Speed and Accuracy Enhancement of Linear ICA Techniques Using Rational Nonlinear Functions

Petr Tichavský[1], Zbyněk Koldovský[1,2], and Erkki Oja[3]

[1] Institute of Information Theory and Automation, Pod vodárenskou věží 4, P.O. Box 18, 182 08 Praha 8, Czech Republic
p.tichavsky@ieee.org
http://si.utia.cas.cz/Tichavsky.html
[2] Faculty of Mechatronic and Interdisciplinary Studies
Technical University of Liberec, Hálkova 6, 461 17 Liberec, Czech Republic
[3] Adaptive Informatics Research Centre, Helsinki University of Technology, P.O. Box 5400, 02015 HUT, Finland
Erkki.Oja@hut.fi

Abstract. Many linear ICA techniques are based on minimizing a nonlinear contrast function and many of them use a *hyperbolic tangent (tanh)* as their built-in nonlinearity. In this paper we propose two *rational* functions to replace the *tanh* and other popular functions that are tailored for separating supergaussian (long-tailed) sources. The advantage of the rational function is two-fold. First, the rational function requires a significantly lower computational complexity than *tanh*, e.g. *nine* times lower. As a result, algorithms using the rational functions are typically *twice* faster than algorithms with *tanh*. Second, it can be shown that a suitable selection of the rational function allows to achieve a better performance of the separation in certain scenarios. This improvement might be systematic, if the rational nonlinearities are selected adaptively to data.

1 Introduction

In this paper, a square linear ICA is treated (see e.g. [4,3]),

$$\mathbf{X} = \mathbf{A}\mathbf{S}, \tag{1}$$

where $\mathbf{X}$ is a $d \times N$ data matrix. The rows of $\mathbf{X}$ are the observed mixed signals, thus d is the number of mixed signals and N is their length or the number of samples in each signal. Similarly, the unknown $d \times N$ matrix $\mathbf{S}$ includes samples of the original source signals. $\mathbf{A}$ is an unknown regular $d \times d$ mixing matrix.

As usual in linear ICA, it is assumed that the elements of $\mathbf{S}$, denoted s_{ij}, are mutually independent i.i.d. random variables with probability density functions (pdf) $f_i(s_{ij})$ $i = 1, \ldots, d$. The row variables s_{ij} for all $j = 1, \ldots, N$, having the same density, are thus an i.i.d. sample of one of the independent sources denoted by s_i. It is assumed that at most one of the densities $f_i(\cdot)$ is Gaussian, and the unknown matrix $\mathbf{A}$ has full rank. In the following, let $\mathbf{W}$ denote the demixing matrix, $\mathbf{W} = \mathbf{A}^{-1}$.

M.E. Davies et al. (Eds.): ICA 2007, LNCS 4666, pp. 285–292, 2007.

Many popular ICA methods use a nonlinear contrast function to blindly separate the signals. Examples include FastICA [5], an enhanced version of the algorithm named EFICA [7], and recursive algorithm EASI [2], Extended Infomax [1], and many others. Adaptive choices of the contrast functions were proposed in [6,8].

In practical large-scale problems, the computational speed of an algorithm is a factor that limits its applications. The main goal of this paper is to propose suitable *rational* functions that can be quickly evaluated when used instead of *tanh* and other nonlinearities, and yet achieve the same or better performance. We design such suitable rational nonlinearities for algorithms FastICA and EFICA, based on our recent analytical results on their asymptotic performances, see [5,7]. It is believed that the nonlinearities proposed here will work well when applied to other methods as well.

The structure of the paper is as follows. In section II we present a brief description of algorithms FastICA and EFICA, and the analytic expressions that characterize the asymptotic performance of the methods. In section III we propose A) two general-purpose rational nonlinearities that have similar performance as *tanh*, and B) nonlinearities that are tailored for separation of supergaussian (heavy tailed) sources.

2 FastICA, EFICA, and Their Performance

In general, the FastICA algorithm is based on minimization/maximization of the criterion $c(\mathbf{w}) = \mathrm{E}[G(\mathbf{w}^T\mathbf{Z})]$, where $G(\cdot)$ is a suitable nonlinearity, called a contrast function, applied elementwise to the row vector $\mathbf{w}^T\mathbf{Z}$; see [4]. Next, $\mathbf{w}$ is the unitary vector of coefficients to be found that separates one of the independent components from a mixture $\mathbf{Z}$. Here $\mathbf{Z}$ denotes a mean-removed and decorrelated data, $\mathbf{Z} = \mathbf{C}^{-1/2}\,(\mathbf{X} - \overline{\mathbf{X}})$ where $\widehat{\mathbf{C}}$ is the sample covariance matrix, $\widehat{\mathbf{C}} = (\mathbf{X} - \overline{\mathbf{X}})(\mathbf{X} - \overline{\mathbf{X}})^T/N$ and $\overline{\mathbf{X}}$ is the sample mean of the mixture data.

