Dimension reduction in density based ICA

Anonymous Authors[1]

Abstract

Independent Component Analysis (ICA) - one of the basic tools in data analysis - aims to find a coordinate system in which the components of the data are independent. In many cases the number of sources is unknown and may be less than the number of sensors. 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 often removes important information. In this paper we present a method which is dedicated for determining non-square mixing matrix in ICA maximum likelihood framework.

1. Introduction

2. 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 (; 9), which states that the distribution of the sum (average or linear combination) of N 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 (**?**) is recently gaining popularity (; 8; 0; 2). Then we search for the optimally fitted to data cordinate system B and marginal densities f_i such that the data density factors in base B as the product of maringal 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). As $\mathcal{F}$ we typically choose the super-Gaussian logistic density or other heavy tails distributions.

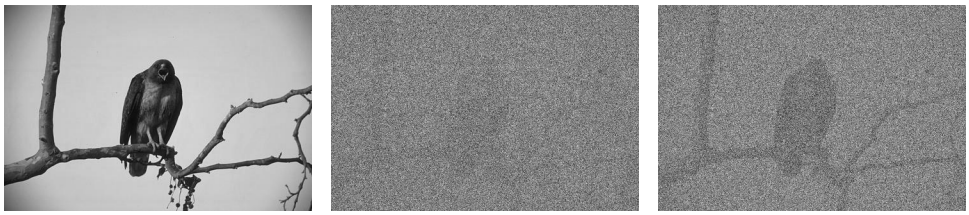
In many applications of ICA we deal with the case when several sensors measure the latents variables and the rest of them record only the noise. This happens when the number of sources is unknown and may be less than the number of sensors (then we are looking for so-called *non-square mixing matrix* W). Such a case is common for example in the identification of brain networks in functional magnetic resonance imaging (fMRI) (; 3). 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 and to reduce computational costs (; 8). Although numerically effective, this approach may fail as it is not invariant with respect to linear transformation, since PCA will find a ''noise" component if it is sufficiently large.

The aim of this paper is to propose a new density based approach to deal with this case which does not have the above mentioned disadvantage. 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 it typically has approximately gaussian distribution. This is why one typically makes the assumption (; 2) that

noise components are comming from a gaussian noise.

Following the density approach 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 (; 2) as $\mathcal{F}$ we take the class SN of split-normal

(a) Original images 42049 and 220075.



(b) ICA_{SG}. (c) FastICA. (d) ProDenICA.

(e) ICA_{SG}. (f) FastICA. (g) ProDenICA.

Figure 1. Comparison of images separation by our method (ICA_{SG}), with FastICA and ProDenICA. 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 ICA_{SG} was able to perfectly recover the two first components.

densities.

Our experiments show, which is illustrated by Figure 1, that ICA_{SG} works as desired and effectively removes the components which contains 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.

3. Related works

4. Related works

In literature exist a few approaches dedicated for non-square ICA problem. Most of such method are dedicated for individual methods. Attias and Schreiner () derived a likelihood based algorithm for separation of general sequences with a frequency domain implementation. Belouchrani and Cardoso () 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 and estimated the mixing matrix using an Expectation-Maximization (EM) approach with both a deterministic and a stochastic formulation. In (; 7) 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 of square mixing is mostly unrealistic in the case of EEG ant FMRI where the number of sources is less than the number of electrodes (; 9; 1). Therefore, many of algorithms dedicated for this task use a probabilistic ICA (; 5). The noisy ICA model can be approximated using a variant of PCA+ICA (), where probabilistic PCA is used to estimate the number of components and achieve dimension reduction (; 5). In () authors developed stochastic EM algorithms to estimate the noisy model and proposed parametric methods.

Other methods exploring non-Gaussian structure in multivariate data include non-Gaussian component analysis (NGCA) and projection pursuit (; 2). NGCA is a more general case of linear non-Gaussian component analysis (LNGCA) (; 8) that allows non-linear dependence between the non-Gaussian components.

