Non-square ICA based on the asymmetry

P. Spurek, J. Tabor, P. Rola, and A. Czechowski

Faculty of Mathematics and Computer Science, Jagiellonian University, Łojasiewicza 6, 30-348 Cracow, Poland
{przemyslaw.spurek,jacek.tabor}@ii.uj.edu.pl
Department of Mathematics of the Cracow University of Economics, Rakowicka 27, 31-510 Cracow, Poland
przemyslaw.rola@outlook.com
Dynniq B.V., Basicweg 16, 3821 BR Amersfoort, The Netherlands
{aleksander.czechowski}@dynniq.com

Abstract. In its basic form Independent Component Analysis (ICA) aims to find an invertible linear transformation so that the transformed data has independent components. In the case when number of sources is less than the number of sensors, the task has an additional step which consists of filtering the possible noise - in such situation we are looking for so-called non-square mixing matrix.
Due to computational constraints, principal component analysis is often used for dimension reduction prior to ICA (PCA+ICA). However, such approach commonly removes important information. In this paper we present a method based on split-gaussians **mozna pisac split-gaussians? troche nieformalnie, moze split-gaussian distributions?** which is dedicated for determining **a** non-square mixing matrix in ICA maximum likelihood framework. The main idea is based on separation of components which are as not-normal **not-normal brzmi dziwnie ale moze to tylko ja? as far from normal as possible?** as possible**przecinek** from the noise, which is assumed to be gaussian. Experiments show that our method obtains better or comparable **comparable or better** results to most state-of-the-art ICA approaches. *abstract* **usunac**

Keywords: ICA, PCA, Source separation

1 Introduction

Independent component analysis (ICA) is similar in many aspects to principal component analysis (PCA). In PCA we look for an orthonormal base in which the data components are not linearly dependent (uncorrelated), while in ICA we search for the coordinate system in which the components are independent. More precisely the aim of ICA is to transform the observed data $\mathbf{X}$ into maximally independent components $\mathbf{S}$ with use of an invertible linear transformation W, called the *transformation matrix*:

$$\mathbf{S} = W^T \mathbf{X}.$$

Popular ICA methodology does not directly attempt to find components that are independent but rather components that are as non-Gaussian as possible. This follows from the fact that one of the theoretical foundations of ICA is given by the dual view at the Central Limit Theorem [15], which states that the distribution of the sum (average or linear combination) of N **identically distributed?** independent random variables approaches Gaussian as $N \to \infty$. Obviously if all source variables are Gaussian, the ICA method will not work.

Another common approach to ICA based on the maximum likelihood estimation [23] is recently gaining popularity [14,26,28]. Then we search for the **a zamiast the** optimally fitted to data **moze usunac optimally fitted to data? troche dlugie zdanie** cordinate **coordinate** system B and marginal densities f_i such that the data density factors in base B as **are zamiast as** the product of maringal **marginal?** densities. To obtain an efficient method and avoid overfitting we have to restrict the marginal densities f_i to a class $\mathcal{F}$ of densities which has not too many parameters which can be easily estimated (clearly from obvious reasons this class has to be different from gaussians). **ostatnie zdanie jest za dlugie i nieczytelne. Wyrzuc which has not too many parameters, zrob ze zdania w nawiasie osobne zdanie, albo wyrzuc, nie uzywaj clearly i for – nie from – obvious reasons obok siebie – maslo maslane** As $\mathcal{F}$ we typically choose the super-Gaussian logistic density or other heavy tails distributions.

In many applications of ICA we deal with the case when several sensors measure the latents variables and the rest of them record only the noise. This happens when the number of sources is unknown and may be less than the number of sensors (then we are looking for so-called *non-square mixing matrix* W). Such a case **Such a situation aby uniknac powtorzen** is common for example in the identification of brain networks in functional magnetic resonance imaging (fMRI) [3,9]. In practice, most approaches deal with this problem by first applying PCA to the observations prior to classic ICA (PCA+ICA) to meet the assumption of square mixing**przecinek** and to reduce computational costs [14]. Although numerically effective, this approach may fail as it is not invariant with respect to linear transformation **transformations**, since PCA will find a "noise" component if it is sufficiently large. **Zdanie za dlugie i niezrozumiale, rozbij na dwa albo powyrzucaj co nie jest istotne**

The aim of this paper is to propose a new**przecinek** density based approach to deal with **zamien deal with na tackle** this case which does not have the above mentioned disadvantage **wyrzuc od which do konca zdanie, wiadomo o co chodzi** . Our idea is to join the two earlier mentioned approaches to solving ICA - one based on the search for non-gaussian components and the other based on density estimation - to deal with the case when the number of sources is smaller then that of sensors. Observe that the noise typically occurs as a sum of many independent factors, and consequently thanks to the central limit theorem **moze usun thanks to the central limit theorem, wszyscy wiedza a za duzo w zdaniu** it typically has approximately gaussian distribution **zamiast approximately gaussian distribution distribution close to gaussian**.

(a) Original images 42049 and 220075 and Gaussian nice. **noise**

(b) $\text{ICA}_{\mathcal{SN}}$.

(c) FastICA.

(d) ProDenICA.

(e) NGPP.

(f) $\text{ICA}_{\mathcal{SN}}$.

(g) FastICA.

(h) ProDenICA.

(i) NGPP.

Fig. 1. Comparison of images separation by our method ($\text{ICA}_{\mathcal{SN}}$), with FastICA, ProDenICA, NGPP. Before mixing by a linear matrix, we added to the first two components given in (a) the third component given by random normal noise. As we see $\text{ICA}_{\mathcal{SN}}$ was able to perfectly recover the two first components.

This is why one typically makes the assumption [8] that

noise components are comming from a gaussian noise.

czy to jakas prawda objawiona? wrzuc noise components.. normalnie w zdanie a nie wyroznione jak rownanie Following the density approach**przecinek** to filter them out we fit the first d-components from a class of $\mathcal{F}$ of densities which is broader then gaussians, while the rest from the gaussians $\mathcal{N}$ (the final choice of the value of parameter $d \in \{1, \ldots, D\}$ can be decided by applying either AIC or BIC criterion). Following [28]**przecinek** as $\mathcal{F}$ we take the class $\mathcal{SN}$ of split-normal densities.

Our experiments show, which is illustrated by Figure 1 **wyrzuc, wsadz (cf. Figure 1)**, that $\text{ICA}_{\mathcal{SN}}$ works as desired**przecinek** and effectively removes the components which contains **contain** gaussian noise. However, we cannot objectively conclude that it is better as compared to other state-of-the-art approaches, since the experiment was conducted in the setting optimal to our method as we assumed that the noise was gaussian. **Musimy sie tak kajac? Przeciez stwierdzamy ze jest lepsza w usuwaniu gaussian noise, wiec ta klaryfikacja ze ogolnie moze nie byc lepsza to troche na wyrost. Spytaj Jacka ewentualnie**

2 Related works

In literature exist a few **jesli masz na mysli malo, to few. a few to duzo** approaches dedicated for non-square ICA problem. Most of such method are dedicated for **to zamiast for** individual methods **dwa razy method w tym zdaniu, nie wiadomo o co chodzi**. Attias and Schreiner [2] derived a likelihood based algorithm for separation of general sequences with a frequency domain implementation. Belouchrani and Cardoso [6] presented a general likelihood approach allowing for additive noise and for non-square mixing matrices. They applied the method to separation of sources taking discrete values**przecinek** and estimated the mixing matrix using an Expectation-Maximization (EM) approach with both a deterministic and a stochastic formulation. In [22] **the** authors used the EM approach for separation of autocorrelated sequences in presence of noise and explored a family of flexible source signal priors based on Gaussian Mixtures.

