall the formula for a geodesic in $SO(m)$ emanating from a point X in direction $X\Omega \in T_X SO(m)$

$$\gamma_X : \mathbb{R} \to SO(m), \qquad t \mapsto X \exp(t\Omega). \tag{11}$$

That is $\gamma_X(0) = X$ and $\dot{\gamma}_X(0) = X\Omega$. By fixing a direction $\Omega \in \mathfrak{so}(m)$, we constrain the ICA contrast function f defined in (3) to a geodesic, such that in each step, a one dimensional optimisation problem

$$f \circ \gamma_X : \mathbb{R} \to \mathbb{R}, \qquad t \mapsto f(X \exp(t\Omega)) \tag{12}$$

has to be solved. Having chosen a predetermined direction Ω_{kl}, we compute the first and second derivatives of $f \circ \gamma_X$ as follows

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t} f(X \exp(t\Omega_{kl}))\big|_{t=0} &= u_{kl}(X) - u_{lk}(X), \\ \tfrac{\mathrm{d}^2}{\mathrm{d}t^2} f(X \exp(t\Omega_{kl}))\big|_{t=0} &= \widehat{h}_{kl}(X) + \widehat{h}_{lk}(X), \end{aligned} \tag{13}$$

where $\widehat{h}_{kl}(X) := h^{(k)}_{ll}(X) - u_{kk}(X)$. Hence, a Newton step $\lambda_{kl} \in \mathbb{R}$ for a fixed direction Ω_{kl} can be computed by

$$\lambda_{kl} := \lambda_{kl}(X) := -\tfrac{u_{kl}(X) - u_{lk}(X)}{\widehat{h}_{kl}(X) + \widehat{h}_{lk}(X)}. \tag{14}$$

Now by the special structure of Ω_{kl}, the matrix $\Phi = (\phi_{pq})_{p,q=1}^m := \exp(\lambda_{kl}\Omega_{kl})$ is computed as being a Jacobi rotation, i.e., it is equal to the identity matrix except for the kk, ll, kl and lk entries with

$$\begin{aligned} \phi_{kk} &= \phi_{ll} = \cos(\lambda_{kl}), \\ \phi_{kl} &= -\phi_{lk} = \sin(\lambda_{kl}). \end{aligned} \tag{15}$$

To ensure the one dimensional Newton direction to point downhill, one can force the denominator defined in (14) to be positive by taking the absolute value of the Hessian of $f \circ \gamma_X$. A Jacobi-type method for optimising the KDE based linear ICA contrast function can be summarised as follows:

Jacobi-type KDE based linear ICA method
Step 1: Given an initial guess $X \in SO(m)$ and a set of predetermined directions $\Delta \subset \mathfrak{so}(m)$;
Step 2: Let $X_{\text{old}} = X$. For $1 \leq k < l \leq m$;
(i) Calculate a Newton stepsize $\lambda_{kl} \in \mathbb{R}$ $$\lambda_{kl} = -\tfrac{u_{kl}(X) - u_{lk}(X)}{\left\lvert \widehat{h}_{kl}(X) + \widehat{h}_{lk}(X) \right\rvert};$$
(ii) Update $X \leftarrow X \exp(\lambda_{kl}\Omega_{kl})$;
Step 3: If $\lVert X_{\text{old}} - X \rVert$ is small enough, Stop. Otherwise, goto Step 2.

To show the local convergence properties of the Jacobi-type KDE based linear ICA method, we first state the following theorem regarding the local quadratic convergence of a Jacobi-type method, which employs a Newton step rotation, for optimising a general cost function on $SO(m)$. This result can be proven by similar arguments explored already in [10, 13]. Due to the page limits, the proof will be omitted here.