In the following, in accordance with the standard notation [4], $g(\cdot)$ and $g'(\cdot)$ denote the first and the second derivative of the function $G(\cdot)$. The application of $g(\cdot)$ and $g'(\cdot)$ to the vector $\mathbf{w}^T\mathbf{Z}$ is elementwise. Classical widely used functions $g(\cdot)$ include “pow3”, i.e. $g(x) = x^3$ (then the algorithm performs kurtosis minimization), “tanh”, i.e. $g(x) = \tanh(x)$, and “gauss”, $g(x) = x\exp(-x^2/2)$.

The algorithm FastICA can be considered either in one unit form, where only one row $\mathbf{w}$ of the estimated demixing matrix $\widehat{\mathbf{W}}$ is computed, or in symmetric form, which estimates the whole matrix $\widehat{\mathbf{W}}$. The outcome of the symmetric FastICA obeys the orthogonality condition meaning that the sample correlations of the separated signals are exactly zeros.

Recently, it was proposed to complete the symmetric FastICA by a test of saddle points that eliminates convergence to side minima of the cost function, which may occur for most nonlinearities $g(\cdot)$ [9]. The test consists in checking if possible saddle points exist for each pair of the signal components exactly halfway between them in the angular sense. This test requires multiple evaluations of the primitive (integral) function of $g(\cdot)$, i.e. $G(\cdot)$. If the intention is to perform the test of saddle points in a fast manner, then it is desired that G is a rational function as well.

We introduced recently a novel algorithm called EFICA [7], which is essentially an elaborate modification of the FastICA algorithm employing a data-adaptive choice of the associated nonlinearities used in FastICA, and thus reaching a very small asymptotic error. The algorithm is initialized by performing a symmetric FastICA with a fixed nonlinearity. After that, the algorithm uses an idea of a generalized symmetric FastICA, and an adaptive choice of nonlinearities, which may be different for each signal component separately. The final demixing matrix does not obey the orthogonality condition. See [7] for details. For the purpose of this paper we shall assume that the adaptive selection of the nonlinearity in the EFICA algorithm is turned off and a fixed nonlinearity $g(\cdot)$ is used instead.

Assume now, for simplicity, that all signal components have the the same probability distribution with the density $f(\cdot)$. It was shown in [9] and in [7] that the asymptotic interference-to-signal ratio of the separated signals (one off-diagonal element of the ISR matrix) for the one-unit FastICA, for the symmetric FastICA and for EFICA is, respectively,

$$\mathrm{ISR}_{1\mathrm{U}} = \frac{1}{N}\frac{\gamma}{\tau^2}, \qquad \mathrm{ISR}_{\mathrm{SYM}} = \frac{1}{2N}\left[\frac{1}{2}+\frac{\gamma}{\tau^2}\right] \tag{2}$$

$$\mathrm{ISR}_{\mathrm{EF}} = \frac{1}{N}\frac{\gamma(\gamma+\tau^2)}{\tau^2\gamma+\tau^2(\gamma+\tau^2)} \tag{3}$$

where

$$\begin{array}{ll} \gamma = \beta - \mu^2 & \mu = \int s\, g(s)\, f(s)\, ds \\ \tau = |\mu - \rho| & \rho = \int g'(s)\, f(s)\, ds \\ & \beta = \int g^2(s)\, f(s)\, ds \end{array}$$

and the integration proceeds over the real line[1].

It can be easily seen that

$$\mathrm{ISR}_{\mathrm{EF}} = \mathrm{ISR}_{1\mathrm{U}}\,\frac{1/N + \mathrm{ISR}_{1\mathrm{U}}}{1/N + 2\,\mathrm{ISR}_{1\mathrm{U}}} \tag{4}$$

and

$$\mathrm{ISR}_{\mathrm{EF}} \le \min\left\{\mathrm{ISR}_{1\mathrm{U}}, \mathrm{ISR}_{\mathrm{SYM}}\right\}. \tag{5}$$

It is well known that all three ISR's are simultaneously minimized, when the nonlinearity $g(\cdot)$ is proportional to the score function of the distribution of the sources, $g(x) = \psi(x) = -f'(x)/f(x)$. To be more accurate the optimum nonlinearity may have the form $g(x) = c_1\psi(x) + c_2 x$, where c_1 and c_2 are arbitrary

[1] Note that it is the orthogonality constraint that makes the ISR of the symmetric FastICA lower bounded by $1/(4N)$.

constants, $c_1 \neq 0$. The choice of the constants c_1, c_2 does not make any influence on the algorithm performance. For this case it was shown EFICA is maximally efficient: the ISR in (3) in fact equals the respective Cramér-Rao-induced lower bound [9,7].