In the paper (; 6) authors propose a novel adaptive twostage

deflation-based FastICA algorithm that allows one to use different nonlinearities for different components and optimizes the order in which the components are extracted.

5. Theoretical part

5.1. Measure of non-gaussianity

We are first going to explain the formula (1). Assume that we have a random vector $\mathbf{X}$ with density $f_{\mathbf{X}}$. A most common measure of non-gaussianity is given by the Lullback-Leibler divergence with respect to the gaussian measure. Recall that for a random vector $\mathbf{X}$ finite covariance matrix, its non-Gaussianness is given by

$$\begin{aligned} D(\mathbf{X}) &= D(f_{\mathbf{X}}) = D(f_{\mathbf{X}} \| \mathcal{N}(\mathrm{mean}_{\mathbf{X}}, \mathrm{Cov}_{\mathbf{X}})) \\ &= \tfrac{1}{2} \log\det(2\pi e \mathrm{Cov}_{\mathbf{X}}) - h(f_{\mathbf{X}}), \end{aligned}$$

where h denotes the entropy and $D(f\|g)$ Kullback-Leibler divergence. It is well-known (; 2) that $D(\mathbf{X}) = 0$ iff $\mathbf{X}$ is Gaussian. Clearly, the above measure is affine invariant.

In practice we have only a finite sample X from $\mathbf{X}$ (that is our data-set). In this case, given a family $\mathcal{F}$ of densites on $\mathbb{R}^D$, we can first obtain estimation of the density by MLE in the class $\mathcal{F}$, and only later compute the measure of non-gaussianity. We use the following notation: given a set X, by $[\![X, \mathcal{F}]\!]$ we denote the MLE estimation of the original density X comes from. Thus in this case

$$D(X, \mathcal{F}) = \frac{1}{2} \log\det(2\pi e \mathrm{Cov}_X) - h(X, \mathcal{F})$$

is the estimation of the nongaussianity, where $h(X, \mathcal{F})$ is the estimation of the entropy of X with the use of best estimation from class $\mathcal{F}$:

$$h(X, \mathcal{F}) = \frac{1}{\mathrm{card} X} \sum_x -\ln([\![X, \mathcal{F}]\!](x)).$$

Observe that to maximize the above, since the first part is constant, we can minimize $h([\![X, \mathcal{F}]\!])$, which is equivalent to maximization of the MLE.

To describe the ability of change of variables, we use the notation based on the push-forward of measures (). Assume that we have a measure on $\mathbb{R}^D$ with density f coming from the random variable $\mathbf{X}$. Then for a linear invertible function V, $V\mathbf{X}$ has the density

$$V_* f(x) = \frac{1}{|V|} f(V^{-1} y) \text{ for } x \in \mathbb{R}^D.$$

Consider now the case when $f(x_1, \ldots, x_n) = f_1(x_1) \cdot \ldots f_n(x_n)$, that is f is a product measure with respect to base coordinates. In other words, to compute $f(x)$, we can write

$$x = x_1 e_1 + \ldots + x_n e_n$$

and compute $f(x) = f_1(x_1) \cdot \ldots \cdot f_n(x_n)$. Now if we want to consider the base given by $V = [\mathrm{v}_1, \ldots, \mathrm{v}_n]$ and the analogue of the above, we take $x = x_1 \mathrm{v}_1 + \ldots + x_n \mathrm{v}_n$ and

$$\frac{1}{\det V} f(x_1) \cdot \ldots \cdot f(x_n) = V_* f(x).$$

This is exactly the above.

In our case we will focus our attention on the task of finding d components which are as non-gaussian as possible. For the simplest case, assume that we first consider the first d coordinates, and that we want to measure how X is far from being gaussian on the first coordinates. To do so we first need a family $\mathcal{F}_d$ of densities on $\mathbb{R}^d$, which is larger then gaussians. Now we consider

$$D(X, \mathcal{F}_d \otimes \mathcal{N}_{D-d}),$$

where the tensor product of densities is defined by $(f \otimes g)(x, y) = f(x) \cdot g(y)$.

