# ON THE RELATION BETWEEN ICA AND MMSE BASED SOURCE SEPARATION

*Benedikt Loesch and Bin Yang*

Chair of System Theory and Signal Processing, University of Stuttgart
Email: {benedikt.loesch,bin.yang}@LSS.uni-stuttgart.de

## ABSTRACT

This paper aims at deriving a relationship between minimum mean square error (MMSE) based source separation and independent component analysis (ICA) based on the Kullback-Leibler divergence (KLD) for a linear noisy mixing model. Starting from a description of the demixing task and two well-known solutions, inverse mixing matrix and MMSE solution, we derive an analytic expression for the demixing matrix of KLD-based ICA in the presence of noise. The derivation is done by using a perturbation analysis valid for small noise variance. Furthermore, we provide an analytic expression for the mean square error (MSE) of the demixed signals using KLD-based ICA. We show that for a wide range of the shape parameter of the generalized Gaussian distribution (GGD), the MSE of KLD-based ICA is very close to the MMSE. Simulations verify this and show that in practice the variance of the ICA estimation due to limited amount of data also influences the achievable performance.

***Index Terms***— Blind source separation, Independent component analysis, Minimum mean square error, Kullback-Leibler divergence, Perturbation analysis

## 1. INTRODUCTION

Up to now, most research concerning ICA considered the noiseless mixing model. The presence of noise leads to a bias in the estimation of the mixing matrix. [1] introduced measures to reduce this bias. [2] studied the maximum likelihood (ML) estimation of both the mixing matrix and the signals. It was shown that in the presence of noise, the ML estimate of the signals is a nonlinear function of the observations. [3,4] drawed parallels between MMSE estimation and ICA. They derived the bias of several variants of the FastICA algorithm, a fixed-point method, from the MMSE solution. [4] also discussed that many ICA algorithms use an orthogonal constraint resulting in decreased separation quality in the presence of noise.

In this paper, we focus on gradient-based ICA using the KLD without such an orthogonal constraint. We derive an analytic expression for the demixing matrix and the MSE of KLD-based ICA in the presence of noise. We compare the MSE of ICA with the MMSE and the MSE of the inverse solution in order to study the performance loss of blind demixing with respect to nonblind demixing.Furthermore, simulations with limited amount of data show that for small to moderate signal-to-noise ratios (SNR), natural gradient (NG) ICA based on KLD achieves an MSE close to the MMSE for a wide range of the shape parameter $\beta$ of the GGD.

## 2. SIGNAL MODEL

We assume the square linear noisy mixing model

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{v} \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^N$ are linear combinations of the $N$ original signals $\mathbf{s} \in \mathbb{R}^N$ with additive noise $\mathbf{v} \in \mathbb{R}^N$. We make the following assumptions:

1. The mixing matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ is deterministic and invertible.
2. $\mathbf{s} = [s_1, \cdots, s_N]^T \in \mathbb{R}^N$ are $N$ independent non-Gaussian random variables with zero mean and unit variance (after scaling the rows of $\mathbf{A}$ suitably). The probability density functions (pdfs) $q_i(s_i)$ of $s_i$ can be different. $q_i(s_i)$ is three times continuously differentiable and all expectations required in the derivation of (10) exist. This excludes some distributions, such as $\alpha$-stable distributions with $\alpha < 2$ or GGD with $\beta < \frac{1}{2}$.
3. $\mathbf{v} = [v_1, \cdots, v_N]^T \in \mathbb{R}^N$ are $N$ random variables with zero mean and covariance matrix $E[\mathbf{v}\mathbf{v}^T] = \sigma^2\mathbf{R}_\mathbf{v}$. $\sigma^2 = \frac{1}{N}\text{tr}\left[E(\mathbf{v}\mathbf{v}^T)\right]$ is the average variance of $\mathbf{v}$ and $\text{tr}(\mathbf{R}_\mathbf{v}) = N$. The pdf of $\mathbf{v}$ is arbitrary but symmetric, i.e. $q(\mathbf{v}) = q(-\mathbf{v})$. This implies $E(v_1^{k_1}\cdots v_N^{k_N}) = 0$ for $k_1 + \cdots + k_N$ odd.
4. $\mathbf{s}$ and $\mathbf{v}$ are independent.

The task is to demix the signals by a linear transform $\mathbf{W} \in \mathbb{R}^{N\times N}$

$$\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{A}\mathbf{s} + \mathbf{W}\mathbf{v} \tag{2}$$

such that $\mathbf{y}$ is "as close to $\mathbf{s}$" as possible according to some metric.

## 3. APPROACHES FOR DEMIXING

To find a demixing matrix $\mathbf{W}$, several approaches exist, which will now be briefly reviewed.

### 3.1. Inverse solution

The inverse solution

$$\mathbf{W}_{\text{inv}} = \mathbf{A}^{-1}, \quad \mathbf{y}_{\text{inv}} = \mathbf{s} + \mathbf{A}^{-1}\mathbf{v} \tag{3}$$

has the following properties:

- It is a perfect demixing $\mathbf{y}_{\text{inv}} = \mathbf{s}$ if there is no noise ($\mathbf{v} = \mathbf{0}$).
- There is a danger of noise amplification $\mathbf{A}^{-1}\mathbf{v}$ if $\mathbf{v} \neq \mathbf{0}$. This is especially serious if $\mathbf{A}$ is close to singular.
- In digital communication, this solution is called zero-forcing.
- It is only possible if we know $\mathbf{A}$ in advance.