3 Optimality Issues

From the end of the previous section it is clear that it is not possible to suggest a nonlinearity that would be optimum for all possible probability distributions of the sources. The opposite is true, however: for each nonlinearity g there exists a source distribution f_g such that all other nonlinearities, that are not linear combinations of g and x, perform worse in separating the data having this distribution (in the sense of mean ISR). The density f_g can be found by solving the equation

$$g(x) = -c_1 \frac{f_g'(x)}{f_g(x)} + c_2\, x = -c_1 \frac{d}{dx}[\log f_g(x)] + c_2\, x \tag{6}$$

and has the solution

$$f_g(x) = \exp\left\{-\frac{1}{c_1} \int g(x)dx + \frac{c_2}{2c_1} x^2 + c_3\right\}. \tag{7}$$

The constants c_1, c_2, and c_3 should be selected in the way that f is a valid pdf, i.e. is nonnegative, its integral over the real line is one and have zero mean and the variance one.

For example, the nonlinearity *tanh* is optimum for the source distributions of the form

$$f_{\tanh} = C_0 \exp(-C_1 x^2)(\cosh x)^{C_2} \tag{8}$$

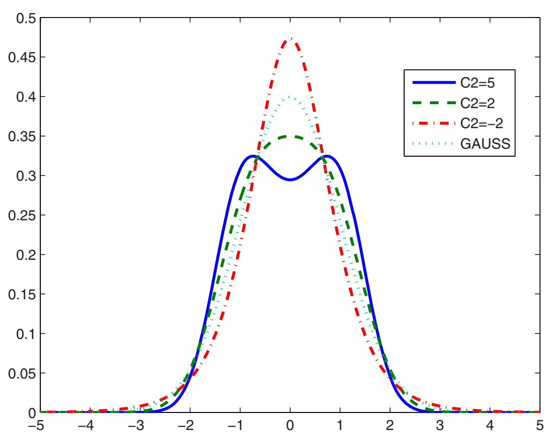


Fig. 1. Probability density functions (8) for which *tanh* is the optimum nonlinearity

where C_0, C_1, and C_2 are suitable constants. It can be shown that for any C_2 it is possible to find C_0 and C_1 such that $f_{\tanh}$ is a valid density function.

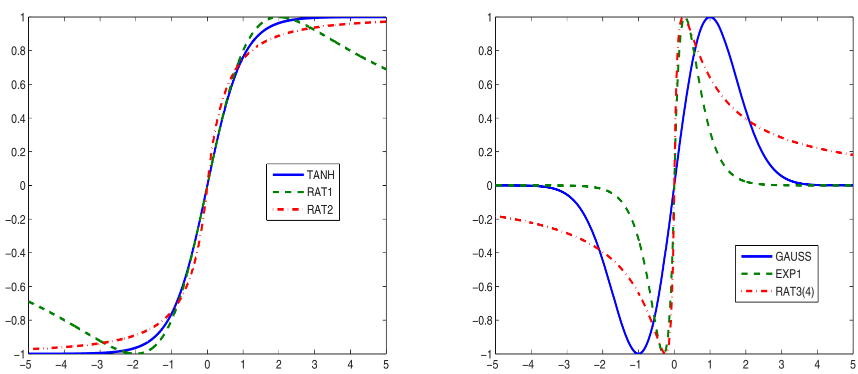
Examples of probability densities for which the *tanh* is the optimum nonlinearity are shown in Figure 1. The pdf's are compared with the standard Gaussian pdf, which would be obtained for $C_2 = 0$. The figure explains why *tanh* works very well for so many different pdf's: it includes supergaussian distributions for $C_2 < 0$ and subgaussian, even double modal distributions for $C_2 > 0$.

4 All-Purpose Nonlinearities

In this subsection we propose two rational functions that can replace *tanh* in FastICA and in other ICA algorithms,

$$g_1(x) = \frac{x}{1 + x^2/4}, \qquad g_2(x) = \frac{x(2 + |x|)}{(1 + |x|)^2} \ . \tag{9}$$

The former one has very similar behavior as *tanh* in a neighborhood of zero, and the latter one has a global behavior that is more similar to *tanh*, see Figure 2. For example, if $x \to \pm\infty$, then $g_2(x)$ approaches ± 1. These rational functions will be called RAT1 and RAT2, for easy reference.