However, in general working with high-dimensional densities, is nontrivial, and therefore for simplicity we fix a family of one dimensional densities that is larger then gaussians $\mathcal{F}$, and take tensor product of d-elements of $\mathcal{F}$, that is $\mathcal{F}^{\otimes d}$ (in our case it would be the family of split-gaussians). Thus we search for

$$\begin{aligned} &\operatorname*{argmax}_V D(X; V_*(\mathcal{F}^{\otimes d} \otimes \mathcal{N}_{D-d})) \\ &= \operatorname*{argmin}_V h([\![X, V_*(\mathcal{F}^{\otimes d} \otimes \mathcal{N}_{D-d}]\!]). \end{aligned}$$

Thus if we reduce to split gaussians of first d-dimensions, put $W = V^{-1}$, we get the formulation we described in the introduction.

PROBLEM. *Let $X \subset \mathbb{R}^D$ be a data set, and $d \leq D$ be given. Find a matrix $W = [\mathrm{w}_1, \ldots, \mathrm{w}_D]$, densities $f_1, \ldots, f_d \in \mathcal{SN}$, $f_{d+1}, \ldots, f_D \in \mathcal{N}$, such that the likelihood of drawing X from the density*

$$x \to \det(W) \cdot f_1(\mathrm{w}_1^T x) \cdot \ldots \cdot f_D(\mathrm{w}_D^T x)$$

is maximized.

6. ICA based on the asymmetry

7. Theoretical foundations of ICA$_{SG}$

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

7.1. Statement of the problem

Since this general idea of the search for ICA with the use of maximum likelihood 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 there is a basis B (we put $W^T = B^{-1}$) such that in that base the components of $\mathbf{X}$ become independent. Observe that where $\mathrm{w}_i^T\mathrm{x}$ is the i-th coefficient of x in the basis B (w_i denotes the i-th column of W), and therefore there exist densities $f_1, \ldots, f_D$ such that

$$F(\mathrm{x}) = \det(W) \cdot f_1(\mathrm{w}_1^T\mathrm{x}) \cdot \ldots \cdot f_d(\mathrm{w}_d^T\mathrm{x}). \qquad (1)$$

Given W and densities $(f_i)_{i=1}^D$ we introduce notation to represent RHS of the above equation:

$$F_W(f_1, \ldots, f_D)(\mathrm{x}) = \det(W) \cdot f_1(\mathrm{w}_1^T\mathrm{x}) \cdot \ldots \cdot f_d(\mathrm{w}_d^T\mathrm{x}).$$

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

GENERAL ICA PROBLEM (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 densities 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 (; 2) as 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, instead at fitting the distribution with respect to heavy tails, fits the asymmetry of the data.

As mentioned in the introduction, we assume that components which we would like to filter-out, are coming from a gaussian noise, and the aim it to fit the first d-components from a larger class of densities, while the rest from the gaussians $\mathcal{N}$. Thus our final problem can be stated as follows.

ICA FOR DENSITY CLASS $\mathcal{F}$ WITH d SOURCES.
Find matrix W, densities

$$f_1, \ldots, f_d \in \mathcal{F} \text{ and normal densities } f_{d+1}, \ldots, f_D \in \mathcal{N},$$

so that the value of

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

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

$$\begin{aligned} &\int \ln F_W(f_1, \ldots, f_D)(\mathrm{x}) f(\mathrm{x}) d\mathrm{x} \\ &= -H(f, F_W(f_1, \ldots, f_D)), \end{aligned}$$

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

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

over all matrices W and densities $f_1, \ldots, f_D \in \mathcal{F}$. 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 7.1. *Let F be a density such that there exist matrix $\overline{W}$ and densities*

