ESANN'2005 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 27-29 April 2005, d-side publi., ISBN 2-930307-05-6.

To Apply Score Function Difference Based ICA Algorithms to High-Dimensional Data

Kun Zhang and Lai-Wan Chan *

Department of Computer Science and Engineering,
The Chinese University of Hong Kong
Shatin, Hong Kong

Abstract. Recently, the score function difference (SFD) has been applied to develop ICA algorithms. But such algorithms are not suitable for high-dimensional data because the SFD estimation in a high-dimensional space is problematic. In this paper, by investigating the relationship between mutual independence and pairwise independence, we develop an approach for ICA with linear instantaneous mixtures and convolutive mixtures based on pairwise independence. This approach only involves the computation of the 2-dimensional SFD and can be directly applied to high-dimensional data. The experimental result illustrates the usefulness of this approach.

1 Introduction

Independent component analysis (ICA) is the main technique to perform blind source separation (BSS). ICA is a method for finding underlying factors or sources from multivariate observed data under the assumption that the underlying sources are statistically independent, and further provided that at most one source is Gaussian when using spatial ICA algorithms.

The outputs of ICA are as independent as possible. Then in ICA we need to exploit an independence measure. Mutual information is a canonical measure of independence. The mutual information between random variables $y_1, ..., y_n$ is defined as $I(y_1, ..., y_n) = \sum_{i=1}^{n} H(y_i) - H(\mathbf{y})$, where $\mathbf{y} = (y_1, .., y_n)^T$ and $H(\cdot)$ denotes the (differential) entropy. ICA can then be performed by minimizing the mutual information between outputs $y_1, ..., y_n$. For simplicity, in this paper we assume that the number of sources is equal to that of the observations.

Assume $\mathbf{x} = (x_1, ..., x_n)^T$ is the observation generated from the vector of independent variables $\mathbf{s} = (s_1, ..., s_n)^T$ by $\mathbf{x} = \mathbf{f}(\mathbf{s})$. Here we assume all the variables are zero-mean. Denote the de-mixing procedure by $\mathbf{y} = \mathbf{g}(\mathbf{x}|\theta)$. If the transformation $\mathbf{g}$ is one-to-one and admits a continuous Jacobian matrix $J_{\mathbf{g}}$, we have $H(Y) = H(\mathbf{x}) + E \log |\det J_{\mathbf{g}}(\mathbf{x})|$. Since $H(\mathbf{x})$ is fixed, the gradient of $I(y_1, ..., y_n)$ with respect to θ only involves the marginal densities [1].

However, in some scenarios, for instance, in ICA with convolutive mixtures or convolutive post-nonlinear (CPNL) mixtures, $\mathbf{g}$ is not one-to-one. Therefore, in minimizing $I(y_1, ..., y_n)$, the estimation of joint densities can not be avoided. In fact, gradient-based algorithms for these problems involve the score function

*This work was partially supported by a grant from the Research rants Council of the Hong Kong Special Administration Region, China.

difference (SFD), which is the variation of the mutual information resulting from a small variation in its argument [2]. The SFD can be easily exploited to develop gradient-based algorithms in mutual information optimization problems. For the definition of the SFD see Section 2.

In the estimation of the SFD, the estimation of the joint score function, or the joint probability density function (pdf) $p_{\mathbf{y}}(\mathbf{y})$ is needed. Due to the "curse of dimensionality", the estimation of the SFD required large number of samples and can be highly biased for high-dimensional data. The algorithms involving SFD are usually applied to 2-dimensional data. Thus it will be greatly beneficial if we can avoid the SFD in the learning rule when the data space is high-dimensional.