(a) (b)

Fig. 2. Comparison of nonlinearities (a) TANH, RAT1 and RAT2 and (b) GAUSS, EXP3 and RAT3(4), discussed in Section 5. In diagram (b), the functions were scaled to have the same maximum value, 1.

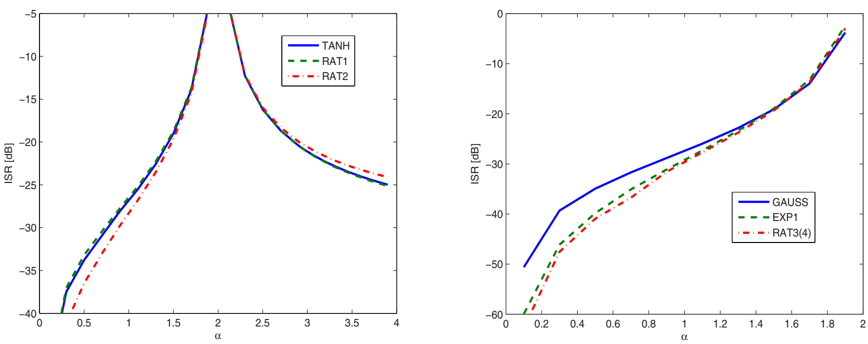
The speed of evaluation of tanh and the rational functions can be compared as follows. In the matlab notation, put $x = randn(1, 1000000)$. It was found that the evaluation of the command $y = tanh(x)$; takes 0.54 s, evaluation of RAT1 via command $y = x./(1 + x.\hat{}2/4)$; requires 0.07 s and evaluation of RAT2 via the pair of commands $h = x.*sign(x)+1$; and $y = x.*(h+1)./h.\hat{}2$; requires 0.11 s.

The computations were performed on HP Unix workstation, using a matlab profiler. We can conclude that evaluation of RAT1 is nine times faster than *tanh*, and evaluation of RAT2 is 5 times faster than *tanh*. As a result, FastICA using nonlinearity RAT1 is about twice faster that FastICA using *tanh*.

Performance of the algorithms using nonlinearities RAT1 and RAT2 appears to be very similar to that of the same algorithms using *tanh* for many probability distributions of the sources.

Assume, for example, that the source distribution belongs to the class of generalized Gaussian distribution with parameter α, which will be denoted GG(α) for easy reference. The pdf of the distribution is proportional to $\exp(-\beta_\alpha|x|)^\alpha$ where β_α is a suitable function of α such that the distribution has unit variance.

The asymptotic variance of one-unit FastICA (2) with the three nonlinearities is plotted as a function of α in Figure 3 (a). The variance is computed for $N = 1000$. We can see that performance of the algorithm with nonlinearity RAT1 is very similar to that of nonlinearity TANH. Performance of RAT2 is slightly better than the previous two ones, if the sources are supergaussian (spiky), i.e. for $\alpha < 2$, and slightly worse when the distribution is subgaussian ($\alpha > 2$).



(a) (b)

Fig. 3. Performance of one unit FastICA with nonlinearities (a) TANH, RAT1 and RAT2 and (b) GAUSS, EXP1 and RAT3(4), discussed in Section 5, for sources with distribution GG(α) as a function of the parameter α

The advantage of RAT2 compared to RAT1 is that while the primitive function of $g_1(x)$ is $G_1(x) = 2\log(1 + x^2/4)$ and it is relatively complex to evaluate, the primitive function of $g_2(x)$ is rational, $G_2(x) = x^2/(1+|x|)$ and can be evaluated faster. This might be important for the test of saddle points. It is, however, possible to combine both approaches and use RAT1 in the main iteration, and the primitive function of RAT2 in the test of saddle points.

It can be shown that the asymptotic variance $\mathrm{ISR}_{1\mathrm{U}}$ goes to infinity for any nonlinearity in rare cases, when the source pdf is a linear combination of a supergaussian and a subgaussian distributions (τ in (2) is zero). An example is shown in Figure 4, where $\mathrm{ISR}_{1\mathrm{U}}$ is plotted for sources $\mathbf{s} = \beta\mathbf{b} + \sqrt{1-\beta^2}\mathbf{l}$ as a function of parameter β, where $\mathbf{b}$ and $\mathbf{l}$ stand for binary (BPSK) and Laplacean random variables, respectively. Performances of nonlinearities TANH and RAT1 appear to be very similar, while a performance of RAT2 is slightly different.