### 3.2. MMSE solution

The MMSE solution is given by

$$\begin{aligned}\mathbf{W}_{\text{MMSE}} &= \arg\min_{\mathbf{W}} \|\mathbf{W}\mathbf{x} - \mathbf{s}\|^2 \\ &= E\left(\mathbf{s}\mathbf{x}^T\right)\left[E\left(\mathbf{x}\mathbf{x}^T\right)\right]^{-1} = \mathbf{A}^T\left(\mathbf{A}\mathbf{A}^T + \sigma^2\mathbf{R}_\mathbf{v}\right)^{-1} \\ &= \left[\mathbf{I} - \sigma^2\mathbf{A}^{-1}\mathbf{R}_\mathbf{v}\mathbf{A}^{-T}\right]\mathbf{A}^{-1} + O(\sigma^4) \qquad (4) \\ &= \left[\mathbf{I} - \sigma^2\mathbf{R}_{\tilde{\mathbf{v}}}\right]\mathbf{A}^{-1} + O(\sigma^4) \qquad (5)\end{aligned}$$

where $\tilde{\mathbf{v}} = \mathbf{A}^{-1}\mathbf{v}, \mathbf{R}_{\tilde{\mathbf{v}}} = E\left[\tilde{\mathbf{v}}\tilde{\mathbf{v}}^T\right]$ and the last two lines are a first-order Taylor series approximation of $\mathbf{W}_{\text{MMSE}}(\sigma^2)$ at $\sigma^2 = 0$. It has the following properties:

- $\mathbf{W}_{\text{MMSE}} = \mathbf{W}_{\text{inv}} = \mathbf{A}^{-1}$ if $\sigma^2 = 0$ (no noise).
- It can only be calculated if we know $\{\mathbf{A}, \mathbf{R}_\mathbf{v}, \sigma^2\}$ or $E\left(\mathbf{s}\mathbf{x}^T\right)$ and $E\left(\mathbf{x}\mathbf{x}^T\right)$ can be estimated from measurements of $\mathbf{x}$ and $\mathbf{s}$.

### 3.3. KLD-based ICA

In blind demixing or blind source separation, neither $\mathbf{A}$ nor $\mathbf{s}$ are known. Hence, neither $\mathbf{W}_{\text{inv}}$ nor $\mathbf{W}_{\text{MMSE}}$ can be caclulated.

In this paper, we focus on the ICA solution based on the KLD

$$D_{pq}(\mathbf{W}) = \int p_{\mathbf{y}}(\mathbf{y};\mathbf{W}) \log \frac{p_{\mathbf{y}}(\mathbf{y};\mathbf{W})}{q(\mathbf{y})} d\mathbf{y}, \tag{6}$$

with $q(\mathbf{s}) = \prod_{i=1}^{N} q_i(s_i)$ being the true pdf of $\mathbf{s}$. KLD utilizes the full pdf and hence yields different solutions than the MMSE criterion which is solely based on second order moments. KLD is closely linked to mutual information, the well-known information maximization (INFOMAX) principle and for the noiseless case to ML estimation [5]. Hence the following study applies to all ICA algorithms that use this type of cost function. The derivative of $D_{pq}(\mathbf{W})$ with respect to $\mathbf{W}$ is

$$\frac{dD_{pq}(\mathbf{W})}{d\mathbf{W}} = \left[\frac{dD_{pq}(\mathbf{W})}{dw_{ij}}\right]_{ij} = \left[E(\boldsymbol{\varphi}(\mathbf{y})\mathbf{y}^T) - \mathbf{I}\right]\mathbf{W}^{-T} \tag{7}$$

with

$$\boldsymbol{\varphi}(\mathbf{y}) = [\varphi_1(y_1), \cdots, \varphi_N(y_N)]^T, \quad \varphi_i(y_i) = -\frac{q_i'(y_i)}{q_i(y_i)} \tag{8}$$

Hence, a necessary condition for the ICA solution $\mathbf{W}_{\text{ICA}} = \arg\min_{\mathbf{W}} D_{pq}(\mathbf{W})$ is given by

$$E(\boldsymbol{\varphi}(\mathbf{y}_{\text{ICA}})\mathbf{y}_{\text{ICA}}^T) \stackrel{!}{=} \mathbf{I} \tag{9}$$

with $\mathbf{y}_{\text{ICA}} = \mathbf{W}_{\text{ICA}}\mathbf{x} = \mathbf{W}_{\text{ICA}}\mathbf{A}\mathbf{s} + \mathbf{W}_{\text{ICA}}\mathbf{v} = \hat{\mathbf{y}} + \mathbf{W}_{\text{ICA}}\mathbf{v}$. Note, that (9) characterizes the stationary points of the KLD cost function (6) and hence holds for the noiseless as well as noisy case.

