On the Optimal Non-linearities for Gaussian Mixtures in FastICA

Joni Virta[1] and Klaus Nordhausen[1,2]

[1] University of Turku, Department of Mathematics and Statistics, Turku, Finland
[2] University of Tampere, School of Health Sciences, Tampere, Finland
joni.virta@utu.fi,klaus.nordhausen@utu.fi

Abstract. In independent component analysis we assume that the observed vector is a linear transformation of a latent vector of independent components, our objective being the estimation of the latter. Deflation-based FastICA estimates the components one-by-one by repeatedly maximizing the expected value of some function measuring non-Gaussianity, the derivative of which is called the non-linearity. Under some weak assumptions, the asymptotically optimal non-linearity for extracting sources with a specific density is given by the location score function of the density. In this paper we look into the consequences of this result from the viewpoint of estimating Gaussian location and scale mixtures. As one of our results we justify the common use of hyperbolic tangent, *tanh*, as a non-linearity in blind clustering by showing that it is optimal for estimating certain Gaussian mixtures. Finally, simulations are used to show that the asymptotic optimality results hold in various settings also for finite samples.

Keywords: Asymptotic optimality, hyperbolic tangent, independent component analysis.

1 Introduction

In independent component analysis (ICA) one assumes that the observed k-vectors $\mathbf{x}_i$, $i = 1, \ldots, n$, are independent realizations of a random vector $\mathbf{x}$ which is a linear transformation of an unobserved vector $\mathbf{z}$ of independent source signals. This corresponds to the model

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Omega}\mathbf{z}, \tag{1}$$

where $\boldsymbol{\mu} \in \mathbb{R}^k$, the mixing matrix $\boldsymbol{\Omega} \in \mathbb{R}^{k \times k}$ is non-singular and the latent vector $\mathbf{z}$ has mutually independent components satisfying the following two assumptions: (i) The components of $\mathbf{z}$ are standardized in the sense that $\mathrm{E}(\mathbf{z}) = \mathbf{0}$ and $Cov(\mathbf{z}) = \mathbf{I}$, (ii) at most one of the components of $\mathbf{z}$ is Gaussian.

Assumption (i) fixes both the location $\boldsymbol{\mu}$ and the scales of the columns of $\boldsymbol{\Omega}$ in (1) and (ii) ensures that there are no orthogonally invariant column blocks in the matrix $\boldsymbol{\Omega}$ [7]. After these assumptions the signs and the order of the components of $\mathbf{z}$ are still not fixed but this is usually satisfactory in applications.

In ICA one wants to find an estimate for the inverse of the unmixing matrix, $\boldsymbol{\Omega}^{-1} =: \boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k)^T$, after which, e.g. the first estimated independent component is obtained as $\boldsymbol{w}_1^T \mathbf{x}$. In FastICA [8] this is done by first standardizing the observed vector, $\boldsymbol{x} \mapsto \boldsymbol{x}_{st} := Cov(\boldsymbol{x})^{-1/2}(\boldsymbol{x} - \mathrm{E}(\boldsymbol{x}))$, which leaves $\boldsymbol{x}_{st}$ a rotation away from the vector $\boldsymbol{z}$ [1]. Then, for estimating this rotation one chooses a non-linearity function $g : \mathbb{R} \mapsto \mathbb{R}$ for which we denote $\mathbf{g}(\mathbf{x}) := (g(x_1), g(x_2), \ldots, g(x_k))^T$. The estimation is formalized in the following definition which, albeit a bit unorthodox way of defining FastICA, nicely captures all variants of it.

Definition 1. *L_p-FastICA finds an orthogonal matrix $\boldsymbol{U}$ satisfying*

$$\boldsymbol{U} = \underset{\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}}{argmax} \| E(\boldsymbol{g}(\boldsymbol{U}\boldsymbol{x}_{st})) \|_p,$$

where $\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^k |x_i|^p\right)^{\frac{1}{p}}$ is the L_p-norm, $p \geq 1$.

Remark 1. L_1-FastICA is equivalent to the symmetric FastICA [8] and L_2-FastICA is equivalent to the squared symmetric FastICA [11].

