Minimum Discrimination Information for Independent Component Analysis

YunPeng. Li*
Department of Automation
Tsinghua University
liyp18@mails.tsinghua.edu.cn

Abstract

We drive a class of algorithms for solving the ICA problem based on minimum discrimination information.Given the projection direction,we model the components' density in Gibbs distribution with particular prior.Since there is a strong connection with regularization regression,our method can make a tradeoff between data fitting and model complexity.The proposed method enjoys fast convergence and robustness,several mainstream methods can be regarded as its special cases.Simulations confirm the feasibility and performance when compared to the presently known algorithms.

1 Introduction

The goal of independent component analysis (ICA) is to recover a latent random components vector $\mathbf{s} = (s_1, \cdots, s_m)^T$ from m observed mixtures $\mathbf{x} = (x_1, \cdots, x_m)^T$,the components of $\mathbf{s}$ are assumed to be zero-mean,mutually independent and the linear mixing process is expressed as the matrix $\mathbf{A}$,thus the mixing relationship can be modeled as:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{1}$$

The estimation $\mathbf{y} = (y_1, \cdots, y_m)^T$ of the original components $\mathbf{s}$ can be estimated via the search of the demixing matrix $\mathbf{B^T}$:

$$\mathbf{y} = \mathbf{B^T}\mathbf{x} \tag{2}$$

Each $y_i = b_i^T\mathbf{x}$ is the scaling version of unique component $s_{i'}$ due to the ambiguities in ordering and scaling,b_i is the ith column vector in $\mathbf{B}$ .Since there often exists a centering and whitening preprocessing stage for the observation $\mathbf{x}$,without loss of generality,we assume that $E(\mathbf{x}) = 0$ and $Cov(\mathbf{x}) = \mathbf{I}$.
In this paper ,we shall assume that:

- the components $\mathbf{s}$ are zero mean,mutually statistically independent,each s_i is independent distributed and at most one of the s_i is gaussian.
- each y_i is constrained in unit scale.
- both $\mathbf{A}$ and $\mathbf{B}$ are invertible matrix.

These assumptions imply that $\mathbf{B}$ is unit orthogonal matrix,so that $E(\mathbf{y}) = 0$ and $Cov(\mathbf{y}) = \mathbf{I}$.What's more,as far as we maintain the demixing matrix $\mathbf{B}^{(k)}$ to be orthogonal normal in kth iteration,the estimation $\mathbf{y}^{(k)}$ always has zero mean and unit covariance.Supposing that the sample size N is large

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

enough,the empirical distribution of $\mathbf{y}^{(k)}$ shares the same features.
Given the expectation of $\Phi(x) = (\phi_1(x), \cdots, \phi_n(x))^T$ on distribution $p(x)$, minimum discrimination information (MDI)[3] is used to determine the closest distribution $p(x)$,which has the minimum KL divergence to particular prior $p_0(x)$.

$$\begin{aligned} \min_{p(x)} \quad & D_{KL}(p||p_0) = \int p(x) \log \frac{p(x)}{p_0(x)} dx \\ \text{s.t.} \quad & \int p(x)\phi_i(x)dx = c_i \qquad\qquad i = 1, 2, \ldots, n \\ & \int p(x)dx = 1 \end{aligned} \tag{3}$$

The solution to (3) is a Gibbs distribution $p(x)$ with prior $p_0(x)$,parameter $\mathbf{w}$,and feature vector $\Phi(x)$ is

$$p(x) = \frac{p_0(x)e^{\mathbf{w}^T\Phi(x)}}{\int p_0(x)e^{\mathbf{w}^T\Phi(x)}dx} \tag{4}$$

where the denominator is the *partition function* for normalization, the (3) is nonnegative and beocomes zero only when $p(x)$ is the prior $p_0(x)$.To simplify the optimization problem,the *partition function* is restricted to unit,where $\int p_0(x)e^{\mathbf{w}^T\Phi(x)}dx = 1$,$p_0(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$,then the distribution in (4) is actually the *exponentially tilted Gaussian* used in *ProDenICA*[4]. we model the components' density in the form of (4) in the stage of data fitting and denote by $\mathcal{H}$ the function space spanned by the ϕ_i,so that $f(x) = \mathbf{w}^T\Phi(x)$ is in this space.We alternately optimize the mode $f(x)$ in $\mathcal{H}$ during data fitting and search the demixing matrix $\mathbf{B}$ for the purpose of projection pursuit.