The assumption is that **usun is that** of square mixing is mostly unrealistic in the case of EEG ant FMRI **przecinek** where the number of sources is less than the number of electrodes [4,25,27]. Therefore, many of **usun of** algorithms dedicated for **to zamiast for** this task use a probabilistic ICA [29]. The noisy ICA model can be approximated using a variant of PCA+ICA [4], where probabilistic PCA is used to estimate the number of components and achieve dimension reduction **moze po prostu reduce dimension** [29]. In [1] **the** authors developed stochastic EM algorithms to estimate the noisy model **przecinek** and proposed parametric methods.

Other methods exploring non-Gaussian structure in multivariate data include non-Gaussian component analysis (NGCA) and projection pursuit [7,18]. NGCA is a more general case of linear non-Gaussian component analysis (LNGCA) [24] that allows non-linear dependence between the non-Gaussian components.

In the paper [21] **the** authors propose a novel **wiadomo ze novel, zamiast a novel po prostu an** adaptive twostage **two-stage?** deflation-based FastICA algorithm **przecinek** that allows one to use different nonlinearities for different components **przecinek** and optimizes the order in which the components are extracted.

3 Theoretical foundations of ICA$_{\mathcal{SN}}$

In this section we present **the** theoretical foundations of the method. We begin with the statement of the problem, next we focus our attention on the presentation of **zamiast tego wszystkiego, next we present** the class of densities we discuss **used zamiast we discuss**. Last **przecinek** we show **compute zamiast show** the gradient of the method **przecinek** which is needed in the optimization procedure.

3.1 Statement of the problem

Since this general **the zamiast this general** idea of the search for ICA **for independent components zamiast for ICA** with the use of maximum likeli-

hood is essential in our further considerations, for the convenience of the reader we first describe it briefly. Assume that the random vector $\mathbf{X}$ in $\mathbb{R}^D$ has the density function $F(\mathrm{x})$. Suppose that the components of $\mathbf{X}$ are not independent, but that we know (or suspect) that **moze usunac we know or suspect that?** there is a basis B (we put $W^T = B^{-1}$) such that in that base the components of $\mathbf{X}$ become independent. Observe **przecinek** that where **then zamiast where** $\omega_i^T \mathrm{x}$ is the i-th coefficient of x in the basis B (ω_i denotes the i-th column of W), and therefore there exist densities $f_1, \ldots, f_D$ such that

$$F(\mathrm{x}) = \det(W) \cdot f_1(\omega_1^T \mathrm{x}) \cdot \ldots \cdot f_d(\omega_d^T \mathrm{x}). \tag{1}$$

Given W and densities $(f_i)_{i=1}^D$ we introduce notation to represent RHS **we denote the right-hand side** of the above equation **as follows**:

$$F_W(f_1, \ldots, f_D)(\mathrm{x}) = \det(W) \cdot f_1(\omega_1^T \mathrm{x}) \cdot \ldots \cdot f_d(\omega_d^T \mathrm{x}).$$

Thus we may **Let us now zamiast Thus we may** state the density based formulation of ICA in the case we have only a sample X from random vector $\mathbf{X}$.

THE? GENERAL ICA PROBLEM (**the** maximum likelihood formulation).
Find densities f_i and matrix W, so that F given by (1) *optimally fits the data $X = (\mathrm{x}_i)$ with respect to the likelihood, that is that the value*

$$\sum_i \log F_W(f_1, \ldots, f_D)(\mathrm{x}_i)$$

is maximized.

Since the search over the space of all densities is not feasible, and could lead to overfitting, we naturally have to reduce to a subclass of all densities on $\mathbb{R}$ parametrized by a finite amount of parameters. Clearly, since ICA does not work if the data are gaussian, we have to choose a family $\mathcal{F}$ of densities which is distant from Gaussian ones.

ICA FOR DENSITY CLASS $\mathcal{F}$.
*Find **a** matrix W and densities $f_1, \ldots, f_D \in \mathcal{F}$, such that the value of*

$$\sum_i \log F_W(f_1, \ldots, f_D)(\mathrm{x}_i)$$

is maximized.

Similarly to [28] as a class **usun a class** $\mathcal{F}$ we are going to take the class of split-gaussians, as as it is easy to deal with (small number of parameters) and is resistant to outliers[1].

[1] The reason is that split gaussians **moze split gaussian distributions**, instead at fitting the distribution with respect to heavy tails, fits the asymmetry of the data. **troche to zdanie ciezkie. Moze po prostu split gaussians are fitting well to asymmetrical data?**

As mentioned in the introduction, we assume that components which we would like to filter-out, **tu bez przecinka** are coming from a gaussian noise, and the **then zamiast the** aim it to fit the first d-components from a larger class of densities, while the rest from the gaussians $\mathcal{N}$.**Poprzednie zdanie jest troche niezrozumiale, przepisz prosze (moze rozbij na dwa).** Thus our final problem can be stated as follows.

ICA FOR DENSITY CLASS $\mathcal{F}$ WITH d SOURCES.
Find **a** *matrix* W*, densities* $f_1, \dots, f_d \in \mathcal{F}$ *and normal densities* $f_{d+1}, \dots, f_D \in \mathcal{N}$*, so that the value of*

$$\sum_i \ln F_W(f_1, \dots, f_D)(\mathrm{x}_i)$$

is maximized.

Observe that the solution to the above problem is linearly invariant, that is if W is optimal for X an A is linear, then W_A is optimal for AX, where $W_A = (A^{-1})^T W$.

The continuous version of the condition we maximize in the case we know the density f of the random variable $\mathbf{X}$ limits to

$$\int \ln F_W(f_1, \dots, f_D)(\mathrm{x}) f(\mathrm{x}) d\mathrm{x} = -H(f, F_W(f_1, \dots, f_D)),$$

where the cross entropy $H(f, g)$ is given by the sum of entropy $H(f)$ and Kullback-Leibler divergence $D_{KL}(f, g)$. Thus the continuous version of the ICA problem with d sources reduces to the minimization of

$$D_{KL}(f, F_W(f_1, \dots, f_D))$$

over all matrices W and densities $f_1, \dots, f_D \in \mathcal{F}$. Since for fixed f Kullback-Leibler divergence is minimized for $g = f$, we arrive at the following result, which says that in the ideal case by the discussed approach we restore the unmixing matrix if it exists.