This paper proposes a scheme to do such a thing. By investigating the relationship between mutual independence and pairwise independence, in ICA with linear instantaneous mixtures or convolutive mixtures, we can successfully separate the sources by ensuring the pairwise independence between outputs. Consequently, to achieve pairwise independence, we just need to estimate the 2-dimensional SFD, instead of the SFD in the original n-dimensional space. This paper is organized as follows. Section 2 reviews some definitions and gives the definition of pairwise score function difference (PSFD). In Section 3, we investigate the relationship between pairwise independence of outputs and mutual independence between them in two scenarios, namely the linear instantaneous ICA and ICA with convolutive mixtures. And based on the relationship, the scheme to avoid the high-dimensional SFD is proposed in Section 4. Section 5 illustrates the validity of the proposed scheme with experimental results.

2 Some definitions

In this section we briefly review the definitions of the joint score function (JSF), the marginal score function (MSF) and the score function difference (SFD) for clarity. For details see [2, 3]. We also define the pairwise score function difference (PSFD), which will be involved in the ICA algorithms based on pairwise independence given in Section 4.

The score function of a scalar variable is the opposite of the log-derivative of its pdf, i.e. $\psi(y) = -\frac{d}{dy}\log p_y(y)$. The *MSF* of $\mathbf{y}$ is defined as $\psi_{\mathbf{y}}(\mathbf{y}) = (\psi_1(y_1), ..., \psi_n(y_n))^T$, where $\psi_i(y_i)$ is the score function of y_i. The *JSF* of $\mathbf{y}$ is $\varphi_{\mathbf{y}}(\mathbf{y}) = (\varphi_1(\mathbf{y}), ..., \varphi_n(\mathbf{y}))^T$, where $\varphi_i(\mathbf{y}) = -\frac{d}{dy_i}\log p_{\mathbf{y}}(\mathbf{y})$. And the *SFD* of $\mathbf{y}$ is the difference between its MSF and JSF, i.e. $\beta_{\mathbf{y}}(\mathbf{y}) = \psi_{\mathbf{y}}(\mathbf{y}) - \varphi_{\mathbf{y}}(\mathbf{y})$.

We also define the *PSFD* of $\mathbf{y}$ in terms of the SFD of each pair of it components. The *PSFD* is a vector function $\gamma_{\mathbf{y}}(\mathbf{y}) = (\gamma_1(\mathbf{y}), ...\gamma_n(\mathbf{y}))^T$, where

$$\gamma_i(\mathbf{y}) = \sum_{j=1, j\neq i}^{n} \beta_1(y_i, y_j) = (n-1)\psi_{y_i}(y_i) - \sum_{j=1, j\neq i}^{n} \varphi_1(y_i, y_j) \qquad (1)$$

From the definition, we can see the SFD of each pair of y_i is needed to construct the n-dimensional PSFD $\gamma_{\mathbf{y}}(\mathbf{y})$. Therefore for estimating $\gamma_{\mathbf{y}}(\mathbf{y})$, we need to estimate a set of 2-dimensional SFD's with $C_n^2 = \frac{n(n-1)}{2}$ elements.

3 Mutual independence vs. pairwise independence

Mutual independence implies pairwise independence, while pairwise independence does not necessarily imply mutual independence. One can easily construct such examples. But in the following two scenarios, pairwise independence of $y_1, ...y_n$ can guarantee their mutual independence, and consequently can be exploited to develop ICA algorithms, which is shown as follows.

3.1 In linear instantaneous ICA

The linear instantaneous ICA model is the basic ICA model. In this model, the observation $\mathbf{x}$ is assumed to be generated by linear transformation of the vector of independent sources $\mathbf{s}$, i.e. $\mathbf{x} = A\mathbf{s}$, where A is a non-singular square constant matrix. ICA aims at producing a vector of statistically independent signals $\mathbf{y}$ by linear transformation $\mathbf{y} = W\mathbf{x}$ such that $\mathbf{y}$ is an estimate of $\mathbf{s}$.