Remark 2. Also deflation-based FastICA [8] has a similar formulation using vector norms. Namely, it can be seen as a repeated application of L_∞-FastICA, where $\|\mathbf{x}\|_\infty = \max_i(|x_i|)$. In the first step we search for a single component that maximizes $|\mathrm{E}(g(x))|$ and repeat the process $(k-1)$ times in the orthogonal complement of the already found directions.

The estimating equations of deflation-based FastICA, see e.g. [14, 11], show that the non-linearity g is in deflation-based FastICA invariant to its linear part (hence its name) and also to scaling and sign-change of its argument. However, the same does not hold for either symmetric or squared symmetric FastICA.

Lemma 1. *Deflation-based FastICA is invariant under transformations $g(x) \mapsto ag(sx) + bx + c$, where $a, b, c \in \mathbb{R}, a \neq 0, s \in \{-1, 1\}$, of the used non-linearity g.*

Remark 3. The result of Lemma 1 holds also if one uses the alternative, modified Newton-Raphson algorithm, see [7, 11].

If two non-linearities, g_1 and g_2, are equal up to the invariance specified in Lemma 1 we denote it as $g_1 \equiv g_2$. In addition to this invariance, deflation-based FastICA has also another interesting feature; given a component with a regular enough density function, in a certain sense optimal non-linearity for extracting it can be stated. This is formalized in the following lemma, the proof of which can be found in [5, 10, 11].

Lemma 2. *Let the random variable z_1 in (1) have a twice continuously differentiable density function $f : \mathbb{R} \mapsto [0, \infty)$. Then, assuming that z_1 is in deflation-based FastICA extracted first, the non-linearity $g(x) = -f'(x)/f(x)$ minimizes the sum of asymptotic variances of the elements of $\hat{\boldsymbol{w}}_1$.*

In the following all uses of the word *optimal* are in the sense of Lemma 2. For other criteria for choosing the non-linearity in deflation-based FastICA see e.g. [2].

The result of Lemma 2 holds conditional on the component of interest being the first to be extracted. This is trivially satisfied in the case the components of $\boldsymbol{z}$ are identically distributed and in a more general case its extraction first can be forced by choosing the starting value of the algorithm appropriately as done for example in reloaded FastICA [13] and adaptive FastICA [10].

In standard FastICA mainly four non-linearity functions are used in practice. They are usually denoted *skew*, *pow3*, *tanh* and *gauss* and correspond to the functions $g(x) = x^2$, $g(x) = x^3$, $g(x) = \tanh(x)$ and $g(x) = x\exp(-x^2/2)$, respectively [6]. The first two are based on the classical use of higher-order cumulants in projection pursuit [4] and the last two provide robust approximations for the negentropy [7], the most popular of the four being *tanh*.

While "robustness" issues are irrelevant when choosing the non-linearity as FastICA will never be robust due to the whitening based on the covariance matrix [14], it is still of interest to ask why some non-linearities seem to work better than others in various situations. Reversing the thinking of Lemma 2 we can then ask, given a non-linearity g, is it possibly optimal for any density f? Solving of the trivial first-order differential equation in combination with Lemma 1 yields the following result.

Lemma 3. *A differentiable and integrable function $g : \mathbb{R} \to \mathbb{R}$ is the optimal non-linearity for independent components with densities $f : \mathbb{R} \to \mathbb{R}_+$ satisfying $f(x) \propto exp(a\int_0^{sx} g(y)dy + bx^2 + cx)$, $\int_{-\infty}^{\infty} xf(x)dx = 0$ and $\int_{-\infty}^{\infty} x^2 f(x)dx = 1$ for some $a, b, c \in \mathbb{R}, a \neq 0, s \in \{-1, 1\}$.*