Theorem 1 *Let $f : SO(m) \to \mathbb{R}$ be a smooth cost function, $X^* \in SO(m)$ be a local minimum of f and $\{\Omega_1, \ldots, \Omega_N\}$ be a basis of $\mathfrak{so}(m)$. If the Hessian $\mathsf{H}_f(X^*)$ is nondegenerated and if $X^*\Omega_i, X^*\Omega_j \in T_{X^*}SO(m)$ are orthogonal with respect to $\mathsf{H}_f(X^*)$ for all $1 \leq i < j \leq N$, then the Jacobi-type method using a Newton step converges locally quadratically fast.*

The local convergence properties of the Jacobi-type KDE based linear ICA method can be stated as follows

Corollary 1 *Let $X^* \in SO(m)$ be a correct demixing matrix of a linear ICA problem. Then the Jacobi-type KDE based linear ICA method is locally quadratically convergent to X^*.*

Sketch of proof. Let Δ be a standard basis of $\mathfrak{so}(m)$. It can be shown that $\mathrm{D}^2 f(X^*)(X^*\Omega_{kl}, X^*\Omega_{pq}) = 0$ for any $\Omega_{kl} \neq \Omega_{pq}$. Then by the assumption that X^* is a nondegenerate critical point, the result follows. □

3.2 Approximate Newton-like ICA Method

For a Newton type method, computing the Hessian is generally too expensive to evaluate at each iteration. Therefore we propose a cheap approximation of the true Hessian according to the explicit structure of the Hessian at a correct demixing matrix.

As suggested in (9), the Hessian of f at a correct demixing matrix X^* is indeed diagonal. We therefore might approximate the Hessian of f considered now as the linear operator $T_X SO(m) \to T_X SO(m)$ by approximating $\tfrac{\mathrm{d}^2}{\mathrm{d}t^2} f(Xe^{t\Omega})\big|_{t=0}$ using the expression

$$\sum_{1 \leq k < l \leq m} \omega_{kl}^2 \left(\widehat{u}_{kk}(X) + \widehat{u}_{ll}(X)\right). \tag{16}$$

It is easily seen that the approximation (16) gives the true Hessian of f as in (9) if $X = X^*$. Hence, for an approximate Newton method, the approximate Newton direction $\Omega = (\omega_{kl})_{k,l=1}^m \in \mathfrak{so}(m)$ can be cheaply computed by

$$\omega_{kl} = \tfrac{u_{kl}(X) - u_{lk}(X)}{\widehat{u}_{kk}(X) + \widehat{u}_{ll}(X)}. \tag{17}$$

Now, an approximate Newton iteration can be closed by projecting the Newton direction Ω back onto $SO(m)$ using the matrix exponential.

However, in contrast to the Jacobi-type update, the matrix exponential of an arbitrary $\Omega \in \mathfrak{so}(m)$ requires an eigenvalue decomposition being an unnecessarily expensive iterative process [16]. However, one can overcome this issue by using a first order approximation of the matrix exponential via a QR decomposition, which is still orthogonal as follows. Let

$$I + \Omega = Q(\Omega)R(\Omega), \tag{18}$$

be the unique QR decomposition of the matrix $I + \Omega$ with $\Omega \in \mathfrak{so}(m)$. Note that, the determinant of $I + \Omega$ is always positive. Therefore the QR decomposition is indeed unique with diagonal entries of $R(\Omega)$ being positive and $\det Q(\Omega) = 1$.

To summarise, we construct the Approximate Newton-like KDE based linear ICA method as follows:

Approximate Newton-like KDE based linear ICA method
Step 1: Given an initial guess $X \in SO(m)$; Step 2: Let $X_{\text{old}} = X$. Compute the direction $\Omega = (\omega_{kl})_{k,l=1}^{m} \in \mathfrak{so}(m)$ by $\omega_{kl} = \frac{u_{kl}(X) - u_{lk}(X)}{\widehat{u}_{kk}(X) + \widehat{u}_{ll}(X)}$; Step 3: Update $X \leftarrow XQ(\Omega)$; Step 4: If $\|X_{\text{old}} - X\|$ is small enough, Stop. Otherwise, goto Step 2.

Finally, the local convergence properties of the approximate Newton-like KDE based linear ICA method are summarised in the following result.

Theorem 2 *Let $X^* \in SO(m)$ be a correct demixing matrix of a linear ICA problem. Then the approximate Newton-like KDE based linear ICA method is locally quadratically convergent to X^*.*

Sketch of proof. Since the approximate Newton direction defined in (17) is locally a smooth map on $SO(m)$ around X^*. Each iteration of the approximate Newton-like KDE based linear ICA method is then locally a smooth map on $SO(m)$ around X^* as well. It can be shown that the first derivative of the algorithm considered now as locally a smooth map around X^* vanishes at X^*. The result follows by a Taylor series argument. □

4. NUMERICAL EXPERIMENTS

In this section, we compare our new methods with an existing parallelised approximate Newton linear ICA algorithm, PANNICA proposed in [19], which can be considered as an nonparametric version of FastICA. Firstly, the local convergence properties of all algorithms are investigated by applying them to an ideal dataset and a simple speech dataset. The performance are further compared in terms of both separation quality and convergence speed. Without fine tuning the kernel bandwidth h, we set $h = 1$.