2 Minimum discrimination information in ICA

According to (3)(4),the minimum $D_{KL}(p||p_0)$ is achieved when the $p(x)$ is in Gibbs distribution above:

$$\begin{aligned} D_{KL}^{min} &= \int \frac{p_0(x)e^{\mathbf{w}^T\Phi(x)}}{\int p_0(x)e^{\mathbf{w}^T\Phi(x)}dx} \log \frac{\frac{p_0(x)e^{\mathbf{w}^T\Phi(x)}}{\int p_0(x)e^{\mathbf{w}^T\Phi(x)}dx}}{p_0(x)} dx \\ &= \frac{\int p_0(x)e^{\mathbf{w}^T\mathbf{\Phi}(\mathbf{x})\mathbf{w}^T\Phi(x)dx}}{\int p_0(x)e^{\mathbf{w}^T\Phi(x)}dx} - \log \int p_0(x)e^{\mathbf{w}^T\Phi(x)}dx \\ &= \frac{\int p_0(x)e^{f(x)}f(x)dx}{\int p_0(x)e^{f(x)}dx} - \log \int p_0(x)e^{f(x)}dx \\ &= \int p_0(x)e^{f(x)}f(x)dx \end{aligned} \tag{5}$$

thus,for all $f(x) \in \mathcal{H}$,D_{KL}^{min} is the lower bound,actually D_{KL}^{min} is the expectation in distribution $p(x)$,which works as as the *negentropy*.The principle of our algorithm under the MDI is that:*In every iteration, If the D_{KL}^{min} keep increasing, then the $p(x)$ should keep away form the $p_0(x)$,and $f(x)$ belongs to a more restricted subspace in $\mathcal{H}$.*.A model $f(x)$ is searched in $\mathcal{H}$ for the sake of most nongaussian.

$$\begin{aligned} \max_{f\in\mathcal{H}} \quad & D_{KL}^{min} = \int p_0(x)e^{f(x)}f(x)dx \\ \text{s.t.} \quad & \int p_0(x)e^{f(x)}dx = 1 \end{aligned} \tag{6}$$

Under certain conditions,entropy-based approaches like *FastICA*[7] and maximum likelihood approaches like *ProDenICA*[4] can be described in (6).

2.1 FastICA

FastICA can be derived from the optimization problem (6) under several assumptions .On one hand we suppose that $p(x)$ is close to the $p_0(x)$,so that first order approximation $p(x) = p_0(x)e^{f(x)} \approx$

$p_0(x)(1+f(x))$ is available.On the other hand the $\phi_i(x)$s form an orthogonal normal basis for $\mathcal{H}$ in the measure of $p_0(x)$ where

$$\begin{aligned} \int p_0(x)\phi_i(x)\phi_j(x)dx &= \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases} \\ \int p_0(x)\phi_i(x)dx &= 0 \end{aligned} \tag{7}$$

The constraint in (6) is satisfied due to the particular choice of $\phi_i(x)$

$$\begin{aligned} \int p_0(x)e^{f(x)}dx &\approx \int p_0(x)(1+f(x))dx \\ &= \int p_0(x)dx + \int p_0(x)f(x)dx \\ &= 1 + \sum_{i=1}^{n} w_i \int p_0(x)\phi_i(x)dx \\ &= 1 \end{aligned} \tag{8}$$

The expectation of $\phi_i(x)$ is actually equal to the w_i in the weight vector $\mathbf{w}$.

$$\begin{aligned} c_i &= \int p(x)\phi(x)dx \\ &= \int p_0(x)e^{f(x)}\phi_i(x)dx \\ &\approx \int p_0(x)(1+\sum_{j=1}^{n} w_j\phi_j(x))\phi_i(x)dx \\ &= \int p_0(x)\phi_i(x)dx + w_i \int p_0(x)\phi_i(x)\phi_i(x)dx \\ &= w_i \end{aligned} \tag{9}$$