An analogous result for deflation-based FastICA, symmetric FastICA and EFICA [9] was given already in [16]. However, our version enjoys an extra degree of freedom in its parameters as restricting to deflation-based FastICA only allows, based on Lemma 1, the inclusion of the linear term cx in Lemma 3. The last two conditions in Lemma 3 reflect our assumption (i) that the independent components are standardized. Using Lemma 3 we see that *pow3* is optimal for sources with power exponential density, $f(x) = 2^{5/4}\sqrt{\pi}\Gamma(1/4)^{-2}exp(-2\pi^2\Gamma(1/4)^{-4}x^4)$, where $\Gamma(\cdot)$ is the Gamma function. The non-linearities *skew* and *gauss* are optimal for densities satisfying respectively $f(x) \propto exp(ax^3 + bx^2 + cx)$ and $f(x) \propto exp(a\, exp(-x^2/2) + bx^2 + cx)$ and, to the authors' knowledge, no common probability distributions defined on the whole $\mathbb{R}$ have such densities. However, an interesting remark can still be made. Namely, define sub-Gaussian (super-Gaussian) densities as those $f(x) = exp(-h(x))$ for which $h'(x)/x$ is increasing (decreasing) in $(0, \infty)$ [15]. Then Lemma 2 says that if a non-linearity $g(x)$ is optimal for some density, then that density is sub-Gaussian (super-Gaussian) if $g(x)/x$ is increasing (decreasing) in $(0, \infty)$, verifying the heuristics of using *pow3* for extracting sub-Gaussian sources and *gauss* for extracting super-Gaussian sources [6]. Note however that the definitions of sub- and super-Gaussian densities in [6] are based on kurtosis values and not on density functions.

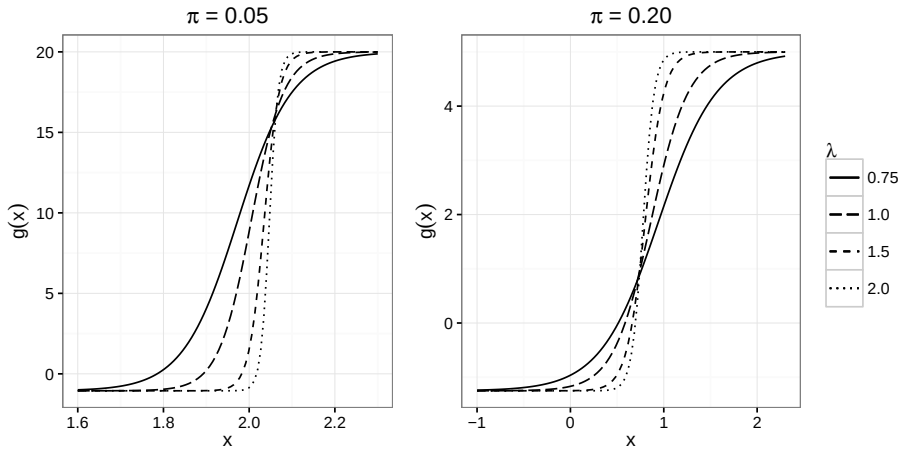

Fig. 1. The optimal non-linearities $g(x)$ for extracting $\mathcal{L}(\pi, \lambda)$-distributed components for various values of π and λ.

2 Optimal Non-linearities for Gaussian Mixtures

It is well-known that Gaussian mixture distributions are suitable for approximating other distributions, see e.g. [3] who show that elliptical distributions can be seen as scale mixtures of Gaussian distributions. Motivated by this we will in the following consider two special cases of Gaussian mixture distributions.

2.1 Gaussian Location Mixtures

Consider the following two-parameter mixture distribution family, $\mathcal{L}(\pi, \lambda)$.

$$\pi\mathcal{N}(\frac{\lambda_1}{\sqrt{4+\lambda_1\lambda_2}}, \frac{4}{4+\lambda_1\lambda_2}) + (1-\pi)\mathcal{N}(\frac{-\lambda_2}{\sqrt{4+\lambda_1\lambda_2}}, \frac{4}{4+\lambda_1\lambda_2}), \tag{2}$$

where the mixing proportion $\pi \in (0, 1)$, the location parameter $\lambda \in (0, \infty)$ and for brevity we denote $\lambda_1 := \lambda/\pi$ and $\lambda_2 := \lambda/(1-\pi)$. It is easily checked that the random variable $z_1 \sim \mathcal{L}(\pi, \lambda)$ satisfies $\mathrm{E}(z_1) = 0$ and $\mathrm{Var}(z_1) = 1$ for any permissible choices of the parameters and the family $\mathcal{L}(\pi, \lambda)$ then contains every standardized two-group Gaussian location mixture distribution where the two groups have the same variance. Applying then Lemma 2 to this family yields

Theorem 1. *Let $z_1 \sim \mathcal{L}(\pi, \lambda)$ for some $\pi \in (0, 1)$, $\lambda \in (0, \infty)$. Then the optimal non-linearity for extracting z_1 satisfies*

$$g(x) \equiv \left(\pi + \left(e^{t(x)} - 1\right)^{-1}\right)^{-1},$$

where $t(x) = (\lambda_1 + \lambda_2)(2x\sqrt{4+\lambda_1\lambda_2} - \lambda_1 + \lambda_2)/8$.