4.1 Convergence Properties

The convergence is measured by the distance of the accumulation point X^* to the current iterate X_k, i.e., by the Frobenius norm $\|X_k - X^*\|_F$. All three methods are initialised by the same demixing matrix. It is worthwhile to notice that, by the nature of the linear ICA problem, all algorithms can converge to different correct demixing matrices, where one is a column-wise permutation of the other. Here we only study the cases where they converge to the same demixing matrix.

First, we construct an ideal dataset where the properties of the Hessian given in (9) hold true. It consists of two simple signals, one being a square wave and the other a sine wave. The numerical results shown in Figure 1 verify the local quadratic convergence properties of both methods presented in Corollary 1 and Theorem 2. It also indicates that PANNICA is locally quadratically convergent as well. Based on our implementations, all algorithms have comparable computational burden.

Now, let us study the local convergence properties of these methods on a real speech dataset, which consists of three signals with $n = 5,000$ samples. The results in Figure 2 suggest that all three algorithms appear to converge only linearly fast. This is most likely caused by the fact that, for this speech dataset, the ideal properties of the Hessian shown in (9) are not fulfilled. Nevertheless from this experiment, one can still conclude that the Jacobi-type method converges much faster than the other two approximate Newton based algorithms within a real environment.

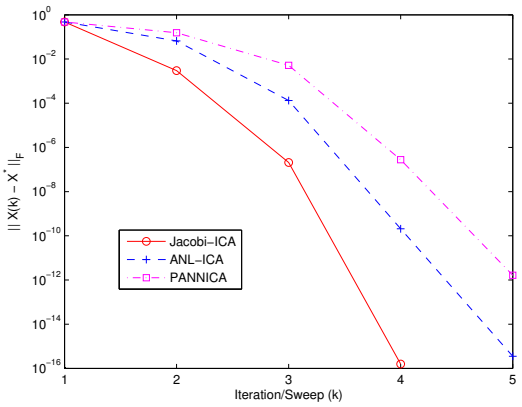

Figure 1: Convergence on an ideal dataset.

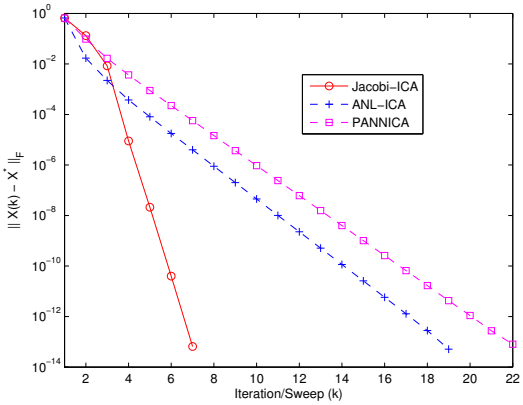

Figure 2: Convergence on a simple speech dataset.

4.2 Comparison of performance

In this experiment, all methods are tested on the benchmark speech signal dataset from the Brain Science Institute, RIKEN, see http://www.bsp.brain.riken.jp/data. The dataset consists of 200 natural speech signals sampled at 4 kHz.

The task of this experiment is to separate $m = 3$ signals which are randomly chosen out of 200 sources, with a fixed sample size $n = 1,000$. The separation performance is measured by the average Signal-to-Interference-Ratio (SIR) index, refer to [2] for the definition. Usually, a large SIR index indicates a better separation. The convergence performance is compared by the number of iterations/sweeps required by each algorithm to reach the same level of error, e.g., $\|X_k - X^*\| < 10^{-13}$. Each method is initialised by the same randomly generated demixing matrix.