The equality constrained problem in (6) becomes a unconstrained minimization problem below.

$$\begin{aligned} \int p_0(x)e^{f(x)}f(x)dx &\approx \int p_0(x)(1+f(x))(f(x))dx \\ &= \int p_0(x)f(x) + p_0(x)(\sum_{i=1}^{n} w_i\phi_i(x))(\sum_{j=1}^{n} w_j\phi_j(x))dx \\ &= 0 + \sum_{i=1}^{n} w_i^2 \int p_0(x)\phi_i(x)\phi_i(x)dx \\ &= \sum_{i=1}^{n} w_i^2 \\ &= \sum_{i=1}^{n} E^2(\phi_i(x)) \end{aligned} \tag{10}$$

The conclusion is the same in [7],actually the *negentropy* here is $\|f\|_{\mathcal{H}}^2 = \int p_0(x)f(x)f(x)dx$,where the function space $\mathcal{H}$ satisfies constraints in (8).*FastICA*[7] control the model complexity via restriction method,deciding before-hand to limit the class of functions.Since only one or several particular robust $\phi_i(x)$ are selected ,the capability of model are limited when it comes to complex data.Usually,only the projection pursuit stage is conducted via a fix-point update during the whole routine,and *FastICA* works well in most cases.

2.2 ProDenICA

Different from *FastICA*,*ProDenICA* makes a tradeoff between data and model complexity via regularization methods,where a large basis are used,and the penalty term $\lambda J(f) = \int \{f^{''}(x)\}^2 dx$

translates to a penalty on the spline coefficients $\mathbf{w}$,where $\lambda \geq 0$.During each iteration,the data fitting and projection pursuit are alternately conducted under the maximum likelihood principle.
During the data fitting stage,we add the $\int p_0(x)e^{f(x)} \log p_0(x)dx$ and penalty term on D_{KL}^{min} in (6)

$$\begin{aligned} & D_{KL}^{min} + \int p_0(x)^{f(x)} \log p_0(x)dx - \lambda J(f) \\ =& \int p_0(x)e^{f(x)}\{\log p_0(x)e^{f(x)}\}dx - \lambda J(f) \\ =& E(\log p_0(x)e^{f(x)}) - \lambda J(f) \end{aligned} \tag{11}$$

If we replace the theoretical distribution $p(x)$ with empirical distribution,then the $\int p_0(x)e^{f(x)} \log p_0(x)dx \approx \frac{1}{N}\sum_{j=1}^{N} \log p_0(x_j)e^{f(x_j)}$ is constant,and the (11) is actually the same objection function in ***ProDenICA***[4],so the optimization problem in (6) with penalty term below is equivalent to ***ProDenICA*** in empirical distribution.

$$\begin{aligned} \max_{f\in\mathcal{H}} \quad & \int p_0(x)e^{f(x)}f(x)dx - \lambda J(f) \\ \text{s.t.} \quad & \int p_0(x)e^{f(x)}dx = 1 \end{aligned} \tag{12}$$

As in [4],during each iteration ,the model $f(x)$ is determined via iteratively reweighted penalized least squares regression,and the demixing matrix $\mathbf{B}$ is optimized in a same fix-point way,with the second derivative available.
ProDenICA is more powerful than ***FastICA*** and ***KernelICA*** in simulations[4],this mainly due to the compromise of data fitting and model complexity.However the tunning parameter λ is difficult to determine,even (12) can be fit using a Newton algorithm in $O(N)$ operations,the iterations in data fitting increase the computational effort,In this paper we propose an algorithm based on minimum discrimination information to solve the problem in (12),a regularization regression is directly acquired due to second order approximation,the data fitting stage can be fulfiled in just one turn,and the result coincides with the original principle in projection pursuit[2].