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

such that

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

Then

$$\begin{aligned} &\overline{W}, \hat{f}_1, \ldots, \hat{f}_D \\ &= \operatorname{argmin}\{F_W(f_1, \ldots, f_D) : \\ &\qquad W, f_1, \ldots, f_d \in \mathcal{F}, f_{d+1}, \ldots, f_D \in \mathcal{N}\}. \end{aligned}$$

7.2. Split normal distribution

In this section we discuss the class $\mathcal{F}$ we will use in our final algorithm of ICA$_{SG}$. The density of SN, the one-dimensional split normal distribution (; 6), is given by the formula

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

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

As we see the split normal distribution comes from merging two opposite halves of two normal distributions in their common mode. The main advantage of split normal distributions over normal one is that it allows data asymmetry. In 1982 John (; 0) showed 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 7.1. *A density of the multivariate split normal d and normal $D-d$ distribution is given by*

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

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

Observe that the above density probability function has mode in m. As a consequence of result of John (; 0) we can maximize the likelihood of the above function on data X with respect to σ and τ.

Theorem 7.2. *Let $\mathrm{x}_1,\dots,\mathrm{x}_n$ 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*

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

where

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

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

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

Proof. See Section 6 (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 7.1. *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, we need the formula for $\nabla\ln l$ (gradient of the cost function).

Theorem 7.3. *Let $X\subset\mathbb{R}^d$, $\mathrm{m} = (\mathrm{m}_1,\dots,\mathrm{m}_d)^T\in\mathbb{R}^d$, $W = (\mathrm{w}_{ij})_{1\le i,j\le d}$ non-singular be given. Then $\nabla_{\mathrm{m}}\ln l(X;\mathrm{m},W) = \left(\frac{\partial\ln l(X;\mathrm{m},W)}{\partial\mathrm{m}_1},\dots,\frac{\partial\ln l(X;\mathrm{m},W)}{\partial\mathrm{m}_d}\right)^T$, where*

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

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

$$-\frac{2}{3}(\mathrm{w}^{-1})_{pk}^T+\frac{2}{3(s_{1p}^{\frac{1}{3}}+s_{2p}^{\frac{1}{3}})}\left(s_{1p}^{-\frac{2}{3}}\sum_{i\in I_p}\mathrm{w}_p^T(\mathrm{x}_i-\mathrm{m})(\mathrm{x}_{ik}-\mathrm{m}_k)+s_{2p}^{-\frac{2}{3}}\sum_{i\in I_p^c}\mathrm{w}_p^T(\mathrm{x}_i-\mathrm{m})(\mathrm{x}_{ik}-\mathrm{m}_k)\right)+\frac{2}{3(s_{1p}+s_{2p})}\left(\sum_{i\in I_p}\mathrm{w}_p^T(\mathrm{x}_i-\mathrm{m})(\mathrm{x}_{ik}-\mathrm{m}_k)+\sum_{i\in I_p^c}\mathrm{w}_p^T(\mathrm{x}_i-\mathrm{m})(\mathrm{x}_{ik}-\mathrm{m}_k)\right)$$

and

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

Proof. See Section 7 (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.

8. Experiments

9. Experiments

To compare ICA_{SG} to other state-of-the-art approaches we use Tucker's congruence coefficient (; 3) 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` (; 5), `PearsonICA` (; 1), `ProDenICA` (; 4), `tsBSS` (; 4). The most popular method used in practice is FastICA (; 7; 6) algorithm, which uses negentropy. In this context we can use three different functions to estimate neg-entropy: log-cosh, exp and kurtosis. We also compare our method with algorithm using Information-Maximization (Infomax) approach (). Similarly to FastICA we consider three possible non-linear functions: hyperbolic tangent, logistic and extended Infomax.

9.1. 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. The results of this experiment are presented in Tab. 1 where we present Tucker's congruence coefficients which shows that almost in all cases ICA_{SG} obtains best results. This is illustrated in Figure 1, where we can see that ICA_{SG} almost perfectly recovered source signal. 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, this shows that ICA_{SG} works as desired and deals well with removing gaussian components from the data.