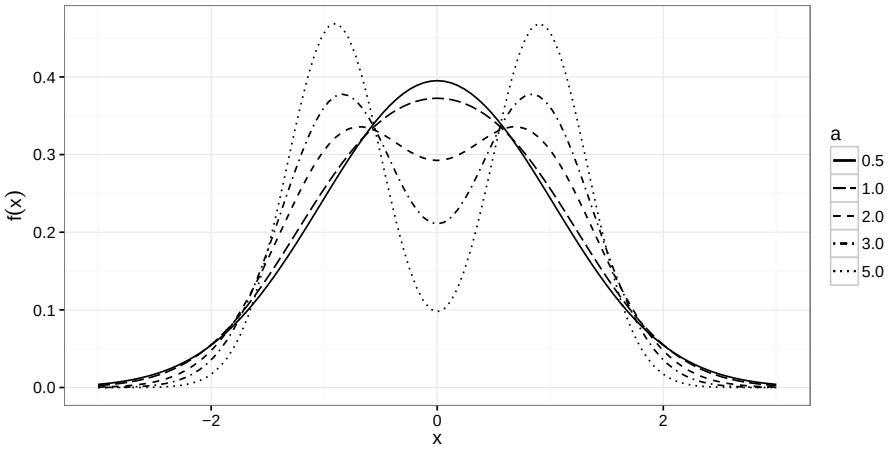

Fig. 2. The plot shows, for different values of $a \in (0, \infty)$, the densities of the symmetric Gaussian location mixtures for which the non-linearity $g(x) = tanh(ax)$ is optimal.

The resulting optimal non-linearity in Theorem 1 is quite complex and not of any standard functional form. Its graph for some select choices of parameters is depicted in Figure 1, all cases exhibiting a sigmoid-like shape. However, considering the symmetric case, $\pi = 1/2$, simplifies the formulae greatly.

Corollary 1. *Let* $z_1 \sim \mathcal{L}(1/2, \lambda)$ *for some* $\lambda \in (0, \infty)$. *Then the optimal non-linearity for extracting* z_1 *satisfies*

$$g(x) \equiv tanh(\lambda\sqrt{1+\lambda^2}x).$$

Corollary 1 says that the widely-used hyperbolic tangent is actually optimal for estimating symmetric two-group Gaussian location mixtures, justifying its use in FastICA when we have expectations to find symmetric bimodal components. A similar optimality result for *tanh* was given already in [16] but the resulting family of distributions was not studied further and Corollary 1 now goes to show that the family is for deflation-based FastICA actually $\mathcal{L}(1/2, \lambda)$. As a non-linearity *tanh* is usually given in the form $g(x) = tanh(ax)$ where $a \in (0, \infty)$ is a tuning parameter and we have, using Corollary 1, plotted in Figure 2 the densities of the distributions for which $g(x) = tanh(ax)$ is the optimal non-linearity for various values of a. The plot implies that the more separated the groups one wants to find, the higher the value of a should be. See Section 3 for simulations of this heuristic. Curiously, the standard case $a = 1$ is optimal for components $z_1 \sim \mathcal{L}(1/2, \sqrt{\phi})$, where $\phi := (\sqrt{5} - 1)/2$ is the golden ratio.