Theorem 1. *Let* F *be a density such that there exist* **a** *matrix* $\overline{W}$ *and densities*

$$\hat{f}_1, \dots, \hat{f}_d \in \mathcal{F} \text{ and } \hat{f}_{d+1}, \dots, \hat{f}_D \in \mathcal{N}$$

such that

$$F = F_{\overline{W}}(\hat{f}_1, \dots, \hat{f}_D).$$

Then

$$\overline{W}, \hat{f}_1, \dots, \hat{f}_D = \operatorname{argmin}\{F_W(f_1, \dots, f_D) : W, f_1, \dots, \bar{f}_d \in \mathcal{F}, f_{d+1}, \dots, f_D \in \mathcal{N}\}.$$

3.2 Split normal distribution

In this section we discuss the class $\mathcal{F}$ we will use in our final algorithm of $\mathrm{ICA}_{\mathcal{SN}}$. The density of $\mathcal{SN}$, the one-dimensional split normal distribution [30], is given by the formula

$$\mathcal{SN}(x; m, \sigma^2, \tau^2) = \begin{cases} c \cdot \exp[-\frac{1}{2\sigma^2}(x-m)^2], x \leq m \\ c \cdot \exp[-\frac{1}{2\tau^2\sigma^2}(x-m)^2], x > m \end{cases}$$

where $c = \sqrt{\frac{2}{\pi}}\sigma^{-1}(1+\tau)^{-1}$.

As we see **przecinek** the split normal distribution comes from merging two opposite halves of two normal distributions in their common mode. The main advantage of **in using zamiast of** split normal distributions over normal one **moze over regular ones?** is that it **they zamiast it** allows **allow zamiast allows** data asymmetry. In 1982 John [16] showed **Moze It was shown by John.. i bez pisania roku** that the likelihood function can be expressed in a form in which the scale parameters σ and τ are an explicit function of the location parameter m. In the case when $\mathcal{F} = \mathcal{SN}$ the density class considered in the previous subsection is given in the explicit form by the following observation.

Observation 31 ***Czemu Observation 31? Przenumeruj to jakos*** *A density of the multivariate split normal d and normal $D-d$ distribution is given by*

$$\mathcal{SN}_d\mathcal{N}_{D-d}(\mathrm{x}; \mathrm{m}, W, \sigma^2, \tau^2) =$$

$$\det(W) \prod_{j=1}^{d} \mathcal{SN}(\omega_j^T(\mathrm{x}-\mathrm{m}); 0, \sigma_j^2, \tau_j^2) \cdot \prod_{j=d+1}^{D} \mathcal{N}(\omega_j^T(\mathrm{x}-\mathrm{m}); 0, \sigma_j^2),$$

where ω_j is the j-th column of non-singular matrix W, $\mathrm{m} = (m_1, \ldots, m_d)^T$, $\sigma = (\sigma_1, \ldots, \sigma_d)$ and $\tau = (\tau_1, \ldots, \tau_{D-d})$.

Observe that the above density probability function has mode in m. As a consequence of result of John [16] we can maximize the likelihood of the above function on data X with respect to σ and τ.

Theorem 2. *Let $\mathrm{x}_1, \ldots, \mathrm{x}_n$ be given, and let $\mathrm{m} \in \mathbb{R}^D$ and matrix W be fixed. Then the likelihood maximized w.r.t. σ and τ is*

$$\hat{L}(X; \mathrm{m}, W) = \frac{2^{(d-D/2)n} n^{dn/2}}{(\pi e)^{Dn/2}} \left(\frac{1}{|\det(W)|^{\frac{2}{3}}} \prod_{j=1}^{d} g_j(\mathrm{m}, W) \right)^{-3n/2} \left(\prod_{j=d+1}^{D} \frac{(s_1+s_2)}{n} \right)^{-n/2}, \tag{2}$$

where

$$g_j(\mathrm{m}, W) = s_{1j}^{1/3} + s_{2j}^{1/3},$$
$$s_{1j} = \sum_{i \in I_j} [\omega_j^T(\mathrm{x}_i - \mathrm{m})]^2, I_j = \{i \colon \omega_j^T(\mathrm{x}_i - \mathrm{m}) \leq 0\},$$
$$s_{2j} = \sum_{i \in I_j^c} [\omega_j^T(\mathrm{x}_i - \mathrm{m})]^2, I_j^c = \{i \colon \omega_j^T(\mathrm{x}_i - \mathrm{m}) > 0\},$$

and the maximum likelihood estimators of σ_j^2 *and* τ_j *are*

$$\begin{aligned}\hat{\tau}_j(\mathrm{m}, W) &= \left(\frac{s_{2j}}{s_{1j}}\right)^{1/3}, \qquad 1 \leq j \leq d \\ \hat{\sigma}_j^2(\mathrm{m}, W) &= \begin{cases} \frac{1}{n} s_{1j}^{2/3} g_j(\mathrm{m}, W), & 1 \leq j \leq d \\ \frac{1}{n}(s_{1j} + s_{2j}), & d < j \leq D \end{cases}\end{aligned} \tag{3}$$

Proof. See Section 5 (Appendix A).

Thanks to the above theorem we can reduce the search for the maximum of the log-likelihood function for two parameters $\mathrm{m} \in \mathbb{R}^d$ and $W \in \mathcal{M}(\mathbb{R}^d)$.

$$l(X; \mathrm{m}, W) = \frac{1}{|\det(W)|^{\frac{2}{3}}} \prod_{j=1}^{d} g_j(\mathrm{m}, W) \prod_{j=d+1}^{D} (s_{1j} + s_{2j})^{\frac{1}{3}} \tag{4}$$

where w_j stands for the j-th column of matrix W. Consequently, maximization of likelihood function is equivalent to minimization of $\ln l$.

Corollary 1. *Let* $X \subset \mathbb{R}^d$, $\mathrm{m} \in \mathbb{R}^d$, $W \in \mathcal{M}(\mathbb{R}^d)$ *be given, then*

$$\operatorname*{argmax}_{\mathrm{m},W} \hat{L}(X; \mathrm{m}, W) = \operatorname*{argmin}_{\mathrm{m},W} \ln l(X; \mathrm{m}, W).$$

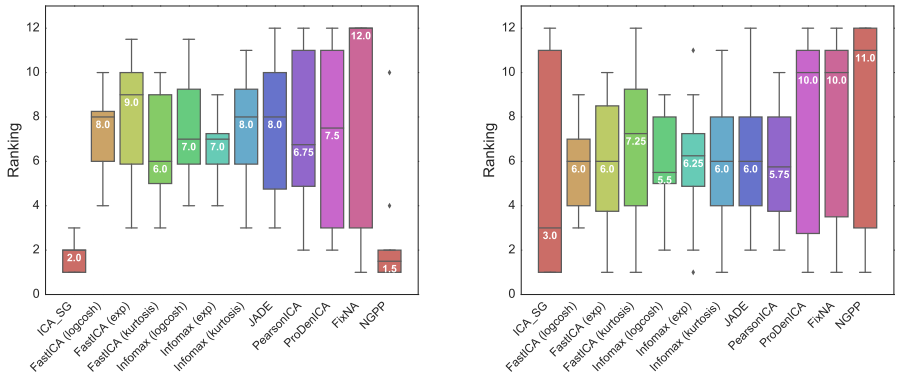
To minimize $\ln l$ with the use classical gradient descent method **moze standard gradient methods? We wtyczce uzywamy BFGSa na przyklad to jakas wariacja gradient descentu**, we need the formula for $\nabla \ln l$ (**the** gradient of the cost function).