NG-ICA uses a gradient descent search with the modified gradient $\boldsymbol{\Delta}\mathbf{W} = \left[E(\boldsymbol{\varphi}(\mathbf{y}_{\text{ICA}})\mathbf{y}_{\text{ICA}}^T) - \mathbf{I}\right]\mathbf{W}$. This improves the convergence but does not change the final ICA solution $\mathbf{W}_{\text{ICA}}$.
The properties of the ICA solution are:

- $\mathbf{W}_{\text{ICA}} = \mathbf{W}_{\text{inv}} = \mathbf{A}^{-1}$ if $\sigma^2 = 0$ (no noise).
- We do not need to know $\mathbf{A}$ or $\mathbf{s}$. Only the pdf $q(\mathbf{s}) = \prod_{i=1}^{N} q_i(s_i)$ is required and $q_i(s_i)$ needs to be non-Gaussian.
- There is no permutation ambiguity if $q_i(s_i) \neq q_j(s_j) \forall i \neq j$.
- There is no scaling ambiguity if $q_i(s_i)$ is known $\forall i$. Only a sign ambiguity remains if $q_i(s_i)$ is symmetric.

Now, we derive an analytic expression for $\mathbf{W}_{\text{ICA}}$ in the presence of noise by using a perturbation analysis. Motivated by $\mathbf{W}_{\text{ICA}} \stackrel{\sigma^2=0}{=} \mathbf{A}^{-1}$, we assume that $\mathbf{W}_{\text{ICA}}$ can be written as $\mathbf{W}_{\text{ICA}} = \mathbf{A}^{-1} + \sigma^2\mathbf{B} + O(\sigma^4)$ (see appendix A for a rigorous justification) and obtain $\mathbf{B}$ by a two-step perturbation analysis:

1. Taylor series approximation of $E(\boldsymbol{\varphi}(\mathbf{y})\mathbf{y}^T)$ in (9) at $\mathbf{y} = \hat{\mathbf{y}} = \mathbf{W}_{\text{ICA}}\mathbf{A}\mathbf{s}$
2. Taylor series approximation of the result of the above step by exploiting $\mathbf{W}_{\text{ICA}} = \mathbf{A}^{-1} + \sigma^2\mathbf{B} + O(\sigma^4)$ and $\hat{\mathbf{y}} = \mathbf{s} + \sigma^2\mathbf{B}\mathbf{A}\mathbf{s} + O(\sigma^4)$

In this way, we determine explicitely the deviation $\sigma^2\mathbf{B}$ of $\mathbf{W}_{\text{ICA}}$ from the inverse solution $\mathbf{A}^{-1}$.

As shown in appendix A, the final ICA solution is

$$\mathbf{W}_{\text{ICA}} = (\mathbf{I} + \sigma^2\mathbf{C})\mathbf{A}^{-1} + O(\sigma^4) = (\mathbf{I} - \sigma^2\mathbf{M} \odot \mathbf{R}_{\tilde{\mathbf{v}}})\mathbf{A}^{-1} + O(\sigma^4). \tag{10}$$

$\odot$ denotes elementwise multiplication, $\mathbf{R}_{\tilde{\mathbf{v}}} = \mathbf{A}^{-1}\mathbf{R}_{\mathbf{v}}\mathbf{A}^{-T}$and

$$M_{ii} = \frac{\kappa_i + \frac{1}{2}\lambda_i}{1 + \rho_i}, \quad M_{ij} = \frac{\kappa_j(\kappa_i - 1)}{\kappa_i\kappa_j - 1} \quad i \neq j$$

$$\boldsymbol{\kappa} = [\kappa_1, \cdots, \kappa_N]^T, \quad \kappa_i = E(\varphi_i'(s_i))$$

$$\boldsymbol{\rho} = [\rho_1, \cdots, \rho_N]^T, \quad \rho_i = E(\varphi_i'(s_i)s_i^2)$$

$$\boldsymbol{\lambda} = [\lambda_1, \cdots, \lambda_N]^T, \quad \lambda_i = E(\varphi_i''(s_i)s_i)$$

$\mathbf{W}_{\text{ICA}}$ in (10) depends on the mixing matrix $\mathbf{A}$, the noise covariance matrix $\mathbf{R}_{\mathbf{v}}$ and the pdfs $q_i(s_i)$. $\kappa$ is a measure of non-Gaussianity. $\kappa \geq 1$ for all pdfs. $\kappa = 1$ if and only if $s$ is Gaussian [6].

Note, that the expression in (10) has been derived by evaluating the expectations exactly. Hence, strictly speaking, it is valid only for infinite amount of data.

## 4. MSE OF THE DIFFERENT SOLUTIONS

In this section, we derive analytic expressions for the mean square error $\text{MSE} = E\left(\|\mathbf{y} - \mathbf{s}\|^2\right)$ of the demixed signals $\mathbf{y} = \mathbf{W}\mathbf{x}$. It is straightforward to show that

$$\text{MSE} = \text{tr}\left[\mathbf{W}(\mathbf{A}\mathbf{A}^T + \sigma^2\mathbf{R}_{\mathbf{v}})\mathbf{W}^T + (\mathbf{I} - 2\mathbf{W}\mathbf{A})\right]. \tag{11}$$

For the inverse solution and the MMSE solution, we get after some calulations

$$\text{MSE}_{\text{MMSE}} = \sigma^2\text{tr}(\mathbf{R}_{\tilde{\mathbf{v}}}) - \sigma^4\text{tr}(\mathbf{R}_{\tilde{\mathbf{v}}}^2) + O(\sigma^6), \tag{12}$$

$$\text{MSE}_{\text{inv}} = \sigma^2\text{tr}(\mathbf{R}_{\tilde{\mathbf{v}}}). \tag{13}$$