2.2 Gaussian Scale Mixtures

We next consider Gaussian scale mixtures via a two-parameter mixture distribution family, $\mathcal{S}(\pi, \theta)$, that contains every standardized two-group Gaussian scale

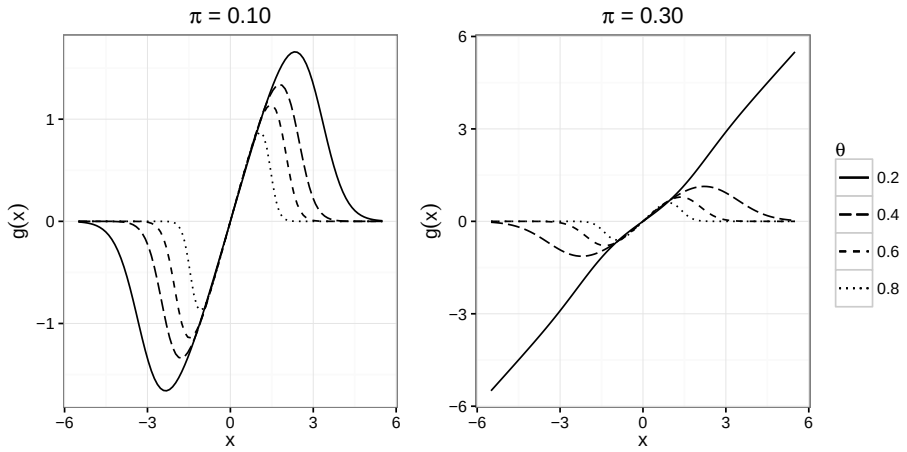

Fig. 3. The optimal non-linearities $g(x)$ for extracting $\mathcal{S}(\pi, \theta)$-distributed components for various values of π and θ.

mixture distribution where the two groups have the same expected value:

$$\pi\mathcal{N}\left(0, \frac{\theta}{\pi}\right) + (1-\pi)\mathcal{N}\left(0, \frac{1-\theta}{1-\pi}\right), \tag{3}$$

where the mixing proportion $\pi \in (0,1)$ and the scale parameter $\theta \in (0,1)$. Again the random variable $z_1 \sim \mathcal{S}(\pi, \theta)$ satisfies $\mathrm{E}(z_1) = 0$ and $\mathrm{Var}(z_1) = 1$ for all combinations of the parameters, yielding the following result via Lemma 2.

Theorem 2. *Let $z_1 \sim \mathcal{S}(\pi, \theta)$ for some $\pi, \theta \in (0,1)$. Then the optimal non-linearity for extracting z_1 satisfies*

$$g(x) \equiv x\left(1 + \left(\frac{\pi}{1-\pi}\right)^{3/2}\left(\frac{1-\theta}{\theta}\right)^{1/2} e^{t(x)}\right)^{-1},$$

where $t(x) = x^2(\theta - \pi)/(2\theta(1-\theta))$.

Examples of the non-linearity in Theorem 2 are plotted in Figure 3. In order to obtain a simpler formula with only one tuning parameter notice that choosing $\theta = 1 - \pi$ corresponds for extreme values of π to a heavy-tail model and in this special case the result of Theorem 2 simplifies as follows.

Corollary 2. *Let $z_1 \sim \mathcal{S}(\pi, 1-\pi)$ for some $\pi \in (0,1)$. Then the optimal non-linearity for extracting z_1 satisfies*

$$g(x) \equiv x\left(1 + \left(\frac{\pi}{1-\pi}\right)^{2} e^{t(x)}\right)^{-1},$$

where $t(x) = x^2(1-2\pi)/(2\pi(1-\pi))$.