Theorem 3. *Let* $X \subset \mathbb{R}^d$, $\mathrm{m} = (\mathrm{m}_1, \ldots, \mathrm{m}_d)^T \in \mathbb{R}^d$, $W = (\omega_{ij})_{1 \leq i,j \leq d}$ *non-singular be given. Then* $\nabla_{\mathrm{m}} \ln l(X; \mathrm{m}, W) = \left(\frac{\partial \ln l(X;\mathrm{m},W)}{\partial \mathrm{m}_1}, \ldots, \frac{\partial \ln l(X;\mathrm{m},W)}{\partial \mathrm{m}_d}\right)^T$, *where*

$$\frac{\partial \ln l(X;\mathrm{m},W)}{\partial \mathrm{m}_k} = \sum_{j=1}^{d} \frac{-2}{3(s_{1j}^{\frac{1}{3}} + s_{2j}^{\frac{1}{3}})} \left(\frac{1}{s_{1j}^{\frac{2}{3}}} \sum_{i \in I_j} \omega_j^T(\mathrm{x}_i - \mathrm{m})\omega_{jk} + \frac{1}{s_{2j}^{\frac{2}{3}}} \sum_{i \in I_j^c} \omega_j^T(\mathrm{x}_i - \mathrm{m})\omega_{jk} \right) +$$
$$\sum_{j=d+1}^{D} \frac{-2}{3(s_{1j} + s_{2j})} \cdot \left(\sum_{i \in I_j} \omega_j^T(\mathrm{x}_i - \mathrm{m})\omega_{jk} + \sum_{i \in I_j^c} \omega_j^T(\mathrm{x}_i - \mathrm{m})\omega_{jk} \right).$$

Moreover, $\nabla_W \ln l(X; \mathrm{m}, W) = \left[\frac{\partial \ln \tilde{l}(X;\mathrm{m},W)}{\partial \omega_{pk}}\right]_{1 \leq p,k \leq d}$, *where*

$$\frac{\partial \ln l(X;\mathrm{m},W)}{\partial \omega_{pk}} = -\frac{2}{3}(\omega^{-1})_{pk}^T +$$
$$\frac{2}{3(s_{1p}^{\frac{1}{3}} + s_{2p}^{\frac{1}{3}})} \left(s_{1p}^{-\frac{2}{3}} \sum_{i \in I_p} \omega_p^T(\mathrm{x}_i - \mathrm{m})(\mathrm{x}_{ik} - \mathrm{m}_k) + s_{2p}^{-\frac{2}{3}} \sum_{i \in I_p^c} \omega_p^T(\mathrm{x}_i - \mathrm{m})(\mathrm{x}_{ik} - \mathrm{m}_k) \right) +$$
$$\frac{2}{3(s_{1p} + s_{2p})} \left(\sum_{i \in I_p} \omega_p^T(\mathrm{x}_i - \mathrm{m})(\mathrm{x}_{ik} - \mathrm{m}_k) + \sum_{i \in I_p^c} \omega_p^T(\mathrm{x}_i - \mathrm{m})(\mathrm{x}_{ik} - \mathrm{m}_k) \right),$$



(a) Results of ICA methods in the case of mixture of images and Gaussian noise.

(b) Results of ICA methods in the case of mixture of images and salt and pepper **"salt and pepper"** noise.

Fig. 2. Results of ICA methods in the case of mixture of images and noise. **Moze nie powtarzac, jest juz w subcaptions to samo**

and

$$s_{1j} = \sum_{i \in I_j} [\omega_j^T (\mathrm{x}_i - \mathrm{m})]^2, I_j = \{1 \leq i \leq n \colon \omega_j^T (\mathrm{x}_i - \mathrm{m}) \leq 0\},$$
$$s_{2j} = \sum_{i \in I_j^c} [\omega_j^T (\mathrm{x}_i - \mathrm{m})]^2, I_j^c = \{1 \leq i \leq n \colon \omega_j^T (\mathrm{x}_i - \mathrm{m}) > 0\}.$$

Proof. See Section 6 (Appendix B).

Thanks to the above Theorem we are able to use in our experiments the gradient descent for finding the minimum of our cost function.

4 Experiments

To compare $\mathrm{ICA}_{\mathcal{SN}}$ to other state-of-the-art approaches we use Tucker's congruence coefficient [19] which values range between -1 and $+1$. It can be used to study the similarity of extracted factors across different samples. Generally, a congruence coefficient of 0.9 indicates a high degree of factor similarity, while a coefficient of 0.95 or higher indicates that the factors are virtually identical.

We evaluate our method in the context of 2D and hyperspectral images. For comparison we use R package `ica` [11], `PearsonICA` [17], `ProDenICA` [10], `tsBSS` [20], `NGPP` [31]. The most popular method used in practice is FastICA [13,12] algorithm, which uses negentropy. In this context we can use three different functions to estimate neg-entropy: logcosh, exp and kurtosis. We also compare our method with algorithm using Information-Maximization (Infomax) approach [5]. Similarly to FastICA we consider three possible non-linear functions: hyperbolic tangent, logistic and extended Infomax.

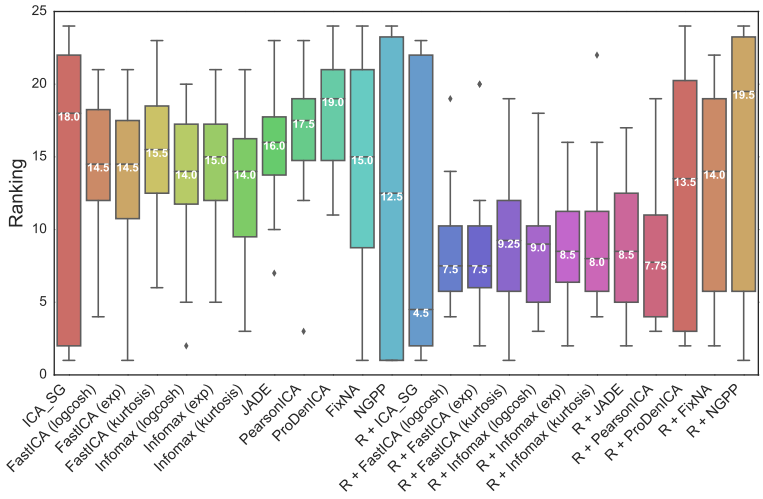


Fig. 3. Results of ICA methods in the case of mixture of images and noise salt and pepper noise with scale (R+) and without preprocessing.

	ICA_{SN}	FastICA logcosh	FastICA exp	FastICA kurtosis	Infomax tanh	Infomax tangent	Infomax logistic	JADE	PearsonICA	ProDenICA	FixNA	NGPP
Source 1	0.3427	0.2778	0.2820	0.2865	0.2823	0.2731	0.2979	0.2901	0.2837	0.2065	0.1827	0.1820
Source 2	0.3212	0.4122	0.3954	0.3690	0.4168	0.4193	0.1972	0.3611	0.2679	0.1318	0.3020	0.3176
Source 3	0.3180	0.1857	0.1878	0.1285	0.1845	0.1819	0.3782	0.0053	0.0252	0.2565	0.0426	0.2962
Source 4	0.2822	0.0357	0.0364	0.1743	0.0339	0.0208	0.0317	0.1780	0.2802	0.1724	0.3196	0.1768

Table 1. Tucker's congruence coefficients between average edges form reference layers and various ICA results.