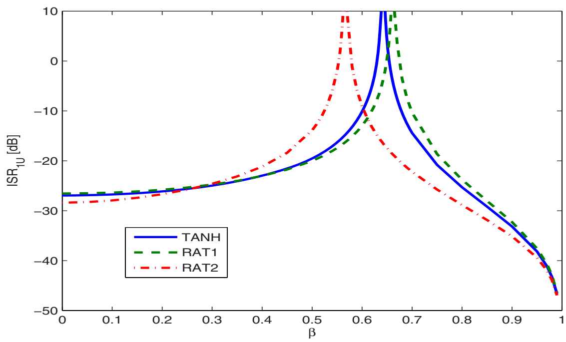


Fig. 4. Performance of one unit FastICA with nonlinearities TANH, RAT1 and RAT2 for sources of the type $s = \beta\mathbf{b} + \sqrt{1-\beta^2}\mathbf{l}$ and $N = 1000$

5 Nonlinearities for Supergaussian Sources

In [5] the following nonlinearity was proposed for separation of supergaussian (long-tailed) sources,

$$g(x) = x \exp(-x^2/2). \tag{10}$$

For a long time, this nonlinearity was considered the best known one for the separation of supergaussian sources. In [7] it was suggested to use the nonlinearity

$$g(x) = x\ \exp(-\eta|x|) \tag{11}$$

where $\eta = 3.348$ was selected. This nonlinearity will be referred as EXP1. This constant is the optimum constant for the nonlinearity of the form (11) provided that the distribution of the sources is Laplacean, i.e. GG(1). It was shown that the latter nonlinearity outperforms the former one for most of distributions GG(α) where $0 < \alpha < 2$. It was also shown in [7] that for the sources with the distribution GG(α) with $\alpha \in (0, 1/2]$ the asymptotic performance of the algorithm monotonically grows with increasing η.

In this paper we suggest to use the following nonlinearity, denoted as RAT3(b), for easy reference,

$$g_{3b}(x) = \frac{x}{(1+b|x|)^2}. \tag{12}$$

We note that like in the case of the nonlinearity EXP, the slope of the function at $x = 0$ increases with growing parameter b. This phenomenon improves the asymptotic performance of the algorithm in separation of highly supergaussian (long-tailed) sources, but makes the convergence of the algorithm more difficult.

We found that the choice $b = 4$ is quite good a trade-off between the performance and the ability to converge.

Evaluation of the nonlinearity RAT3(b) was found to be about five times faster than evaluation of EXP1. Performance of the algorithm using the 3 nonlinearities in separating sources with the distribution GG(α), $\alpha < 2$, is shown in Figure 3(b).

6 Conclusions

The rational nonlinearities were shown to be highly viable alternatives to classical ones in terms of speed and accuracy. Matlab code of EFICA, utilizing these nonlinearities can be downloaded at the second author's web page.

Acknowledgements. This work was supported by Ministry of Education, Youth and Sports of the Czech Republic through the project 1M0572 and by Grant Agency of the Czech Republic through the project 102/07/P384.

References

1. Bell, A.J., Sejnowski, T.J.: An information-maximization approach to blind separation and blind deconvolution. Neural Computation 7, 1129–1159 (1995)
2. Cardoso, J.-F., Laheld, B.: Equivariant adaptive source separation. IEEE Tr. Signal Processing 45, 434–444 (1996)
3. Cichocki, A., Amari, S.-I.: Adaptive signal and image processing: learning algorithms and applications. Wiley, New York (2002)
4. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley-Interscience, New York (2001)
5. Hyvärinen, A., Oja, E.: A Fast Fixed-Point Algorithm for Independent Component Analysis. Neural Computation 9, 1483–1492 (1997)
6. Karvanen, J., Eriksson, J., Koivunen, V.: Maximum likelihood estimation of ICA model for wide class of source distributions. Neural Networks for Signal Processing X 1, 445–454 (2000)
7. Koldovský, Z., Tichavský, P., Oja, E.: Efficient Variant of Algorithm FastICA for Independent Component Analysis Attaining the Cramér-Rao Lower Bound. IEEE Tr. Neural Networks 17, 1265–1277 (2006)
8. Pham, D.T., Garat, P.: Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. IEEE Trans. Signal Processing 45, 1712–1725 (1997)
9. Tichavský, P., Koldovský, Z., Oja, E.: Performance Analysis of the FastICA Algorithm and Cramér-Rao Bounds for Linear Independent Component Analysis. IEEE Tr. on Signal Processing 54, 1189–1203 (2006)
10. Vrins, F.: Contrast properties of entropic criteria for blind source separation, Ph.D. Thesis, Université catholique de Louvain (March 2007)