The Darmois-Skitovich theorem [4] provides the separability of the linear instantaneous ICA. As a direct application of the Darmois-Skitovich theorem, Theorem 11 in [5] states that in the linear instantaneous ICA, pairwise independence of $y_1, ..., y_n$ is equivalent to the mutual independence between them, and they can both be exploited to separate the sources successfully.

3.2 In ICA with convolutive mixtures

In ICA with convolutive mixtures, each element of the mixing matrix A is a linear time-invariant filter, and $s_i(t)$ are assumed to be spatially independent stochastic sequences. The generation model is described in the matrix form $\mathbf{x}(t) = [\mathcal{A}(z)]\mathbf{s}(t)$ [3]. And the separation system is $\mathbf{y}(t) = [\mathcal{W}(z)]\mathbf{x}(t)$. ICA aims to produce the outputs $y_i(t)$ as a filtered version of the sources $s_i(t)$.

The theorem, as the extension of the Darmois-Skitovich theorem to the case of convolutive mixtures, is given in [6], provided that the sequences $s_i(t)$ are temporally and spatially independent. Further provided that $s_i(t)$ are non-Gaussian, pairwise spatial independence of $\mathbf{y}(t)$ can be exploited to do ICA.

Furthermore, as a consequence of the theorem mentioned above, the Fundamental Theorem, given in [7], relaxes the assumption that $s_i(t)$ are whitened, and focuses on the colored sources $s_i(t)$ whose innovation sequences are spatially and temporally independent. From this theorem, we can see in ICA with convolutive mixtures, in general the pairwise spatial independence of the output sequences $y_i(t)$ is equivalent to the mutual spatial independence between $y_i(t)$, and it allows us to estimate the original sources up to their filtered version.

4 Use of PSFD in ICA

As discussed in Section 3, in ICA with linear instantaneous mixtures or convolutive mixtures, sources can be successfully separated by enforcing their pairwise independence. Using pairwise independence of outputs as the objective, the rules for ICA in these two scenarios involve the PSFD, instead of the SFD.

4.1 For linear instantaneous ICA

Denote the (i,k)-th entry of the de-mixing matrix W by w_{ik}. We can calculate the gradient of the mutual information between y_i and y_j with respect to w_{ik}:

$$\frac{\partial I(y_i,y_j)}{\partial w_{ik}} = \frac{\partial H(y_i)}{\partial w_{ik}} - \frac{\partial H(y_i,y_j)}{\partial w_{ik}} = E\{\beta_1(y_i,y_j)\cdot x_k\} \tag{2}$$

The i-th and j-th rows of W can be updated according to Eq. 2.

We can use the "pairwise processing" mode to update W—in each iteration we randomly choose a pair of the components of $\mathbf{y}$ and update the corresponding rows of the de-mixing matrix W according to Eq. 2 until convergence.

Alternatively, we can use the sum of mutual information between every pair of the components of $\mathbf{y}$ as the objective function:

$$J_1 = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} I(y_i,y_j) \tag{3}$$

instead of $I(y_1,...,y_n)$. J_1 is always nonnegative and is zero iff the components of $\mathbf{y}$ are pairwise independent. Combine Eq. 2 and Eq. 3, we can obtain the gradient of J_1 with respect to W, in which the PSFD $\gamma_{\mathbf{y}}(\mathbf{y})$ is involved:

$$\frac{\partial J_1}{\partial W} = E\{\gamma_{\mathbf{y}}(\mathbf{y})\mathbf{x}^T\} \tag{4}$$

This rule is the same as the one with the SFD involved [2], except that here the SFD, $\beta_{\mathbf{y}}(\mathbf{y})$, is replaced by the PSFD, $\gamma_{\mathbf{y}}(\mathbf{y})$. The equivariant algorithm can be obtained by multiplying Eq. 4 with W^TW. Eq. 4 is more stable than the "pairwise processing" mode. From the definition of the PSFD (Eq. 1) we can see that the complexity of the PSFD based algorithms is a quadratic function of n in each iteration.