4.1 Separation of images

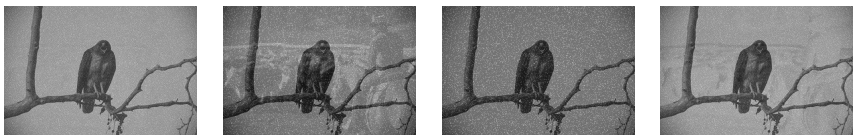
One of the most popular application of ICA is the separation of images. In our experiments we use four images from the USC-SIPI Image Database of size 256×256 pixels (4.1.01, 4.1.06, 4.1.02, 4.1.03) and eight of size 512×512 pixels (4.2.04, 4.2.02, boat.512, elaine.512, 5.2.10, 5.2.08, 5.3.01, 4.2.03). We also use 8 images from the Berkeley Segmentation Dataset of size 482×321 with indexes (#119082, #42049, #43074, #38092, #157055, #220075, #295087, #167062).

We make random pairs of above images and one component with noise (random sample from Gaussian distribution $\mathcal{N}(0, 1)$) and use them as a source signal combined by the mixing matrix $A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \end{bmatrix}$. Our goal was to reconstruct two original images by using only the knowledge about mixed ones. The visualization of this process we present in Fig. 1.

(a) Original images 42049 and 220075 and Gaussian nice **noise**.

(b) Mixture of sources after rescaling.

(c) $\text{ICA}_{\mathcal{SN}}$. (d) FastICA. (e) ProDenICA. (f) NGPP.



(g) $\text{ICA}_{\mathcal{SN}}$. (h) FastICA. (i) ProDenICA. (j) NGPP.

Fig. 4. Comparison of images separation by our method ($\text{ICA}_{\mathcal{SN}}$), with FastICA, ProDenICA, NGPP. **Ale to jest to samo co Fig.1 – po co powtarzac?**

As a summary from the experiment, in Figure 2(a) we present a boxplot **czy miales na mysli box plots?** of **the** ranks obtained by the methods. The $\text{ICA}_{\mathcal{SN}}$ almost perfectly recovered source signal and it is one of the best in ranking. Although this is not surprising as the experiments were in fact conducted in the setting which favored our approach, as we chose the noise to be gaussian, **Znowu sie kajamy, czy potrzebnie to spytaj Jacka. Ale jakos niezgrabnie?** this shows that $\text{ICA}_{\mathcal{SN}}$ works as desired and deals well with removing **any** gaussian components from the data.

In the next experiments we reaped **repeated?** the procedure with salt and pepper **"salt and pepper"** noise. But **usun But** before applying ICA methods we rescale images, see Fig. 4. Such procedure can be understand as a preprocessing in the ICA framework. After rescaling of images non-gaussian noise become more normal and therefore we are able to remove them. In Figure 2(a) we present

a boxplot **box plot? box plots? box-plot?** of ranks obtained on rescale data **on the rescaled data**.

The rescaling of images give sensational **lepiej vastly zamiast sensational** better results in the case of non-gaussian noise end increase **and increases the** score of all methods, see Figure 3.

4.2 Hyperspectral Unmixing

Independent component analysis has been recently **can be zamiast has been recently?** applied into **to zamiast into** hyperspectral unmixing [32] as a result of **lepiej due to** its low computation time and its ability to perform without prior information. In this subsection we apply **present a zamiast apply** simple example which suggests that our method also can by used for spectral data.

Urban data [33,35,34] is one of the most widely used hyperspectral data-sets used **usun used** in the hyperspectral **usun hyperspectral** unmixing study. Each image has 307×307 pixels, each of which corresponds to a 2×2 m area. In this image, there are 210 wavelengths ranging from 400 nm to 2500 nm, resulting in a spectral resolution of 10 nm. After the channels 1–4, 76, 87, 101–111, 136–153 and 198–210 are removed (due to dense water vapor and atmospheric effects), there remain 162 channels (this is a common preprocess for hyperspectral unmixing analyses). There is ground truth [33,35,34], which contains 4 channels: #1 Asphalt, #2 Grass, #3 Tree and #4 Roof.

A highly mixed area is cut from the original data set in this experiment (similar example was showed in [32]), with the size of 200×150 pixels.

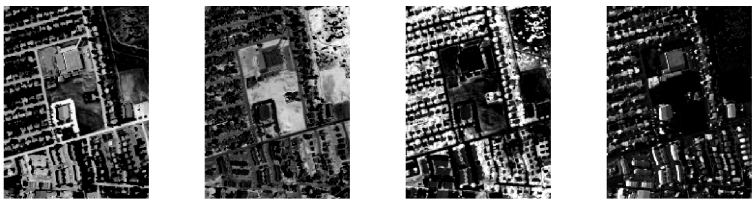
In our experiment we compared $\mathrm{ICA}_{\mathcal{SN}}$ to other popular ICA methods, see Fig. 5. Observe that $\mathrm{ICA}_{\mathcal{SN}}$, NGPP and ProDenICA give layers which seem to contain more information then FastICA, as the last component in FastICA contains mainly noise. To verify which method separate sources from noise better we calculate how obtained layers corresponds to original signals. Since there is no classical measure for such task **przecinek** we verified it by calculation **calculating the** correlation coefficient between **the** average muse **co to jest muse? moze literowka a moze ja nie znam slowa** of edges of reference layers (in our experiment we use Canny edge detector **zrob z tego osobne zdanie**) and ICA results, see Tab. 4.1. Our method gives similar **comparable zamiast similar** value of similarity measure on all layers.

5 Appendix A

Proof (Proof of Theorem 2.). **Zrob tak zeby sie proof nie powtarzal 2x w naglowku w tym i innych twierdzeniach w appendixie** Let $X = \{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$. We write

$$\mathrm{z}_i = W(\mathrm{x}_i - m), \quad \mathrm{z}_{ij} = \omega_j^T(\mathrm{x}_i - m),$$

for observation i, where $i = 1, \ldots, n$ and coordinates $j = 1, \ldots, d$.



(a) Ground truth layers which contains 4 channels: #1 Asphalt, #2 Grass, #3 Tree and #4 Roof.

(b) The effect of the $\text{ICA}_{\mathcal{SN}}$ method.

(c) The effect of the FastICA (logcosh) method.

(d) The effect of the ProDenICA method.

(e) The effect of the NGPP method.

Fig. 5. Results of image separation with the uses of various ICA algorithms.