With $\mathbf{W}_{\text{ICA}} = (\mathbf{I} + \sigma^2\mathbf{C})\mathbf{A}^{-1} + O(\sigma^4)$, we get for the ICA solution

$$\begin{aligned}\text{MSE}_{\text{ICA}} &= \sigma^2\text{tr}(\mathbf{U}) + \sigma^4\text{tr}\left(\mathbf{C}\mathbf{C}^T + \mathbf{C}\mathbf{R}_{\tilde{\mathbf{v}}} + \mathbf{R}_{\tilde{\mathbf{v}}}\mathbf{C}^T\right) + O(\sigma^6) \\ &= \text{MSE}_{\text{MMSE}} + \sigma^4\text{tr}\left[((\mathbf{1} - \mathbf{M}) \odot \mathbf{R}_{\tilde{\mathbf{v}}})((\mathbf{1} - \mathbf{M}) \odot \mathbf{R}_{\tilde{\mathbf{v}}})^T\right] \\ &\quad + O(\sigma^6). \end{aligned} \tag{14}$$

where $\mathbf{1}$ denotes a matrix whose elements are all one.

## 5. DISCUSSION

### 5.1. Theoretical results

Comparing (10) and (5), we see that the ICA solution and the MMSE solution are quite similar except for the scaling matrix $\mathbf{M}$. If $\mathbf{M} \approx \mathbf{1}$, the ICA solution is close to the MMSE solution. The elements of $\mathbf{M}$ are determined by the pdf $q(s)$ of the sources:
$M_{ij} \to 1$ if $\kappa_i \cdot \kappa_j \to \infty$ and $M_{ii} \to 1$ if $\frac{\kappa_i + 1/2\lambda_i}{1 + \rho_i} \to 1$.

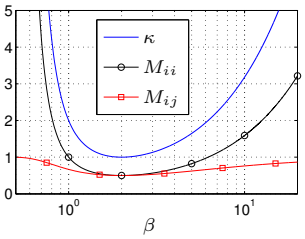
We now consider the generalized Gaussian (GGD) distribution $q(s) = \frac{\beta}{2\alpha\Gamma(\frac{1}{\beta})} e^{-\left(\frac{|s|}{\alpha}\right)^{\beta}}$, where $\Gamma(\cdot)$ is the Gamma function. The GGD is a flexible distribution that incorporates the Gaussian distribution ($\beta = 2$), Laplacian distribution ($\beta = 1$) and uniform distribution ($\beta \to \infty$). For a GGD with variance 1, i.e. $\alpha = \sqrt{\Gamma(\beta^{-1})/\Gamma(3\beta^{-1})}$, we get after some calculations

$$\kappa = \begin{cases} \frac{\Gamma\left(2 - \frac{1}{\beta}\right)\Gamma\left(\frac{3}{\beta}\right)}{\Gamma^2\left(1 + \frac{1}{\beta}\right)} & \beta > \frac{1}{2} \\ \infty & \text{otherwise} \end{cases},$$

$$\rho = \beta - 1,$$

$$\lambda = \begin{cases} (\beta - 2)\kappa & \beta > \frac{1}{2} \\ -\infty & \text{otherwise} \end{cases}. \tag{15}$$

With (15), $M_{ii}$ simplifies to $M_{ii} = \kappa_i/2$. Fig. 1 shows $\kappa$, $M_{ii}$ and $M_{ij}$ when all sources $s_i$ follow a GGD with the same shape parameter $\beta$. In the following, we consider $N = 2$ such sources with mixing matrix coefficients $A_{ii} = 1$ and $A_{ij} = 0.5$ ($i \neq j$) and Gaussian noise with $\mathbf{R}_{\mathbf{v}} = \mathbf{I}$. Fig. 2 plots the relative MSE $\text{MSE}_{\text{rel}} = \text{MSE}/\text{MSE}_{\text{MMSE}}$ of the ICA solution for different values of the SNR $1/\sigma^2$ and $\beta$. The MSE is calculated using the exact expression from (11). The relative MSE of the inverse solution is 1.35 and 1.04 for the two SNRs of 10 dB and 20 dB. We see that for a wide range of $\beta$, the MSE of the ICA solution is close to $\text{MSE}_{\text{MMSE}}$. This range becomes larger as the SNR increases. $\text{MSE}_{\text{ICA}}$ shows a local maximum at $\beta = 2$ and two minima which move towards $\beta = 0$ or $\beta \to \infty$ as the SNR increases.




| $\beta$ | $\kappa$ | $M_{ii}$ | $M_{ij}$ |
|---|---|---|---|
| 0.50 | $\infty$ | $\infty$ | 1.00 |
| 0.75 | 5.73 | 2.87 | 0.85 |
| 1.00 | 2.00 | 2.00 | 0.67 |
| 2.00 | 1.00 | 0.50 | 0.50 |
| 5.00 | 1.65 | 0.82 | 0.62 |
| 10.00 | 3.18 | 1.59 | 0.76 |
| 20.00 | 6.43 | 3.22 | 0.87 |