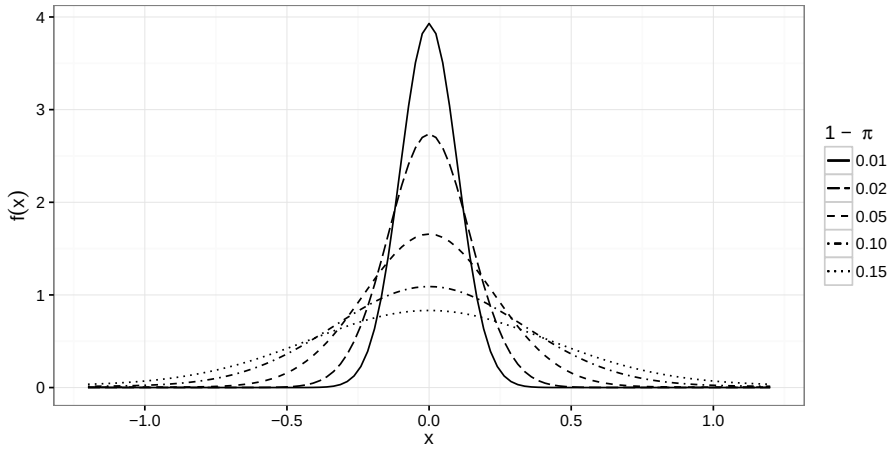

Fig. 4. The plot shows, for different values of $1-\pi \in (0,1)$, the densities of the Gaussian scale mixtures for which the non-linearity *tail* is optimal.

In Figure 4 we have plotted for various values of $1-\pi$ the densities of those Gaussian scale mixtures for which the non-linearity of Corollary 2 (referred to hereafter as *tail*) is optimal. As distributions $\mathcal{S}(\pi, 1-\pi)$ with extreme values of π are basically symmetric, heavy-tailed distributions a reasonable guess is that the non-linearity *tail* is useful for extracting also other heavy-tailed symmetric components. This will be investigated in the next section.

3 Simulations

3.1 The Choice of the Tuning Parameter in *tanh(ax)*

The simulations are divided into two parts: the investigation of the tuning parameter a in $tanh(ax)$ and the testing of the non-linearity *tail* of Corollary 2.

For the first we used two different three-variate settings where all components of $\boldsymbol{z} \in \mathbb{R}^3$ were either $\mathcal{L}(0.5, 2)$- or $\mathcal{L}(0.4, 2)$-distributed and we used deflation-based FastICA to estimate one of the components. We considered the non-linearities, *pow3*, *gauss*, $tanh(x)$, $tanh(3x)$ and $tanh(5x)$, of which the last one should work the best in the first setting and the second setting investigates how the non-linearities handle small deviations from the distribution they are optimal for. *skew* is not included as it carries no information in symmetric settings. The sample size is taken to be $n = 1000, 2000, 4000, 8000, 16000, 32000$ and the number of repetitions is 10000.

As all three i.i.d. components of $\boldsymbol{z}$ are equally likely to be estimated first, we measured the success of the extraction by the criterion $D^2(\hat{\boldsymbol{w}}_1) = \min\{\|\boldsymbol{P}\boldsymbol{J}\hat{\boldsymbol{w}}_1 - \boldsymbol{e}_1\|^2\}$, where $\hat{\boldsymbol{w}}_1$ is the estimated first direction, $\boldsymbol{e}_1 = (1,0,0)^T$ and the minimum is taken over all 3×3 permutation matrices $\boldsymbol{P}$ and 3×3 diagonal matrices $\boldsymbol{J}$ with

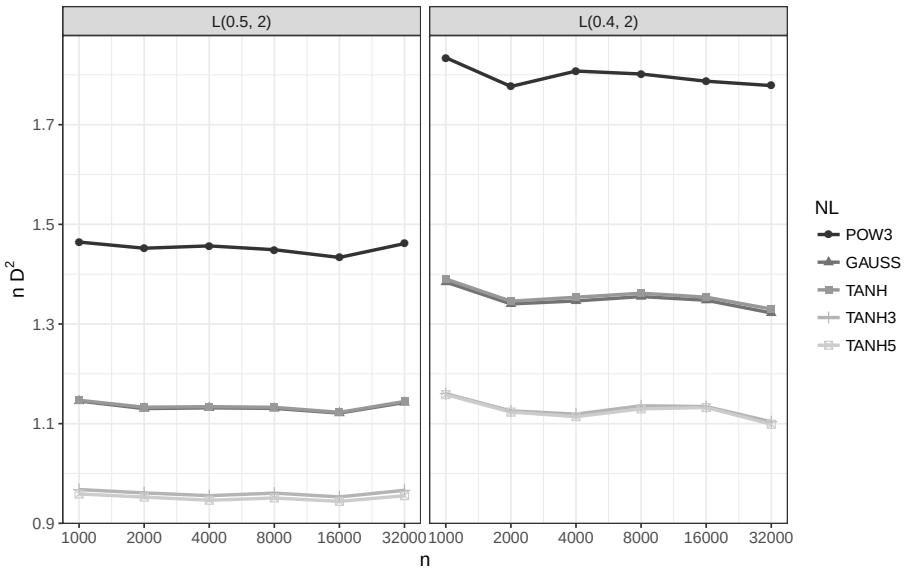

Fig. 5. The results of the first simulation.

diagonal elements equal to ± 1. Thus $D^2 = 0$ means that we succeeded perfectly in estimating one of the components. In the simulations we furthermore scaled D^2 by the sample size n, see the modified minimum distance index in [17].