Let us consider the likelihood function, i.e.

$$L(X;\mathrm{m},W,\sigma,\tau)=\prod_{i=1}^{n} SN_d N_{D-d}(\mathrm{x}_i;\mathrm{m},W,\sigma^2,\tau^2)$$
$$=\prod_{i=1}^{n}|\det(W)|\prod_{j=1}^{d} SN(\omega_j^T(\mathrm{x}_i-\mathrm{m});0,\sigma_j^2,\tau_j^2)\cdot\prod_{j=d+1}^{D} N(\omega_j^T(\mathrm{x}_i-\mathrm{m});0,\sigma_j^2)=$$
$$\Big(c_1|\det(W)|\Big)^n\Big(\prod_{j=1}^{d}\sigma_j(1+\tau_j)\Big)^{-n}\prod_{i=1}^{n}\prod_{j=1}^{d}\exp\Big[-\tfrac{1}{2\sigma_j^2}z_{ij}^2(\mathbb{1}_{\{z_{ij}\leq 0\}}+\tau_j^{-2}\mathbb{1}_{\{z_{ij}>0\}})\Big]$$
$$\Big(\prod_{j=d+1}^{D}\sigma_j\Big)^{-n}\prod_{i=1}^{n}\prod_{j=d+1}^{D}\exp\Big[-\tfrac{1}{2\sigma_j^2}z_{ij}^2\Big],$$

where $c_1=\Big(\sqrt{\frac{2}{\pi}}\Big)^d\cdot\Big(\frac{1}{\sqrt{2\pi}}\Big)^{D-d}$. Now we take the log-likelihood function, i.e.

$$\ln(L(X;\mathrm{m},W,\sigma,\tau))=\ln\Bigg(\Big(c_1|\det(W)|\Big)^n\Big(\prod_{j=1}^{d}\sigma_j(1+\tau_j)\Big)^{-n}\Big(\prod_{j=d+1}^{D}\sigma_j\Big)^{-n}\Bigg)+$$
$$\sum_{i=1}^{n}\sum_{j=1}^{d}\Big[-\tfrac{1}{2\sigma_j^2}z_{ij}^2(\mathbb{1}_{\{z_{ij}\leq 0\}}+\tau_j^{-2}\mathbb{1}_{\{z_{ij}>0\}})\Big]+\sum_{i=1}^{n}\sum_{j=d+1}^{D}\Big[-\tfrac{1}{2\sigma_j^2}z_{ij}^2\Big]$$
$$=\ln\Bigg(\Big(c_1|\det(W)|\Big)^n\Big(\prod_{j=1}^{d}(1+\tau_j)\Big)^{-n}\Big(\prod_{j=1}^{D}\sigma_j\Big)^{-n}\Bigg)-$$
$$\tfrac{1}{2}\sum_{j=1}^{d}\Big(\sigma_j^{-2}\sum_{i\in I_j}z_{ij}^2+\tfrac{\sigma_j^{-2}}{\tau_j^2}\sum_{i\in I_j^c}z_{ij}^2\Big)-\tfrac{1}{2}\sum_{j=d+1}^{D}\sigma_j^{-2}\Big(\sum_{i\in I_j}z_{ij}^2+\sum_{i\in I_j^c}z_{ij}^2\Big)$$
$$=\ln\Bigg(\Big(c_1|\det(W)|\Big)^n\Big(\prod_{j=1}^{d}(1+\tau_j)\Big)^{-n}\Big(\prod_{j=1}^{D}\sigma_j\Big)^{-n}\Bigg)-$$
$$\sum_{j=1}^{d}\tfrac{1}{2\sigma_j^2}\Big(s_{1j}+\tfrac{1}{\tau_j^2}s_{2j}\Big)-\sum_{j=d+1}^{D}\tfrac{1}{2\sigma_j^2}\Big(s_{1j}+s_{2j}\Big).$$

We fix m, W and maximize the log-likelihood function over τ and σ. In such a case **zamiast In such a case uzyj Consequently** we have to **zamiast we have to uzyj we need to** solve the following system of equations

$$\frac{\partial\ln(L(X;\mathrm{m},W,\sigma,\tau))}{\partial\sigma_j}=0,\qquad\frac{\partial\ln(L(X;\mathrm{m},W,\sigma,\tau))}{\partial\tau_j}=0,$$

for $j=1,\ldots,D$. Hence **It follows that zamiast Hence**

$$-\tfrac{n}{\sigma_j}+\sigma_j^{-3}(s_{1j}+\tau_j^{-2}s_{2j})=0,\ \text{for } j=1,\ldots,d,$$
$$-\tfrac{n}{\sigma_j}+\sigma_j^{-3}(s_{1j}+s_{2j})=0,\ \text{for } j>d,$$
$$-\tfrac{n}{1+\tau_j}+\tfrac{s_{2j}}{\tau_j^3\sigma_j^2}=0,\ \text{for } j=1,\ldots,d.$$

By simple calculations we obtain the expressions for the estimators in 3. Substituting it into the log-likelihood function, we get

$$\hat{L}(\mathrm{m},W)=\Big(\tfrac{2}{\pi}\Big)^{\frac{dn}{2}}\Big(\tfrac{1}{2\pi}\Big)^{\frac{(D-d)n}{2}}|\det(W)|^n\Big(\prod_{j=1}^{d}\tfrac{1}{\sqrt{n}}g_j(\mathrm{m},W)^{\frac{3}{2}}\Big)^{-n}e^{-\frac{dn}{2}}\Big(\prod_{j=d+1}^{D}(\tfrac{s_{1j}+s_{2j}}{n})^{\frac{1}{2}}\Big)^{-n}=$$
$$\frac{2^{(d-D/2)n}n^{dn/2}}{(\pi e)^{Dn/2}}\cdot\Big(\frac{1}{|\det(W)|^{\frac{2}{3}}}\prod_{j=1}^{d}g_j(\mathrm{m},W)\Big)^{-\frac{3n}{2}}\Big(\prod_{j=d+1}^{D}\frac{(s_{1j}+s_{2j})}{n}\Big)^{-\frac{n}{2}}$$

6 Appendix B

We will need the following well-known lemma (for the convenience of the reader we provide the **a zamiast the** proof).

Lemma 1. *Let $A = (a_{ij})_{1\leq i,j\leq d}$ be a differentiable map from real numbers to $d \times d$ matrices then*

$$\frac{\partial \det(A)}{\partial a_{ij}} = \mathrm{adj}^T(A)_{ij}, \tag{5}$$

where $\mathrm{adj}(A)$ *stands for the adjugate of A, i.e. the transpose of the cofactor matrix.*

Proof. By the Laplace expansion **formula na wyznacznik jako equation w osobnej linijce** $\det A = \sum_{j=1}^{d}(-1)^{i+j}a_{ij}M_{ij}$ **przecinek** where M_{ij} is the minor of the entry in the i-th row and j-th column. Hence

$$\frac{\partial \det A}{\partial a_{ij}} = (-1)^{i+j}M_{ij} = \mathrm{adj}^T(A)_{ij}.$$