**Fig. 1**: Values of $\kappa$, $M_{ii} = \frac{\kappa}{2}$ and $M_{ij} = \frac{\kappa^2-\kappa}{\kappa^2-1} = \frac{\kappa}{\kappa+1}$

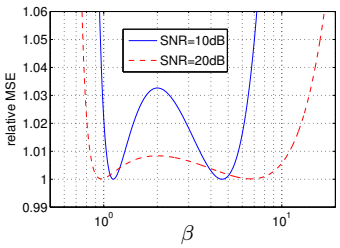



**Fig. 2**: Relative MSE (with respect to $\text{MSE}_{\text{MMSE}}$) of theoretical ICA solution for different values of $\beta$ and different SNRs

### 5.2. Simulation results

In the simulations, we want to study the MSE of demixed signals which is achievable in practice, i.e. with a NG KLD-based ICA algorithm and with limited amount of data. In practice, both the bias of $\mathbf{W}_{\text{ICA}}$ from the MMSE solution and the covariance of $\mathbf{W}_{\text{ICA}}$ contribute to the MSE. We consider the same scenario with $N = 2$ identically distributed sources as in Section 5.1. Since $q_1(s_1) = q_2(s_2)$, ICA suffers from the permutation ambiguity. Furthermore, the variance of the signals demixed by the ICA solution differs from the variance of the signals demixed by the MMSE solution. Hence, we adjust the row permutation and row scaling of $\mathbf{W}_{\text{ICA}}$ such that $\text{diag}(\mathbf{W}_{\text{ICA}}\mathbf{W}_{\text{MMSE}}^{-1} - \mathbf{I}) = 0$. We then use (11) to calculate the MSE for an estimated $\mathbf{W}_{\text{ICA}}$ and average the MSE for 100 independent trials. We use NG adaptation with an adaptive step-size ensuring that the cost-function decreases in each iteration.

First, we consider Laplacian distributed sources and study the behaviour for different SNRs and sample sizes $L$. As shown in Fig. 3, NG-ICA yields an MSE that is close to the theoretical MSE of ICA for moderate SNR and close to $\text{MSE}_{\text{MMSE}}$ for low to moderate SNR. Due to small noise assumptions in the derivation of $\mathbf{W}_{\text{ICA}}$, the theoretical MSE of ICA is only valid for moderate to high SNRs. For low SNR up to 20 dB, ICA outperforms the inverse solution for sample sizes $L \geq 10^3$. In the high SNR region, the MSE of ICA is bounded by the estimation variance due to limited amount of data. This bound is related to the CRB derived in [6], since the ICA solution $\mathbf{W}_{\text{ICA}} = \mathbf{A}^{-1}$ for the noiseless case is unbiased and $E(s_i^2) = 1$:

$$\text{MSE}_{\text{ICA}}|_{\sigma^2=0} \geq \sum_{i \neq j} \text{CRB}\left([\mathbf{WA}]_{i,j}\right) = \sum_{i \neq j} \frac{1}{L} \frac{\kappa_i}{\kappa_i \kappa_j - 1} \quad (16)$$

Next, we study the behaviour for a fixed SNR but vary the shape parameter $\beta$ of the GGD. Fig. 4 shows the averaged relative MSE of ICA as a function of the iteration number of NG-ICA and $\beta$ for an SNR of 10 dB. At this SNR, the MSE of the ICA solution decreases monotonically as we move away from $\beta = 2$ towards $\beta = 0.5$ or $\beta = 10$. This is different from the corresponding theoretical result in Fig. 2 which shows a local minimum of the MSE for $\beta = 4.62$. Two possible explanations are: Firstly, the theoretical MSE of ICA has been calculated using $\mathbf{W}_{\text{ICA}}$ from (10) which neglects terms $O(\sigma^4)$. For an SNR of 10 dB these terms might be important. Secondly, the variance of $\mathbf{W}_{\text{ICA}}$ due to limited amount of data might depend on $\kappa$ and might be smaller for $\kappa \to \infty$ as in the noiseless case (16). Furthermore, Fig. 4 shows a minimum of the relative MSE over the iteration number and its value is very close to one. This behaviour only occurs for symmetric $\mathbf{A}$ and depends on the initialization. For nonsymmetric $\mathbf{A}$, separation of Gaussian sources ($\beta = 2$) fails.

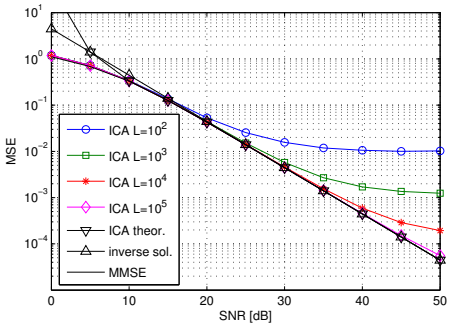



**Fig. 3**: MSE of NG-ICA for $N = 2$ sources with Laplacian pdf

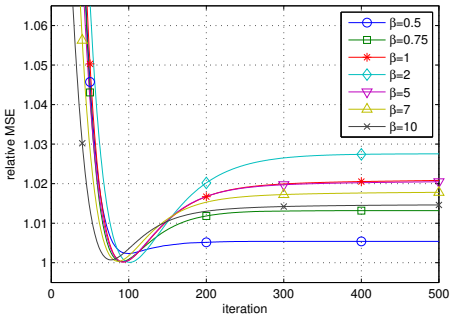



**Fig. 4**: Relative MSE of NG-ICA for $N = 2$ sources with GGD, SNR=10 dB, $L = 10^4$

### 5.3. Mismatch of $\varphi(s)$

Here, we consider the case that the assumed pdfs $q_i(s_i)$ and corresponding $\varphi_i(s_i)$ in the KLD-based ICA do not match the true pdfs of the sources. However, we assume that these "wrong" pdfs still result in the solution $\mathbf{W}_{\text{ICA}} = \mathbf{A}^{-1}$ for the noiseless case. Then our Taylor expansion approach is still valid and (24) will change to