The resulted mean criterion values are given in Figure 5 and show that $tanh(3x)$ and $tanh(5x)$ performed the best in the symmetric setting, *gauss* and $tanh(x)$ not being that far behind. More interestingly, the same conclusions can be drawn also in the asymmetric case. Only the overall level of the extraction is a bit worse.

3.2 Estimating Scale Mixtures and Heavy-tailed Components

To evaluate the performance of *tail* we considered two three-variate settings where the components of $\boldsymbol{z}$ were all either $\mathcal{S}(0.10, 0.90)$- or t_5-distributed (and standardized), where t_5 denotes a t-distribution with 5 degrees of freedom. The sample sizes, the number of repetitions and the criterion function were the same as in the previous simulation and we used six non-linearities, *pow3*, *gauss*, $tanh(x)$, *tail* with $\pi = 0.1$, *tail* with $\pi = 0.3$ and *rat3* with $b = 4$, see below. The third-to-last non-linearity should be superior in the first setting and with the second setting we experiment whether *tail* works for other heavy-tailed distributions also. The non-linearity *rat3*, $g(x) = x/(1+b|x|)^2$, was proposed in [16] to estimate heavy-tailed sources and there $b = 4$ was suggested as a balanced choice for the tuning parameter.

The results in Figure 6 again show that in the first setting the asymptotically optimal non-linearity, *tail* with $\pi = 0.1$, gives the best separation also for finite

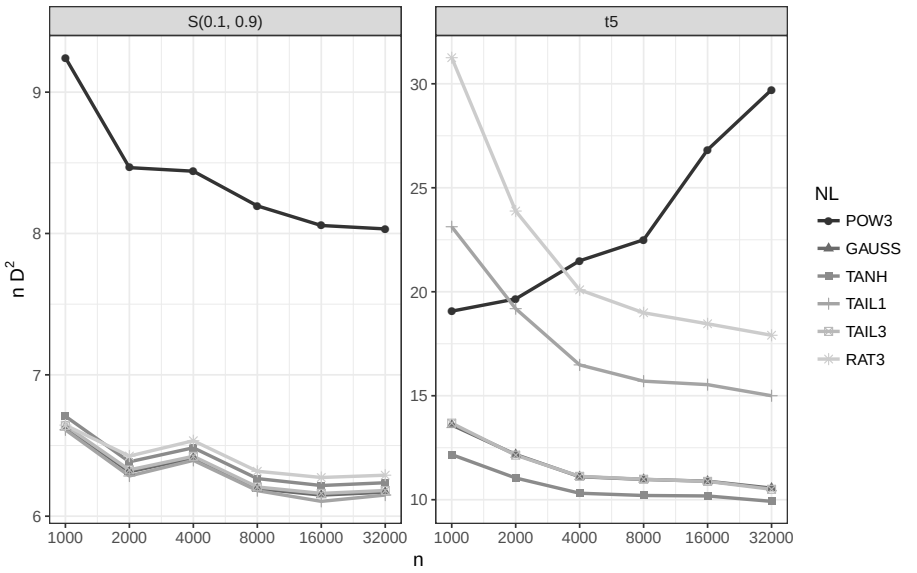

Fig. 6. The results of the second simulation.

samples. In the "experimental" setting with the t_5-distribution $tanh(x)$ proved most useful but also *gauss* and *tail* with $\pi = 0.3$ were quite successful.