9.2. Hyperspectral Unmixing

Independent component analysis has been recently applied into hyperspectral unmixing (; 8;) as a result of its low computation time and its ability to perform without prior information. In this subsection we apply simple example which suggests that our method also can by used for spectral data.

Urban data (; 9; 1; 0) is one of the most widely used hyperspectral data-sets used in the 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 (; 9; 1; 0), 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 (; 8)), with the size of 200×150 pixels.

In our experiment we compared ICA_{SG} to other two popular ICA methods – ProDenICA and FastICA, see Fig. 2. Observe that ICA_{SG} and ProDenICA give layers which seem to contain more information then FastICA, as the last component in FastICA contains mainly noise.

10. Appendix A

Proof of Theorem 4.1. Let $X = \{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$. We write

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

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

Table 1. Tucker's congruence coefficients for reconstruction of two images.

	ICA_{SG}	FastICA logcosh	FastICA exp	FastICA kurtosis	Infomax tanh	Infomax tangent	Infomax logistic	JADE	PearsonICA	ProDenICA	FixNA
4.1.01	**0.6807**	0.0325	0.0321	0.0352	0.0327	0.0334	0.0449	0.029	0.0549	0.1184	0.5487
4.1.02	**0.5491**	0.4588	0.4588	0.459	0.4589	0.4589	0.4593	0.4584	0.4592	0.1781	0.0074
4.1.06	0.4279	0.0143	0.0142	0.0149	0.0142	0.0145	0.0143	0.0147	0.0141	0.0777	**0.4295**
4.1.03	0.2033	0.4151	0.4151	0.415	0.4151	0.4151	0.4151	0.4151	**0.4152**	0.0932	0.0107
4.2.04	**0.4333**	0.0234	0.0233	0.0238	0.0234	0.0236	0.0234	0.0234	0.3681	0.064	0.3647
5.2.10	0.3161	**0.3681**	**0.3681**	**0.3681**	**0.3681**	**0.3681**	**0.3681**	**0.3681**	0.0234	0.0704	0.2652
4.2.02	0.0235	0.0066	0.0067	0.011	0.0066	0.0066	0.0067	0.0109	**0.282**	0.0164	0.1244
5.2.08	0.1279	**0.282**	**0.282**	0.2818	**0.282**	**0.282**	**0.282**	0.2818	0.0066	0.0099	4E-04
BOAT.512	**0.5454**	0.0355	0.0357	0.0329	0.0356	0.0339	0.0355	0.035	0.4631	0.0388	0.2753
5.3.01	0.3368	0.463	0.4631	0.463	0.463	0.463	0.463	0.463	0.0367	0.3282	**0.4778**
ELAINE.512	**0.3134**	0.0213	0.0283	0.0173	0.0213	0.0175	0.0213	0.0197	0.2507	0.0469	0.2296
4.2.03	**0.3216**	0.2509	0.2511	0.2506	0.2509	0.2506	0.2509	0.2508	0.0189	0.1894	0.2567
119082	**0.5744**	0.0428	0.043	0.042	0.0427	0.0423	0.0416	0.0402	0.0412	0.0921	0.371
157055	0.3612	**0.482**	**0.482**	**0.482**	**0.482**	**0.482**	0.4819	0.4818	0.4819	0.0599	0.0032
42049	**0.5287**	0.0398	0.0396	0.039	0.04	0.04	0.0399	0.037	0.0382	0.5014	0.2964
220075	0.3043	0.3668	0.3668	0.3669	0.3668	0.3668	0.3668	0.3671	0.3669	0.061	**0.4956**
43074	**0.3886**	0.0501	0.0692	0.0281	0.0502	0.0331	0.0501	0.037	0.0346	0.0461	0.3168
295087	**0.401**	0.3384	0.3336	0.3407	0.3384	0.3405	0.3384	0.3403	0.3404	0.0676	0.3025
38092	0.3841	0.0504	0.0502	0.0497	0.0504	0.0499	0.0504	0.0497	0.0503	0.2541	**0.4965**
167062	**0.7406**	0.6963	0.6962	0.6962	0.6963	0.6962	0.6963	0.6962	0.6963	0.587	0.5496

Let us consider the likelihood function, i.e.