$$\begin{aligned}\mathbf{C}^T + \text{diag}(\boldsymbol{\kappa})\mathbf{C} + \text{diag}(\boldsymbol{\rho} - \boldsymbol{\kappa})\text{Diag}(\mathbf{C}) \\ = \mathbf{I} - \text{diag}(\boldsymbol{\xi}) - \text{diag}(\boldsymbol{\kappa})\mathbf{R}_{\tilde{\mathbf{v}}} - \frac{1}{2}\text{diag}(\boldsymbol{\lambda})\text{Diag}(\mathbf{R}_{\tilde{\mathbf{v}}}) \quad (17)\end{aligned}$$

with $\boldsymbol{\xi} = [\xi_1, \cdots, \xi_N]^T$ and $\xi_i = E(\varphi_i(s_i)s_i)$. (27) remains unchanged and (25) will be

$$C_{ii} = \frac{1 - \xi_i}{1 + \rho_i} - \frac{\kappa_i + \frac{1}{2}\lambda_i}{1 + \rho_i}\left[\mathbf{R}_{\tilde{\mathbf{v}}}\right]_{ii} \quad (18)$$

## 6. CONCLUSION

In this paper, we have derived an analytic expression for the demixing matrix obtained from KLD-based ICA for the low noise regime. Furthermore, we have derived the corresponding MSE of the demixed signals and have shown its relationship to the MMSE solution. Although KLD and MMSE differ, linear demixing based on these two criteria yields demixed signals with similar MSE. Simulation results have verified that $\text{MSE}_{\text{ICA}}$ is indeed close to $\text{MSE}_{\text{MMSE}}$ for a wide range of the shape parameter of GGD. We also have shown that the estimation variance of $\mathbf{W}_{\text{ICA}}$ plays an important role in practice. We have provided theoretical results for the case when the nonlinearity in the ICA algorithm does not match the true pdf of the sources. In future, these results may be used to derive nonlinear functions that can achieve an MSE even closer to $\text{MSE}_{\text{MMSE}}$.

## A. PROOF OF (10)

### A.1. Taylor series approximation of $E(\boldsymbol{\varphi}(\mathbf{y})\mathbf{y}^T)$ in (9)

For simplicity, we use the notation $\mathbf{W} = \mathbf{W}_{\text{ICA}}$, $\mathbf{y} = \mathbf{y}_{\text{ICA}} = \mathbf{WAs} + \mathbf{Wv} = \hat{\mathbf{y}} + \mathbf{Wv}$. With (8) and $\mathbf{y} - \hat{\mathbf{y}} = \mathbf{Wv}$, we first get the Taylor series approximation for $\varphi_i(y_i)$

$$\boldsymbol{\varphi}(\mathbf{y}) = \boldsymbol{\varphi}(\hat{\mathbf{y}}) + \text{diag}(\boldsymbol{\varphi}'(\hat{\mathbf{y}}))\mathbf{Wv} + \frac{1}{2}\text{diag}(\boldsymbol{\varphi}''(\hat{\mathbf{y}})) \cdot (\mathbf{Wv} \odot \mathbf{Wv}) + O_p(\sigma^3) \quad (19)$$

with $\boldsymbol{\varphi}'(\mathbf{y}) = [\varphi_1'(y_1), \cdots, \varphi_N'(y_N)]^T$, $\boldsymbol{\varphi}''(\mathbf{y}) = [\varphi_1''(y_1), \cdots, \varphi_N''(y_N)]^T$. $\text{diag}(\mathbf{a})$ is a diagonal matrix with the elements of the vector $\mathbf{a}$ and $\odot$ denotes the elementwise multiplication. $O_p(\sigma^3)$ is the order in probability notation, i.e. $\lim_{\sigma\to 0} P(|O_p(\sigma^3)/\sigma^3| \geq \epsilon) = 0$. Multiplying (19) with $\mathbf{y}^T = (\hat{\mathbf{y}} + \mathbf{Wv})^T$ and taking the expectation, we get from (9)

$$\begin{aligned}\mathbf{I} &= E\left[\boldsymbol{\varphi}(\mathbf{y})\mathbf{y}^T\right] \\ &= E\left[\boldsymbol{\varphi}(\hat{\mathbf{y}})\hat{\mathbf{y}}^T\right] + E\left[\text{diag}(\boldsymbol{\varphi}'(\hat{\mathbf{y}}))\mathbf{Wvv}^T\mathbf{W}^T\right] \\ &\quad + \frac{1}{2}E\left[\text{diag}(\boldsymbol{\varphi}''(\hat{\mathbf{y}})) \cdot (\mathbf{Wv} \odot \mathbf{Wv})\,\hat{\mathbf{y}}^T\right] + O(\sigma^3) \quad (20)\end{aligned}$$