Proof (Proof of Theorem 3.). Let us start with the partial derivative of $\ln(l)$ with respect to m. We have

$$\begin{aligned}
\tfrac{\partial \ln l(X;\mathrm{m},W)}{\partial \mathrm{m}_k} &= \sum_{j=1}^{d} \tfrac{\partial \ln(g_j(\mathrm{m},W))}{\partial \mathrm{m}_k} + \sum_{j=d+1}^{D} \tfrac{\partial \ln((s_{1j}+s_{2j})^{\frac{1}{3}})}{\partial \mathrm{m}_k} \\
&= \sum_{j=1}^{d} \tfrac{1}{s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}}} \tfrac{\partial (s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}})}{\partial \mathrm{m}_k} + \sum_{j=d+1}^{D} \tfrac{1}{(s_{1j}+s_{2j})^{\frac{1}{3}}} \tfrac{\partial ((s_{1j}+s_{2j})^{\frac{1}{3}})}{\partial \mathrm{m}_k} \\
&= \sum_{j=1}^{d} \tfrac{1}{s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}}} \left(\tfrac{1}{3s_{1j}^{\frac{2}{3}}} \tfrac{\partial s_{1j}}{\partial \mathrm{m}_k} + \tfrac{1}{3s_{2j}^{\frac{2}{3}}} \tfrac{\partial s_{2j}}{\partial \mathrm{m}_k} \right) + \sum_{j=d+1}^{D} \tfrac{1}{(s_{1j}+s_{2j})^{\frac{1}{3}}} \tfrac{1}{3} \tfrac{1}{(s_{1j}+s_{2j})^{\frac{2}{3}}} \left(\tfrac{\partial s_{1j}}{\partial \mathrm{m}_k} + \tfrac{\partial s_{2j}}{\partial \mathrm{m}_k} \right).
\end{aligned}$$

Now, we need $\frac{\partial s_{1j}}{\partial \mathrm{m}_k}$ and $\frac{\partial s_{2j}}{\partial \mathrm{m}_k}$, therefore **Napisz zamiast tego The derivatives .. are given by i wypisz wzory bez posrednich obliczen**

$$\tfrac{\partial s_{1j}}{\partial \mathrm{m}_k} = \sum_{i\in I_j} \tfrac{\partial [\omega_j^T(\mathrm{x}_i-\mathrm{m})]^2}{\partial \mathrm{m}_k} = \sum_{i\in I_j} 2\omega_j^T(\mathrm{x}_i-\mathrm{m}) \tfrac{\partial \omega_j^T(\mathrm{x}_i-\mathrm{m})}{\partial \mathrm{m}_k} = \sum_{i\in I_j} -2\omega_j^T(\mathrm{x}_i-\mathrm{m})\omega_{jk}.$$

Analogously we get $\frac{\partial s_{2j}}{\partial \mathrm{m}_k} = \sum_{i\in I_j^c} -2\omega_j^T(\mathrm{x}_i-\mathrm{m})\omega_{jk}$. Hence

$$\begin{aligned}
\tfrac{\partial \ln l}{\partial \mathrm{m}_k} = \sum_{j=1}^{d} \tfrac{-1}{s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}}} \left(\tfrac{1}{3s_{1j}^{\frac{2}{3}}} \sum_{i\in I_j} 2\omega_j^T(\mathrm{x}_i-\mathrm{m})\omega_{jk} + \tfrac{1}{3s_{2j}^{\frac{2}{3}}} \sum_{i\in I_j^c} 2\omega_j^T(\mathrm{x}_i-\mathrm{m})\omega_{jk} \right) + \\
\sum_{j=d+1}^{D} \tfrac{-1}{3(s_{1j}+s_{2j})} \cdot \left(\sum_{i\in I_j} 2\omega_j^T(\mathrm{x}_i-\mathrm{m})\omega_{jk} + \sum_{i\in I_j^c} 2\omega_j^T(\mathrm{x}_i-\mathrm{m})\omega_{jk} \right).
\end{aligned}$$

Now we calculate the partial derivative of $\ln l(X;\mathrm{m},W)$ with respect to the matrix W. We have

$$\tfrac{\partial \ln l(X;\mathrm{m},W)}{\partial \omega_{pk}} = \tfrac{\partial \ln |\det(W)|^{-\frac{2}{3}}}{\partial \omega_{pk}} + \sum_{j=1}^{d} \tfrac{\partial \ln(g_j(\mathrm{m},W))}{\partial \omega_{pk}} + \sum_{j=d+1}^{D} \tfrac{\partial \ln((s_{1j}+s_{2j})^{\frac{1}{3}})}{\partial \omega_{pk}}.$$

To calculate the derivative of the determinant we use Jacobi's formula (see Lemma 1). Hence

$$\frac{\partial \ln(\det(W)^{-\frac{2}{3}})}{\partial \omega_{pk}} = \det(W)^{\frac{2}{3}}\Big(-\frac{2}{3}\Big)\det(W)^{-\frac{5}{3}}\frac{\partial \det(W)}{\partial \omega_{pk}}$$
$$= -\frac{2}{3}\det(W)^{-1}\mathrm{adj}^T(W)_{pk} = -\frac{2}{3}\frac{1}{\det(W)}\Big[\det(W)(W^{-1})^T_{pk}\Big] = -\frac{2}{3}(\omega^{-1})^T_{pk},$$

where $(\omega^{-1})^T_{pk}$ is the element in the p-th row and k-th column of the matrix $(W^{-1})^T$. Now we calculate

$$\frac{\partial \ln(g_j(\mathrm{m},W))}{\partial \omega_{pk}} = \frac{1}{s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}}}\frac{\partial(s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}})}{\partial \omega_{pk}} = \frac{1}{s_{1j}^{\frac{1}{3}}+s_{2j}^{\frac{1}{3}}}\left(\frac{1}{3s_{1j}^{\frac{2}{3}}}\frac{\partial s_{1j}}{\partial \omega_{pk}} + \frac{1}{3s_{2j}^{\frac{2}{3}}}\frac{\partial s_{2j}}{\partial \omega_{pk}}\right),$$

where

$$\frac{\partial s_{1j}}{\partial \omega_{pk}} = \sum_{i\in I_j}\frac{\partial[\omega_j^T(\mathrm{x}_i-\mathrm{m})]^2}{\partial \omega_{pk}} = \sum_{i\in I_j} 2\omega_j^T(\mathrm{x}_i-\mathrm{m})\frac{\partial \omega_j^T(\mathrm{x}_i-\mathrm{m})}{\partial \omega_{pk}}$$
$$= \begin{cases} 0, & \text{if } j\neq p \\ \sum\limits_{i\in I_p} 2\omega_p^T(\mathrm{x}_i-\mathrm{m})(\mathrm{x}_{ik}-\mathrm{m}_k), & \text{if } j=p \end{cases}$$

and x_{ik} is the k-th element of the vector x_i. Analogously we get

$$\frac{\partial s_{2j}}{\partial \omega_{pk}} = \begin{cases} 0, & \text{if } j\neq p \\ \sum\limits_{i\in I_p^c} 2\omega_p^T(\mathrm{x}_i-\mathrm{m})(\mathrm{x}_{ik}-\mathrm{m}_k), & \text{if } j=p. \end{cases}$$