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

where $c_1 = \left(\sqrt{\frac{2}{\pi}}\right)^d \cdot \left(\frac{1}{\sqrt{2\pi}}\right)^{D-d}$. Now we take the log-likelihood function, i.e. $\ln(L(X;\mathrm{m},W,\sigma,\tau)) =$

$$\ln\left(\Big(c_1|\det(W)|\Big)^n\Big(\prod_{j=1}^{d}\sigma_j(1+\tau_j)\Big)^{-n}\Big(\prod_{j=d+1}^{D}\sigma_j\Big)^{-n}\right)+$$
$$\sum_{i=1}^{n}\sum_{j=1}^{d}\Big[-\tfrac{1}{2\sigma_j^2}z_{ij}^2(\mathbb{1}_{\{z_{ij}\le 0\}}+\tau_j^{-2}\mathbb{1}_{\{z_{ij}>0\}})\Big]+$$
$$\sum_{i=1}^{n}\sum_{j=d+1}^{D}\Big[-\tfrac{1}{2\sigma_j^2}z_{ij}^2\Big]$$
$$= \ln\left(\Big(c_1|\det(W)|\Big)^n\Big(\prod_{j=1}^{d}(1+\tau_j)\Big)^{-n}\Big(\prod_{j=1}^{D}\sigma_j\Big)^{-n}\right)-$$
$$\tfrac{1}{2}\sum_{j=1}^{d}\Big(\sigma_j^{-2}\sum_{i\in I_j}z_{ij}^2+\tfrac{\sigma_j^{-2}}{\tau_j^2}\sum_{i\in I_j^c}z_{ij}^2\Big)-$$
$$\tfrac{1}{2}\sum_{j=d+1}^{D}\sigma_j^{-2}\Big(\sum_{i\in I_j}z_{ij}^2+\sum_{i\in I_j^c}z_{ij}^2\Big)$$
$$= \ln\left(\Big(c_1|\det(W)|\Big)^n\Big(\prod_{j=1}^{d}(1+\tau_j)\Big)^{-n}\Big(\prod_{j=1}^{D}\sigma_j\Big)^{-n}\right)-$$
$$\sum_{j=1}^{d}\tfrac{1}{2\sigma_j^2}\Big(s_{1j}+\tfrac{1}{\tau_j^2}s_{2j}\Big)-\sum_{j=d+1}^{D}\tfrac{1}{2\sigma_j^2}\Big(s_{1j}+s_{2j}\Big).$$

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

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

for $j=1,\dots,D$. Hence

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

By simple calculations we obtain the expressions for the estimators in 3. Substituting it into the log-likelihood function, we get $\hat{L}(\mathrm{m},W) =$

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

□

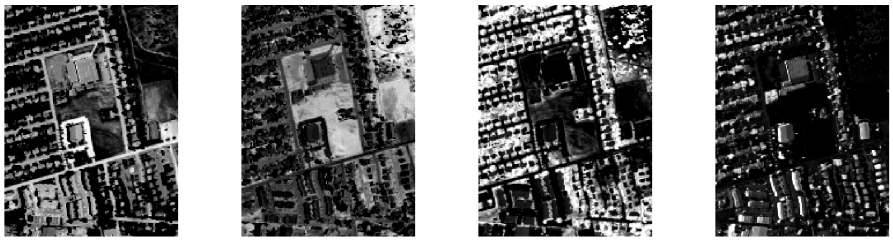
11. Appendix B

12. Appendix B

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

Lemma 12.1. *Let $A=(a_{ij})_{1\le i,j\le d}$ be a differentiable map from real numbers to $d\times d$ matrices then*

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



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

(b) The effect of the ICA$_{SG}$ method.