since $\hat{\mathbf{y}} = \mathbf{WAs}$ is independent of $\mathbf{v}$ and $E(\mathbf{v}) = \mathbf{0}$. After some straightforward manipulations, we get

$$\mathbf{I} = E\left[\boldsymbol{\varphi}(\hat{\mathbf{y}})\hat{\mathbf{y}}^T\right] + \sigma^2 E\left[\text{diag}(\boldsymbol{\varphi}'(\hat{\mathbf{y}}))\right]\mathbf{WR_v}\mathbf{W}^T + \frac{1}{2}\sigma^2 E\left[\boldsymbol{\varphi}''(\hat{\mathbf{y}})\hat{\mathbf{y}}^T\right] \cdot \text{Diag}\left(\mathbf{WR_v}\mathbf{W}^T\right) + O(\sigma^3), \quad (21)$$

where $\text{Diag}(\mathbf{Z})$ deletes all off-diagonal elements of the matrix $\mathbf{Z}$.
In general, if $E(v_1^{k_1}\cdots v_N^{k_N}) = 0 \quad \forall k_1 + \cdots + k_N$ odd (as assumed in Sec. 2), the expectation of all odd terms of $\mathbf{v}$ is zero and the Taylor series (21) and hence also $\mathbf{W}_{\text{ICA}}$ is only a function of $\sigma^2$.

Using a second Taylor series expansion of $\mathbf{W}_{\text{ICA}}(\sigma^2)$ at $\sigma^2 = 0$, we can write $\mathbf{W}_{\text{ICA}}(\sigma^2) = \mathbf{A}^{-1} + \sigma^2\mathbf{B} + O(\sigma^4)$ with $\mathbf{B} = \left.\frac{d\mathbf{W}_{\text{ICA}}(\sigma^2)}{d\sigma^2}\right|_{\sigma^2=0}$. Since we do not know $\mathbf{W}_{\text{ICA}}$ explicitely, it is hard to compute $\mathbf{B}$ directly by derivative. We do that by a second perturbation analysis in the next section.

### A.2. Taylor series approximation of (21) at $\mathbf{W}_{\text{ICA}} = \mathbf{A}^{-1} + \sigma^2\mathbf{B}$

Since $\mathbf{W} = \mathbf{W}_{\text{ICA}} = \mathbf{A}^{-1} + \sigma^2\mathbf{B} + O(\sigma^4)$,

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{WAs} = \mathbf{s} + \sigma^2\mathbf{BAs} + O(\sigma^4) \\ &= \mathbf{s} + \sigma^2\mathbf{Cs} + O(\sigma^4) = \mathbf{s} + \sigma^2\mathbf{b} + O(\sigma^4)\end{aligned}$$

with $\mathbf{C} = \mathbf{BA}$, we get $G_{ij} = \left[E(\boldsymbol{\varphi}(\hat{\mathbf{y}})\hat{\mathbf{y}}^T)\right]_{ij}$ with a second Taylor series:

$$\begin{aligned}G_{ij} &= E\left[(\varphi_i(s_i) + \varphi_i'(s_i)\sigma^2 b_i)(s_j + \sigma^2 b_j)\right] + O(\sigma^4) \\ &= E\left[\varphi_i(s_i)s_j\right] + \sigma^2 E\left[\varphi_i(s_i)b_j\right] + \sigma^2 E\left[\varphi_i'(s_i)b_i s_j\right] + O(\sigma^4)\end{aligned} \quad (22)$$

with $b_j = \sum_{l=1}^N C_{jl}s_l$. Because $s_i$ and $s_j$ are independent and zero mean, it holds

$$E\left[\varphi_i(s_i)s_j\right] = \begin{cases}1 & i = j \\ 0 & i \neq j\end{cases}$$

$$E\left[\varphi_i(s_i)b_j\right] = \sum_{l=1}^N C_{jl}E(\varphi_i(s_i)s_l) = C_{ji}$$

$$E\left[\varphi_i'(s_i)b_i s_j\right] = \sum_{l=1}^N C_{il}E(\varphi_i'(s_i)s_l s_j) = \begin{cases}\rho_i C_{ii} & i = j \\ \kappa_i C_{ij} & i \neq j\end{cases}$$

with $\kappa_i = E(\varphi_i'(s_i))$ and $\rho_i = E(\varphi_i'(s_i)s_i^2)$. The result is thus

$$E(\boldsymbol{\varphi}(\hat{\mathbf{y}})\hat{\mathbf{y}}^T) = \mathbf{I} + \sigma^2(\mathbf{C}^T + \text{diag}(\boldsymbol{\kappa})\mathbf{C}) + \sigma^2\text{diag}(\boldsymbol{\rho} - \boldsymbol{\kappa})\text{Diag}(\mathbf{C}) + O(\sigma^4) \quad (23)$$

with $\boldsymbol{\kappa} = [\kappa_1, \cdots, \kappa_N]^T$ and $\boldsymbol{\rho} = [\rho_1, \cdots, \rho_N]^T$. With $\mathbf{W} = \mathbf{A}^{-1} + \sigma^2\mathbf{B}$ and neglecting $O(\sigma^4)$, we get from (21)