3 Proposed algorithm

3.1 Simplification for ProDenICA

Our algorithem makes the advantage of minimum discrimination information,maximizing the lower bound D_{Kl}^{min} in (5) for the sake of departure from gaussian.Given N observed samples from the m size random vector $\mathbf{x}$.We then alternately maximize the following problem to recover $\mathbf{y}$ as the estimation of original $\mathbf{s}$.

$$\begin{aligned} \max_{f\in\mathcal{H}} \quad & \sum_{j=1}^{m}\left\{\frac{1}{N}\sum_{i=1}^{N}\left[f_j(b_j^T\mathbf{x}_i)\right] - \lambda_j J(f_j)\right\} \\ \text{s.t.} \quad & \int p_0(b_j^T\mathbf{x}_i)e^{f_j(b_j^T\mathbf{x}_i)}dx = 1 \quad j = 1,2,\ldots,m \\ & \mathbf{b}_j^T\mathbf{b}_k = \delta_{jk} \quad j,k = 1,2,\ldots,m \end{aligned} \tag{13}$$

Each component can be acquired separately,the equality constrained problem can be modified into the unconstrained according to the conclusion in [10],for particular independent component $y_j = b_j^T\mathbf{x}$,we maximize the D_{KL}^{min} with the penalty term along with the equality constraint.

$$\max_{f_j\in\mathcal{H}} \quad \frac{1}{N}\sum_{i=1}^{N}\left[f_j(b_j^T\mathbf{x}_i)\right] - \int p_0(b_j^T\mathbf{x}_i)e^{f_j(b_j^T\mathbf{x}_i)}dx - \lambda_j J(f_j) \tag{14}$$

An approximation is conducted to handle the integral in (14),a number of L grid is created with Δ size.We construct a fine grid of L values y_l^* in increments Δ covering the recovered values y_i, and replace the probability with frequency q_l^*.

$$q_l^* = \frac{\#y_i \in [y_l^* - \frac{\Delta}{2}, y_l^* + \frac{\Delta}{2})]}{N} \tag{15}$$

Thus,we approximate (14) by replacing the theory distribution with empirical distribution in D_{KL}^{min},

$$\max_{f_j \in \mathcal{H}} \quad \sum_{l=1}^{L} \left\{ q_l^* f_j(y_l^*) - \Delta p_0(y_l^*) e^{f_j(y_l^*)} \right\} - \lambda_j J(f_j) \tag{16}$$

Since we have derive (16) according to the minimum discrimination information,so it's nature to think that $p_0(y_l^*)e^{f_j(y_l^*)}$ is not far from the prior p_0,$f_j(y_l^*)$ should to be very small,so that we can approximate the $e^{f_j(y_l^*)}$ in second order by

$$e^{f_j(y_l^*)} \approx 1 + f_j(y_l^*) + \frac{1}{2} f_j^2(y_l^*) + o(f_j^2) \tag{17}$$

We transform (16) in a minimization problem along with the second order approximation.

$$\min_{f_j \in \mathcal{H}} \quad \sum_{l=1}^{L} \left\{ \frac{1}{2} \Delta p_0(y_l^*) \left(f_j(y_l^*) - \frac{q_l^* - \Delta p_0(y_l^*)}{\Delta p_0(y_l^*)} \right)^2 \right\} + \lambda_j J(f_j) + Const \tag{18}$$

So we directly derive a regularization regression from the original problem in data fitting stage,we denote the $d_l = \frac{1}{2}\Delta p_0(y_l^*)$,and $z_l = \frac{q_l^* - \Delta p_0(y_l^*)}{\Delta p_0(y_l^*)}$,the irrelevant constant is ignored,(18) can be denoted in a standard form.

$$\min_{f_j \in \mathcal{H}} \quad \sum_{l=1}^{L} d_l \left(f_j(y_l^*) - z_l \right)^2 + \lambda_j J(f_j) \tag{19}$$

the response z_l is acquired in the regression in one turn and only contain the information from data.While in ***ProDenICA***[4],we run the iterations several times to get the response involving both information from the data and penalty term.Although the approximation in (16) seems rather rough,the z_l works comparable in simulations. Since in (19),we try to fit the data,we assume that the empirical distribution should be close to the theory distribution ,which is $q_l^* \approx \Delta p_0(y_l^*)e^{f_j(y_l^*)}$,so we can denote the their relationship

$$\begin{aligned} f_j(y_l^*) \approx & \ln \frac{q_l^*}{\Delta p_0(y_l^*)} \\ = & \ln \left(1 + \frac{q_l^* - \Delta p_0(y_l^*)}{\Delta p_0(y_l^*)} \right) \\ \approx & \frac{q_l^* - \Delta p_0(y_l^*)}{\Delta p_0(y_l^*)} \\ = & z_l \end{aligned} \tag{20}$$

Our algorithm attempts to minimize the distance between the empirical distribution q_l^* and the theory distribution $\Delta p_0(y_l^*)e^{f_j(y_l^*)}$ in certain transformation under weights measure d_l, a penalty term is added to control the model complexity.