By replicating the experiment 100 times, Figure 3 and Figure 4 present the quartile based boxplots of the SIR index and the number of iterations/sweeps for each method to converge respectively. In Figure 3, it is shown that both the Jacobi-type method and the approximate Newton-like method reach almost identical accuracy, while the parallelised approximate Newton method performs slightly worse than the others. By comparing the convergence performance

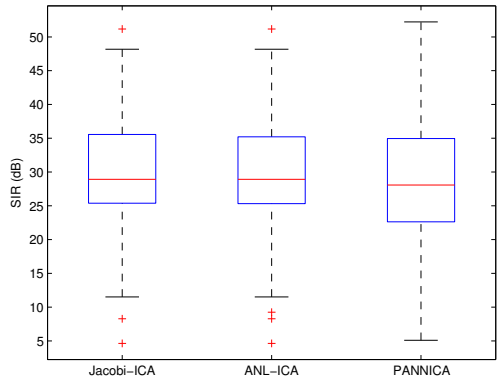

Figure 3: Comparison of separation quality.

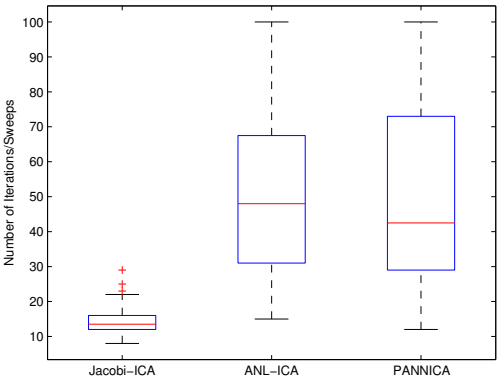

Figure 4: Comparison of convergence speed.

in Figure 4, the Jacobi-type method outperforms the other two methods. In average, it converges about 3 times faster than the others.

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

[2] R. Boscolo, H. Pan, and V. P. Roychowdhury. Independent component analysis based on nonparametric density estimation. *IEEE Transactions on Neural Networks*, 15(1):55–65, 2004.

[3] J.-F. Cardoso. High-order contrasts for independent component analysis. *Neural Computation*, 11(1):157–192, 1999.

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

[5] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

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

[7] G. Golub and C. F. van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, 2nd edition, 1989.

[8] J. Götze, S. Paul, and M. Sauer. An efficient Jacobi-like algorithm for parallel eigenvalue computation. *IEEE Transactions on Computers*, 42(9):1058–1065, 1993.

[9] A. Gretton, R. Herbrich, A. J. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005.

[10] K. Hüper. *A Calculus Approach to Matrix Eigenvalue Algorithms*. Habilitation Dissertation, Department of Mathematics, University of Würzburg, Germany, July 2002.

[11] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, New York, 2001.

[12] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[13] M. Kleinsteuber. *Jacobi-Type Methods on Semisimple Lie Algebras – A Lie Algebraic Approach to the Symmetric Eigenvalue Problem*. PhD thesis, Bayerische Julius-Maximilians-Universität Würzburg, 2006.

[14] E. G. Learned-Miller and J. W. Fisher III. ICA using spacings estimates of entropy. *Journal of Machine Learning Research*, 4:1271–1295, 2003.

[15] J. J. Modi and J. D. Pryce. Efficient implementation of Jacobi's diagonalization method on the DAP. *Numerische Mathematik*, 46(3):443–454, 1985.

[16] C. Moler and C. van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49, 2003.

[17] H. Shen and K. Hüper. Newton-like methods for parallel independent component analysis. In *Proceedings of the* 16^{th} *IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2006)*, pages 283–288, Maynooth, Ireland, 2006.

[18] H. Shen, K. Hüper, and A.-K. Seghouane. Geometric optimisation and FastICA algorithms. In *Proceedings of the* 17^{th} *International Symposium of Mathematical Theory of Networks and Systems (MTNS 2006)*, pages 1412–1418, Kyoto, Japan, 2006.

[19] H. Shen, K. Hüper, and A. J. Smola. Newton-like methods for nonparametric independent component analysis. In *Proceedings of the* 13^{th} *International Conference on Neural Information Processing, Part I, (ICONIP 2006), Hong Kong, China, October 3-6, 2006*, volume 4232 of *Lecture Notes in Computer Science*, pages 1068–1077, Berlin/Heidelberg, 2006. Springer-Verlag.

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

[21] V. Zarzoso and A. K. Nandi. Blind separation of independent sources for virtually any source probability density function. *IEEE Transactions on Signal Processing*, 47(9):2419–2432, 1999.