$$E\left[\text{diag}(\boldsymbol{\varphi}'(\hat{\mathbf{y}}))\right]\mathbf{WR_v}\mathbf{W}^T = \text{diag}(\boldsymbol{\kappa})\mathbf{A}^{-1}\mathbf{R_v}\mathbf{A}^{-T},$$

$$\frac{1}{2}E\left[\boldsymbol{\varphi}''(\hat{\mathbf{y}})\hat{\mathbf{y}}^T\right]\text{Diag}(\mathbf{WR_v}\mathbf{W}^T) = \frac{1}{2}\text{diag}(\boldsymbol{\lambda})\text{Diag}(\mathbf{A}^{-1}\mathbf{R_v}\mathbf{A}^{-T})$$

with $\boldsymbol{\lambda} = [\lambda_1, \cdots, \lambda_N]^T$ and $\lambda_i = E(\varphi_i''(s_i)s_i)$. In summary, (21) simplifies to

$$\mathbf{C}^T + \text{diag}(\boldsymbol{\kappa})\mathbf{C} + \text{diag}(\boldsymbol{\rho} - \boldsymbol{\kappa})\text{Diag}(\mathbf{C}) = -\text{diag}(\boldsymbol{\kappa})\mathbf{R}_{\tilde{\mathbf{v}}} - \frac{1}{2}\text{diag}(\boldsymbol{\lambda})\text{Diag}(\mathbf{R}_{\tilde{\mathbf{v}}}) \quad (24)$$

with $\mathbf{R}_{\tilde{\mathbf{v}}} = \mathbf{A}^{-1}\mathbf{R_v}\mathbf{A}^{-T}$. The diagonal elements $(i,i)$ of (24) are

$$C_{ii} = -\frac{\kappa_i + \frac{1}{2}\lambda_i}{1 + \rho_i}\left[\mathbf{R}_{\tilde{\mathbf{v}}}\right]_{ii}. \quad (25)$$

The non-diagonal elements $(i,j)$ and $(j,i)$ of (24) are

$$\begin{bmatrix}\kappa_i & 1 \\ 1 & \kappa_j\end{bmatrix}\begin{bmatrix}C_{ij} \\ C_{ji}\end{bmatrix} = -\begin{bmatrix}\kappa_i \\ \kappa_j\end{bmatrix}\left[\mathbf{R}_{\tilde{\mathbf{v}}}\right]_{ij} \quad (26)$$

If both $s_i$ and $s_j$ were Gaussian, $\kappa_i = \kappa_j = 1$ and we would get $C_{ij} + C_{ji} = -\left[\mathbf{R}_{\tilde{\mathbf{v}}}\right]_{ij}$. The solution for $C_{ij}$ would then not be unique. Hence, $\mathbf{W}_{\text{ICA}}$ is ambiguous if there is more than one Gaussian source signal $s_i$ in $\mathbf{s}$.
If at least one of $s_i$ and $s_j$ is non-Gaussian, $\kappa_i\kappa_j > 1$ and we get

$$\begin{bmatrix}C_{ij} \\ C_{ji}\end{bmatrix} = -\frac{1}{\kappa_i\kappa_j - 1}\begin{bmatrix}\kappa_j(\kappa_i - 1) \\ \kappa_i(\kappa_j - 1)\end{bmatrix}\left[\mathbf{R}_{\tilde{\mathbf{v}}}\right]_{ij} \quad (27)$$

Hence we can write

$$\mathbf{W}_{\text{ICA}} = (\mathbf{I} + \sigma^2\mathbf{C})\mathbf{A}^{-1} + O(\sigma^4) = (\mathbf{I} - \sigma^2\mathbf{M} \odot \mathbf{R}_{\tilde{\mathbf{v}}})\mathbf{A}^{-1} + O(\sigma^4) \quad (28)$$

with

$$M_{ii} = \frac{\kappa_i + \frac{1}{2}\lambda_i}{1 + \rho_i}, \quad M_{ij} = \frac{\kappa_j(\kappa_i - 1)}{\kappa_i\kappa_j - 1} \quad i \neq j. \quad \blacksquare$$

**Acknowledgment:** The authors would like to thank P. Tichavsky for sharing his MATLAB code for GGD random number generation.

## REFERENCES

[1] S.C. Douglas, A. Cichocki, and S. Amari, "A bias removal technique for blind source separation with noisy measurements," *Eletronics Letters*, vol. 34, pp. 1379–1380, July 1998.

[2] A. Hyvärinen, "Independent component analysis in the presence of Gaussian noise by maximizing joint likelihood," *Neurocomputing*, vol. 22, pp. 49–67, 1998.

[3] Z. Koldovsky and P. Tichavsky, "Blind instantaneous noisy mixture separation with best interference-plus-noise rejection," *Proc. ICA 2007*, Sept. 2007.

[4] Z. Koldovsky and P. Tichavsky, "Asymptotic analysis of bias of FastICA-based algorithms in the presence of additive noise," *Technical Report 2181, UTIA, AV CR, 2007*, 2007.

[5] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[6] P. Tichavsky, Z. Koldovsky, and E. Oja, "Performance analysis of the FastICA algorithm and Cramer-Rao bounds for linear independent component analysis," *IEEE Trans. on Signal Processing*, vol. 54, no. 4, Apr. 2006.