3.2 Parametric model

3.2.1 data fitting

We can denote the $f(x)$ more specific in $\mathbf{w}^T\Phi(x)$,the problem in (16) can be proved to be a convex problem with respect $\mathbf{w}$.In this subsection,we attempt to solve the equivalent problem below via Netwon's method.

$$\min_{\mathbf{w_j}} \quad \sum_{l=1}^{L} \left\{ -q_l^* \mathbf{w}_j^T \Phi(y_l^*) + \Delta p_0(y_l^*) e^{\mathbf{w}_j^T \Phi(y_l^*)} \right\} + \lambda_j \mathbf{w}_j^T \mathbf{w}_j \tag{21}$$

We denote the optimization problem in (21) as $G_j(\mathbf{w}_j)$,the first and second derivatives are:

$$\frac{\partial G_j}{\partial \mathbf{w}_j} = \sum_{l=1}^{L} \left\{ -q_l^* \Phi(y_l^*) + \Delta p_0(y_l^*) e^{\mathbf{w}_j^T \Phi(y_l^*)} \Phi(y_l^*) \right\} + 2\lambda_j \mathbf{w}_j \tag{22}$$

$$\frac{\partial^2 G_j}{\partial \mathbf{w}_j \mathbf{w}_j^T} = \sum_{l=1}^{L} \left\{ \Delta p_0(y_l^*) e^{\mathbf{w}_j^T \Phi(y_l^*)} \Phi(y_l^*) \Phi^T(y_l^*) \right\} + 2\lambda_j \mathbf{I} \tag{23}$$

the hessian matrix $\frac{\partial^2 G_j}{\partial \mathbf{w}_j \mathbf{w}_j^T} \succcurlyeq 0$,(21) is a convex problem and can be solved with second order convergent speed.

$$\mathbf{w}_j^{(k+1)} = \mathbf{w}_j^{(k)} - \left(\frac{\partial^2 G_j}{\partial \mathbf{w}_j^{(k)} \mathbf{w}_j^{(k)T}} \right)^{-1} \frac{\partial G_j}{\partial \mathbf{w}_j^{(k)}} \tag{24}$$

Neural network model with one hidden layer has exactly the same form as the projection pursuit described in $f(x)$[5],we here use sigmoid function $\sigma(\alpha_i y_l^* + \alpha_{i0})$ as feature function $\phi_i(y_l^*)$.We can optimize the parameters in feature function via first derivatives.

$$\phi_i(y_l^*) = \sigma(\alpha_i y_l^* + \alpha_{i0}) \tag{25}$$

$$\frac{\partial G_j}{\partial \alpha_i} = \sum_{l=1}^{L} \left\{ -q_l^* \mathbf{w}_{ji} \sigma^{'}(\alpha_i y_l^* + \alpha_{i0}) y_l^* + \Delta p_0(y_l^*) e^{\sum_{i=1}^{n} \mathbf{w}_{ji} \sigma(\alpha_i y_l^* + \alpha_{i0})} \mathbf{w}_{ji} \sigma^{'}(\alpha_i y_l^* + \alpha_{i0}) y_l^* \right\} \tag{26}$$

$$\frac{\partial G_j}{\partial \alpha_{i0}} = \sum_{l=1}^{L} \left\{ -q_l^* \mathbf{w}_{ji} \sigma^{'}(\alpha_i y_l^* + \alpha_{i0}) + \Delta p_0(y_l^*) e^{\sum_{i=1}^{n} \mathbf{w}_{ji} \sigma(\alpha_i y_l^* + \alpha_{i0})} \mathbf{w}_{ji} \sigma^{'}(\alpha_i y_l^* + \alpha_{i0}) \right\} \tag{27}$$

where $\sigma^{'} = \sigma(1-\sigma)$,since $\sigma(\alpha_i y_l^* + \alpha_0)$ has low complexity than a more general nonparametric $f(y_l^*)$,the neural network might use more functions,which causes the difficulty in convergence.In certain cases e.g. discrete distribution[9],smooth spline might not be able to fit particular functions well,neural network may enjoys better performance according to *universal approximation theory*[6],despite the massive parameters.