4.2 For ICA with convolutive mixtures

As discussed in Subsection 3.2, ICA with convolutive mixtures can be performed by enforcing the independence between the signals $y_i(t)$ and $y_j(t-\tau)$ for all $i \neq j$ and τ.

Assume each element of W is a causal and finite impulse response (FIR) filter. Let $\mathcal{W}(z) = \mathcal{W}_0 + \mathcal{W}_1 z^{-1} + ... + \mathcal{W}_M z^{-M}$. Denote the (i,k)-th entry of $\mathcal{W}_l$ by $w_{ik}^{(l)}$. Analogously as in [3], the gradient of $I(y_i(t), y_j(t-\tau))$ with respect to $w_{ik}^{(l)}$ is

$$\frac{\partial I(y_i(t), y_j(t-\tau))}{\partial w_{ik}^{(l)}} = E\{\beta_1(y_i(t), y_j(t-\tau))\cdot x_k(t-l)\} \tag{5}$$

Then we can use the "pairwise processing" mode—in each iteration, we choose the values of i, j and τ (in [3], τ is randomly chosen from the set $\{-M,...,M\}$), and the corresponding elements of $\mathcal{W}_l$ can be updated according to Eq. 5.

A better way is to minimize the objective function:

$$J_2 = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I(y_i(t-\tau_i), y_j(t-\tau_j)) \tag{6}$$

J_2 is always nonnegative, and it is zero for all values of $\tau = (\tau_1, ..., \tau_n)^T$ iff the stochastic sequences $y_i(t)$ and $y_j(t)$ are spatially independent for $i \neq j$. Combine Eq. 5 and Eq.6, we can get the gradient of J_2 with respect to $\mathcal{W}_l$:

$$\frac{\partial J_2}{\partial \mathcal{W}_l} = E\{\gamma_{\mathbf{y}}^{(\tau)}(t)\mathbf{x}^T(t-l)\} \tag{7}$$

where $\gamma_{\mathbf{y}}^{(\tau)}(t) = \left(\gamma_1^{(\tau)}(t), ..., \gamma_n^{(\tau)}(t)\right)^T$, and $\gamma_i^{(\tau)}(t) = \sum_{j=1, j\neq i}^{n} \beta_1(y_i(t), y_j(t - \tau_j'))$, $\tau_j' = \tau_j - \tau_i$. With the gradient descent method, $\mathcal{W}_l$ are updated according to Eq. 7. For a lower computational load, in our experiment $\tau_1, ...\tau_n$ are all randomly chosen from the set $\{0, ..., M\}$ in each iteration.

5 Simulation

Since the linear instantaneous ICA is a simple case of ICA with convolutive mixtures, due to the space limitation, we only demonstrate our proposed scheme with application in the latter problem. Assuming there is no permutation indeterminacy, for measuring the separation quality, we use the output signal-to-noise ratio (SNR) defined as $SNR_i = 10 \log_{10} \frac{E\{y_i^2\}}{E\{y_i^2|_{s_i=0}\}}$, where $y_i|s_i = 0$ stands for what is at the i-th output, where the corresponding input $s_i(t)$ is zero.

Five artificially generated signals with 2000 samples are used as the input spatially independent sequences. They are a sawtooth signal, a sinusoid signal, an amplitude-modulated signal, a uniformly distributed whitened signal, and a phase-modulated signal. Each element of W is a 4th-order causal filter, i.e. $M = 4$. The learning rate in the gradient descent procedure is $\mu = 0.07$. We use Pham's method [1] to estimate the 2-dimensional SFD, because it is very fast and its result is comparatively accurate. We repeat our method for 20 different runs with the mixing system chosen randomly in each run: $\mathcal{A}_{ii}(z) = 1 + b_{ii}z^{-1} + c_{ii}z^{-2}$, and $\mathcal{A}_{ij}(z) = b_{ij} + c_{ij}z^{-1}$, $j \neq i$, where b_{ik} are drawn from the uniform distribution between 0 and 0.6, $\mathcal{U}(0, 0.6)$, and c_{ik} from $\mathcal{U}(0, 0.3)$.