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

(d) The effect of the ProDenICA method.

Figure 2. Results of image separation with the uses of various ICA algorithms.

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

Proof. By the Laplace expansion $\det A = \sum_{j=1}^{d}(-1)^{i+j}a_{ij}M_{ij}$ 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 of Theorem 4.2. Let us start with the partial derivative of $\ln(l)$ with respect to m. We have

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

Now, we need $\frac{\partial s_{1j}}{\partial \mathrm{m}_k}$ and $\frac{\partial s_{2j}}{\partial \mathrm{m}_k}$, therefore

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

Analogously we get

$$\frac{\partial s_{2j}}{\partial \mathrm{m}_k} = \sum_{i\in I_j^c} -2\mathrm{w}_j^T(\mathrm{x}_i-\mathrm{m})\mathrm{w}_{jk}.$$

Hence

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

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

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

To calculate the derivative of the determinant we use Jacobi's formula (see Lemma 4.1). Hence

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

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

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

where

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

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

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

Moreover,

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

Hence we obtain

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

□

References

Stéphanie Allassonniere and Laurent Younes. A stochastic algorithm for probabilistic independent component analysis. *The Annals of Applied Statistics*, pages 125–160, 2012.

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.

Christian F Beckmann. Modelling with independent components. *Neuroimage*, 62(2):891–901, 2012.

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.

Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

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.

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.

Vladimir I Bogachev. *Measure theory*, volume 1. Springer Science & Business Media, 2007.

Cesar F Caiafa, Emanuele Salerno, Araceli N Proto, and L Fiumi. Blind spectral unmixing by local maximization of non-gaussianity. *Signal Processing*, 88(1):50–68, 2008.

Jean-François Cardoso and Antoine Souloumiac. Blind beamforming for non-gaussian signals. In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 362–370. IET, 1993.

Aiyou Chen, Peter J Bickel, et al. Efficient independent component analysis. *The Annals of Statistics*, 34(6):2825–2855, 2006.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

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.

Trevor Hastie and Rob Tibshirani. *ProDenICA: Product Density Estimation for ICA using tilted Gaussian density estimates*, 2010. R package version 1.0.

Nathaniel E. Helwig. *ica: Independent Component Analysis*, 2015. R package version 1.0-1.

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.

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

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430, 2000.

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.

J. Karvanen. *PearsonICA*, 2008. R package version 1.2-3.

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.

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.

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.

David S Matteson and Ruey S Tsay. Independent component analysis via distance covariance. *Journal of the American Statistical Association*, (just-accepted):1–38, 2016.

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.

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.

Benjamin B Risk, David S Matteson, and David Ruppert. Likelihood component analysis. *arXiv preprint arXiv:1511.01609*, 2015.

Alexander Samarov, Alexandre Tsybakov, et al. Nonparametric independent component analysis. *Bernoulli*, 10(4):565–582, 2004.

Richard J Samworth, Ming Yuan, et al. Independent component analysis via nonparametric maximum likelihood estimation. *The Annals of Statistics*, 40(6):2973–3002, 2012.

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.

P. Spurek et al. Ica based on the data asymmetry. *Pattern Recognition (accepted)*, 2017.

Przemysław Spurek. General split gaussian cross–entropy clustering. *Expert Systems with Applications*, 68:58–68, 2017.

Harald Stögbauer, Alexander Kraskov, Sergey A Astakhov, and Peter Grassberger. Least-dependent-component analysis based on mutual information. *Physical Review E*, 70(6):066123, 2004.

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.

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.

Joni Virta, Klaus Nordhausen, and Hannu Oja. Joint use of third and fourth cumulants in independent component analysis. *arXiv preprint arXiv:1505.02613*, 2015.

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.

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.

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.

Feiyun Zhu, Ying Wang, Bin Fan, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Spectral unmixing via data-guided sparsity. *CoRR*, abs/1403.3155, 2014.