3.2.2 projection pursuit

The model f is fixed after the data fitting stage,so we only to maximize the negentropy $\int p_0(x) e^{f(x)} dx$ for the most non-gaussian projections $\mathbf{w}_j$,a fix-point routine can be efficiently conducted.

$$\boldsymbol{b_j^{(k+1)} = E\{x f_j^{'}(b_j^{(k)T} x)\} - E\{f_j^{''}(b_j^{(k)T} x) b_j^{(k)}\}} \tag{28}$$

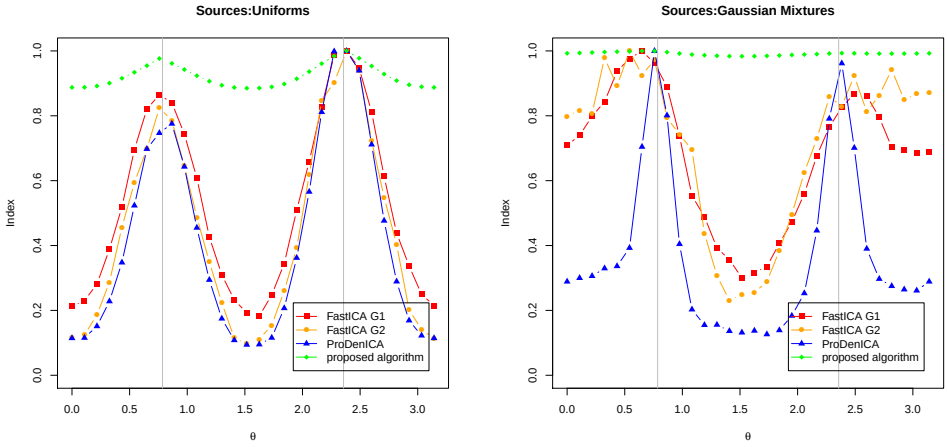
4 Experiments

The same simulations from [4][1] are designed to compare the performance with *FastICA*,*ProDenICA*,and our proposed algorithm..We denote the distance between mixing matrix $\mathbf{A}$ and demixing matrix $\mathbf{B}^T$ via Amari metric.

$$d(\mathbf{A}, \mathbf{B}) = \frac{1}{2m} \sum_{i=1}^{m} \left(\frac{\sum_{j=1}^{m} |r_{ij}|}{max_j |r_{ij}|} - 1 \right) + \frac{1}{2m} \sum_{j=1}^{m} \left(\frac{\sum_{i=1}^{m} |r_{ij}|}{max_i |r_{ij}|} - 1 \right) \tag{29}$$

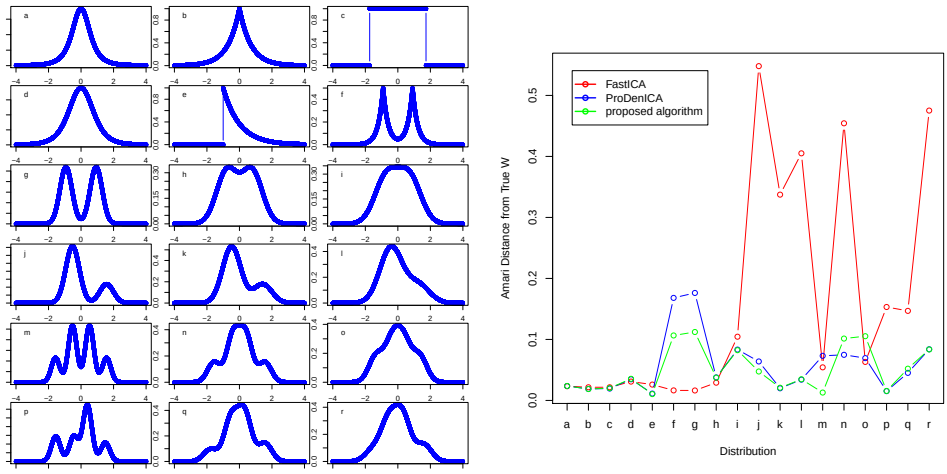
where $r_{ij} = (\mathbf{A}\mathbf{B}^{-1})_{ij}$,18 distributions are generated from official R package "ProDenICA"[8] ,both sample size N is 1024.Three simulations were given.