4 Discussion

In FastICA the choice of the non-linearity, e.g. the popular *tanh*, is usually motivated with heuristic claims and asymptotic arguments showing that a particular non-linearity is optimal for some class of distributions. However, one is usually not interested in a non-linearity that works well in only a few cases but instead in a multitude of situations – and as also our simulations show, *tanh* performs in general quite well, also with distributions for which it is not optimal. And although there exists cases where *tanh* does not work at all [18, 11], this drawback should not be given too much weight; only a few non-linearities are so far shown to work for any combination of sources, assuming that at most one of them has an objective function value of zero, see e.g. [12, 17]. Such un-estimable distributions can actually be crafted for any non-linearity [16].

The use of different non-linearities for different components in FastICA has also been considered, see e.g. EFICA [9] and adaptive deflation-based FastICA [10]. While EFICA tries to estimate the optimal non-linearities from the data, adaptive deflation-based FastICA chooses them out of a set of candidates. It seems thus reasonable to include in this set non-linearities which are known to have optimality properties, such as the ones given in our Corollaries 1 and 2.

Acknowledgements. We would like to thank the anonymous referees for their stimulating comments which enhanced the paper and provided us with existing

results previously unknown to us. This work was supported by the Academy of Finland Grant 268703.

References

1. Cardoso, J.F., Souloumiac, A.: Blind beamforming for non-Gaussian signals. IEE Proceedings F - Radar and Signal Processing 140, 362–370 (1993)
2. Dermoune, A., Wei, T.: FastICA algorithm: Five criteria for the optimal choice of the nonlinearity function. IEEE Transactions on Signal Processing 61(8), 2078–2087 (2013)
3. Gómez-Sánchez-Manzano, E., Gómez-Villegas, M., Marín, J.: Sequences of elliptical distributions and mixtures of normal distributions. Journal of Multivariate Analysis 97(2), 295–310 (2006)
4. Huber, P.J.: Projection pursuit. The Annals of Statistics 13(2), 435–475 (1985)
5. Hyvarinen, A.: One-unit contrast functions for independent component analysis: A statistical analysis. In: Proceedings of the 1997 IEEE Workshop on Neural Networks for Signal Processing. pp. 388–397 (1997)
6. Hyvärinen, A.: Fast and robust fixed-point algorithms for independent component analysis. IEEE Transactions on Neural Networks 10(3), 626–634 (1999)
7. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons, New York, USA (2001)
8. Hyvärinen, A., Oja, E.: A fast fixed-point algorithm for independent component analysis. Neural Computation 9, 1483–1492 (1997)
9. Koldovskỳ, Z., Tichavskỳ, P., Oja, E.: Efficient variant of algorithm FastICA for independent component analysis attaining the Cramer-Rao lower bound. IEEE Transactions on Neural Networks 17(5), 1265–1277 (2006)
10. Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S.: Deflation-based FastICA with adaptive choices of nonlinearities. IEEE Transactions on Signal Processing 62(21), 5716–5724 (2014)
11. Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S., Virta, J.: The squared symmetric FastICA estimator. Signal Processing 131, 402 – 411 (2017)
12. Miettinen, J., Taskinen, S., Nordhausen, K., Oja, H.: Fourth moments and independent component analysis. Statistical Science 30, 372–390 (2015)
13. Nordhausen, K., Ilmonen, P., Mandal, A., Oja, H., Ollila, E.: Deflation-based FastICA reloaded. In: Proceedings of 19th European Signal Processing Conference. pp. 1854–1858 (2011)
14. Ollila, E.: The deflation-based FastICA estimator: Statistical analysis revisited. IEEE Transactions on Signal Processing 58(3), 1527–1541 (2010)
15. Palmer, J., Kreutz-Delgado, K., Rao, B.D., Wipf, D.P.: Variational EM algorithms for non-gaussian latent variable models. In: Advances in neural information processing systems. pp. 1059–1066 (2005)
16. Tichavskỳ, P., Koldovskỳ, Z., Oja, E.: Speed and accuracy enhancement of linear ICA techniques using rational nonlinear functions. In: International Conference on Independent Component Analysis and Signal Separation. pp. 285–292. Springer (2007)
17. Virta, J., Nordhausen, K., Oja, H.: Projection pursuit for non-Gaussian independent components (2016), submitted
18. Wei, T.: On the spurious solutions of the FastICA algorithm. In: IEEE Workshop on Statistical Signal Processing. pp. 161–164 (2014)