Fig. 1 shows the separation result after 1000 iterations of one run, with the output SNR's 23.4dB, 27.5dB, 29.1dB, 23.1dB, and 25.8dB. We can see the original sources are successfully recovered up to their filtered version. The SNR's versus iterations are shown in Fig. 2. In fact for all the 20 runs, the sources are always successfully recovered, with the worst SNR 20.3dB.

6 Conclusion

This paper concerns how to apply ICA algorithms involving the score function difference to high-dimensional data. In the linear instantaneous ICA and ICA

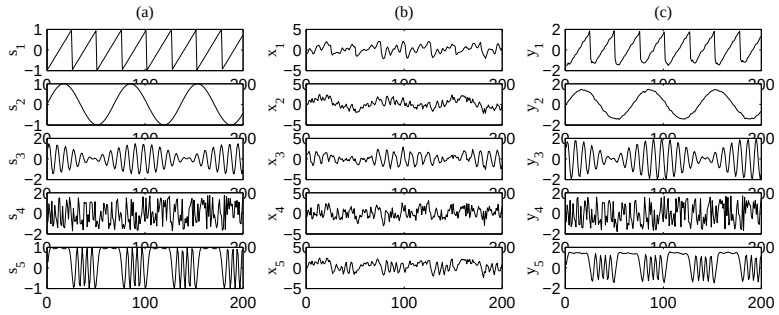

Fig. 1: (a)source sequences; (b)convolutive mixtures; (c)recovered signals. Each signal has 2000 samples. Only the first 200 samples are given for illustration.

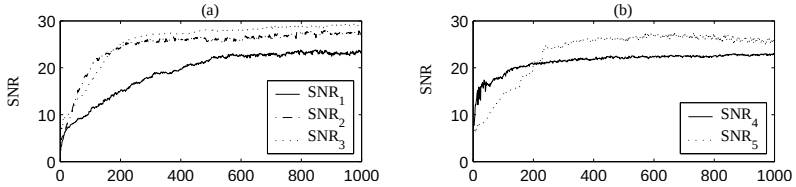

Fig. 2: Output SNR vs. iterations. (a)SNR_1, SNR_2, & SNR_3; (b)SNR_4 & SNR_5.

with convolutive mixtures, pairwise independence and mutual independence between outputs are equivalent. We then use the *pairwise* score function difference to develop ICA algorithms. Consequently only the 2-dimensional score function difference estimation is needed, regardless of the original data dimension.

References

[1] D. T. Pham. Fast algorithm for estimating mutual information, entropies and scores functions. In *Proceeding of ICA 2003*, Nara, Japan, April 2003.

[2] M. Babaie-Zadeh and C. Jutten. Differential of the mutual information. *IEEE Signal Processing Letters*, 11(1):48–51, January 2004.

[3] M. Babaie-Zadeh, C. Jutten, and K. Nayebi. Separating convolutive mixtures by mutual information minimization. In *Proc. IWANN*, volume 2, pages 834–842, Granada, Spain, June 2001.

[4] A. M. Kagan, J. V. Linnik, and C. R. Rao. *Characterization Problems in Mathematical Statistics*. Wiley, New York, 1973.

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

[6] R. W. Liu and H. Luo. Direct blind separation of independent non-Gaussian signals with dynamic channels. In *Proc. Fifth IEEE Workshop on Cellular Neural Networks and their Applications*, pages 34–38, London, England, April 1998.

[7] S. Choi, H. Hong, H. Glotin, and F. Berthommier. Multichannel signal separation for cocktail party speech recognition: a dynamic recurrent network. *Neurocomputing*, 49(1-4):299–314, December 2002.