- We compare the normalization *negentropy* among different methods in uniform and gaussian mixtures distribution,the demixing matrix $\mathbf{B}$ is indexed by the angle θ.
- All distributions are used to generate 2-dim sources $\mathbf{s}$ and observations $\mathbf{x}$ in 30 reps,the average Amari metric is recorded.
- 4-dim $\mathbf{s}$ are picked at random in 300 reps, we compare the Amari metric among different methods.



(a) Uniforms (b) Gaussian mixtures

Figure 1: The *negentropy* and solutions found in Unifroms and GaussianMixtures distribution using *FastICA*,*ProDenICA* and the proposed algorithm. In the left example the independent components are uniformly distributed, in the right a mixture of Gaussians. In the left plot, all the procedures found the correct frame,while the proposed algorithm worked the worst; in the right plot, only the *ProDenICA* acquire the success,our proposed algorithm behaved almost constant. The vertical lines indicate the true angle θ.



(a) distributions (b) 2-dim comparisons

Figure 2: The left panel shows eighteen distributions used for comparisons. The right panel shows the average Amari metric for each method and each distribution, based on 30 simulations for each distribution

Our algorithm has the poor performance in distinguishing the true θ from other angles in Figure.1,this might be the result of rough approximation in single calculation.

The left panel Figure.2 shows the 18 distributions from [1] ,which work as a basis of comparisons,the right panel compare the average Amari metric (29) for different distributions among the tree algorithms.Our algorithm almost shares the same performance as *ProDenICA*,both of them are competitive with *FastICA*.

We pick the 4-dim source $\mathbf{s}$ at random in 300 turns and compare their *ProDenICA*,*FastICA* along with our algorithm via the measure of Amari metric, the results in Figure.3 show that our algorithm

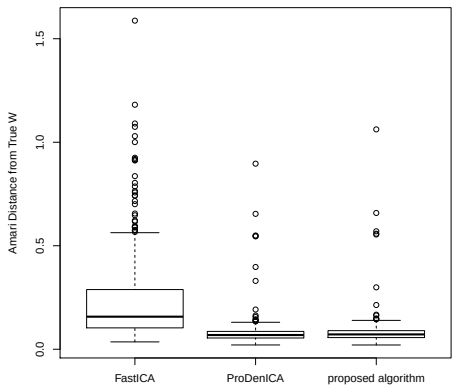


Figure 3: The 4-dim s is picked at random in 300 simulations,the results of Amari metric among different algorithms are shown in boxplot

works as well as the *ProDenICA* in this case.

5 Discussion

References

References

[1] Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *J. Mach. Learn. Res.*, 3(null):1–48, March 2003.

[2] J. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76:817–823, 1981.

[3] D.V Gokhale and S. Kullback. The minimum discrimination information approach in analyzing categorical data. *Communications in Statistics - Theory and Methods*, 7(10):987–1005, 1978.

[4] Trevor Hastie and Rob Tibshirani. Independent components analysis through product density estimation. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 665–672. MIT Press, 2003.

[5] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2009.

[6] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251 – 257, 1991.

[7] A. Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

[8] Trevor Hastie Rob Tibshirani. Prodenica: Product density estimation for ica using tilted gaussian density estimates. `cran.r-project.org/web/packages/ProDenICA/index.html`.

[9] Richard Samworth and Ming Yuan. Independent component analysis via nonparametric maximum likelihood estimation. *The Annals of Statistics*, 40, 06 2012.

[10] B. W. Silverman. On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.*, 10(3):795–810, 09 1982.