Moreover, **it holds that:**

$$\frac{\partial \ln((s_{1j}+s_{2j})^{\frac{1}{3}})}{\partial \omega_{pk}} = \frac{1}{(s_{1j}+s_{2j})^{\frac{1}{3}}}\frac{\partial((s_{1j}+s_{2j})^{\frac{1}{3}})}{\partial \omega_{pk}} = \frac{1}{(s_{1j}+s_{2j})^{\frac{1}{3}}}\frac{1}{3}\frac{1}{(s_{1j}+s_{2j})^{\frac{2}{3}}}\left(\frac{\partial s_{1j}}{\partial \omega_{pk}}+\frac{\partial s_{2j}}{\partial \omega_{pk}}\right),$$

kropka zamiast przecinka w ostatnim rownaniu Hence we obtain

$$\frac{\partial \ln l}{\partial \omega_{pk}} = -\frac{2}{3}(\omega^{-1})^T_{pk} + \frac{1}{s_{1p}^{\frac{1}{3}}+s_{2p}^{\frac{1}{3}}}\left(\frac{1}{3}s_{1p}^{-\frac{2}{3}}\sum_{i\in I_p}2\omega_p^T(\mathrm{x}_i-\mathrm{m})(\mathrm{x}_{ik}-\mathrm{m}_k)\right.$$
$$\left.+\frac{1}{3}s_{2p}^{-\frac{2}{3}}\sum_{i\in I_p^c}2\omega_p^T(\mathrm{x}_i-\mathrm{m})(\mathrm{x}_{ik}-\mathrm{m}_k)\right) + \frac{1}{3(s_{1p}+s_{2p})}\left(\sum_{i\in I_p}2\omega_p^T(\mathrm{x}_i-\mathrm{m})(\mathrm{x}_{ik}-\mathrm{m}_k)+\right.$$
$$\left.\sum_{i\in I_p^c}2\omega_p^T(\mathrm{x}_i-\mathrm{m})(\mathrm{x}_{ik}-\mathrm{m}_k)\right).$$

References

1. Stéphanie Allassonniere and Laurent Younes. A stochastic algorithm for probabilistic independent component analysis. *The Annals of Applied Statistics*, pages 125–160, 2012.
2. H Attias and CE Schreiner. Blind source separation and deconvolution by dynamic component analysis. In *Neural Networks for Signal Processing [1997] VII. Proceedings of the 1997 IEEE Workshop*, pages 456–465. IEEE, 1997.

3. Christian F Beckmann. Modelling with independent components. *Neuroimage*, 62(2):891–901, 2012.
4. Christian F Beckmann and Stephen M Smith. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE transactions on medical imaging*, 23(2):137–152, 2004.
5. Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.
6. Adel Belouchrani, Jean-François Cardoso, et al. Maximum likelihood source separation by the expectation-maximization technique: Deterministic and stochastic implementation. In *Proc. Nolta*, volume 95, pages 49–53. Citeseer, 1995.
7. Gilles Blanchard, Motoaki Kawanabe, Masashi Sugiyama, Vladimir Spokoiny, and Klaus-Robert Mazller. In search of non-gaussian components of a high-dimensional distribution. *Journal of Machine Learning Research*, 7(Feb):247–282, 2006.
8. Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley and Sons, 2012.
9. Christopher G Green, Rajesh R Nandy, and Dietmar Cordes. Pca-preprocessing of fmri data adversely affects the results of ica. In *Proceedings of international society of magnetic resonance in medicine*, volume 10, 2002.
10. Trevor Hastie and Rob Tibshirani. *ProDenICA: Product Density Estimation for ICA using tilted Gaussian density estimates*, 2010. R package version 1.0.
11. Nathaniel E. Helwig. *ica: Independent Component Analysis*, 2015. R package version 1.0-1.
12. Nathaniel E Helwig and Sungjin Hong. A critique of tensor probabilistic independent component analysis: implications and recommendations for multi-subject fmri data analysis. *Journal of neuroscience methods*, 213(2):263–273, 2013.
13. Aapo Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *Neural Networks, IEEE Transactions on*, 10(3):626–634, 1999.
14. Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.
15. Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430, 2000.
16. Sleeba John. The three-parameter two-piece normal family of distributions and its fitting. *Communications in Statistics-Theory and Methods*, 11(8):879–885, 1982.
17. J. Karvanen. *PearsonICA*, 2008. R package version 1.2-3.
18. Motoaki Kawanabe, Masashi Sugiyama, Gilles Blanchard, and Klaus-Robert Müller. A new algorithm of non-gaussian component analysis with radial kernel functions. *Annals of the Institute of Statistical Mathematics*, 59(1):57–75, 2007.
19. Urbano Lorenzo-Seva and Jos MF Ten Berge. Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, 2(2):57–64, 2006.
20. Markus Matilainen, Jari Miettinen, Klaus Nordhausen, Hannu Oja, and Sara Taskinen. *tsBSS: Tools for Blind Source Separation for Time Series*, 2016. R package version 0.2.
21. Jari Miettinen, Klaus Nordhausen, Hannu Oja, and Sara Taskinen. Deflation-based fastica with adaptive choices of nonlinearities. *IEEE Transactions on Signal Processing*, 62(21):5716–5724, 2014.
22. Eric Moulines, J-F Cardoso, and Elisabeth Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 5, pages 3617–3620. IEEE, 1997.

23. Dinh Tuan Pham and Philippe Garat. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *Signal Processing, IEEE Transactions on*, 45(7):1712–1725, 1997.
24. Benjamin B Risk, David S Matteson, and David Ruppert. Likelihood component analysis. *arXiv preprint arXiv:1511.01609*, 2015.
25. Alexander Samarov, Alexandre Tsybakov, et al. Nonparametric independent component analysis. *Bernoulli*, 10(4):565–582, 2004.
26. Richard J Samworth, Ming Yuan, et al. Independent component analysis via nonparametric maximum likelihood estimation. *The Annals of Statistics*, 40(6):2973–3002, 2012.
27. Ran Shi, Ying Guo, et al. Investigating differences in brain functional networks using hierarchical covariate-adjusted independent component analysis. *The Annals of Applied Statistics*, 10(4):1930–1957, 2017.
28. P Spurek, J Tabor, P Rola, and M Ociepka. Ica based on asymmetry. *Pattern Recognition*, 67:230–244, 2017.
29. Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.
30. Mattias Villani and Rolf Larsson. The multivariate split normal distribution and asymmetric principal components analysis. *Communications in StatisticsTheory and Methods*, 35(6):1123–1140, 2006.
31. Joni Virta, Klaus Nordhausen, and Hannu Oja. Projection pursuit for non-gaussian independent components. *arXiv preprint arXiv:1612.05445*, 2016.
32. Nan Wang, Bo Du, Liangpei Zhang, and Lifu Zhang. An abundance characteristic-based independent component analysis for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 53(1):416–428, 2015.
33. Feiyun Zhu, Ying Wang, Shiming Xiang andBin Fan, and Chunhong Pan. Structured sparse method for hyperspectral unmixing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 88:101–118, 2014.
34. Feiyun Zhu, Ying Wang, Bin Fan, Gaofeng Meng, and Chunhong Pan. Effective spectral unmixing via robust representation and learning-based sparsity. *CoRR*, abs/1409.0685, 2014.
35. Feiyun Zhu, Ying Wang, Bin Fan, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Spectral unmixing via data-guided sparsity. *CoRR*, abs/1403.3155, 2014.