Independent Vector Analysis: Algorithms and Applications

by

Zois Boukouvalas

Ph.D. Preliminary Defense Report

Department of Mathematics and Statistics

University of Maryland, Baltimore County

April 23, 2015

Contents

1 Introduction

Independent vector analysis (IVA) is a recent extension of independent component analysis (ICA) that makes full use of the statistical dependence across multiple datasets to achieve source separation. It is an effective solution to the joint blind source separation (JBSS) problem and promises to be an effective solution to many problems where JBSS solutions are needed. These include detection of a target in a given video sequence or multi-spectral remote sensing data, and analysis of brain activity using medical image data collected from multiple subjects among many others. IVA can be formulated in a maximum likelihood (ML) framework such that all available types of diversity are taken into account simultaneously. The key issue that will enable the application of IVA to many problems with effective solutions is development of effective models for the underlying source density and their estimation. In addition, efficient use of available prior information will greatly increase the utility of IVA.

The multivariate generalized Gaussian distribution (MGGD) provides an effective model for IVA's latent multivariate sources, however, its performance highly depends on the estimation of the source parameters. Therefore, in our preliminary work, we discuss techniques to estimate the MGGD parameters and successfully integrate them into IVA. MGGD depends on two parameters, the scatter matrix (which is symmetric positive definite) and the shape parameter. Current methods attempt to estimate the scatter matrix for a given value of the shape parameter. However, their accuracy significantly suffers when the value of the shape parameter increases. We propose two different maximum likelihood (ML) techniques to overcome this issue. The first is based on the Fisher scoring (FS) algorithm and the second on a fixed point (FP) algorithm. We integrate both techniques into IVA to accurately estimate all the parameters of the MGGD sources simultaneously, which in turn leads to effective calculation of the IVA score and cost functions, resulting in better IVA performance.

Using our previous work as a starting point, we propose to extend IVA in two major directions: the design of a more general multivariate probability density function estimator and the investi-

gation of how prior knowledge can be incorporated into the IVA model. To generalize the source model, we propose an estimation method based on the maximum entropy to successfully match multivariate sources from a wide range of distributions. This approach is quite different from the MGGD-based approach described in the preliminary work, where a density model is chosen and the parameters are estimated during the adaption of the demixing matrix. Then, to effectively include prior knowledge, we propose to design a new IVA framework. This can be achieved by constraining the density models through the chosen distribution or by an optimization framework that directly implements each constraint through Lagrange multipliers. We also propose the use of relational constraints within the IVA framework, an approach that will eliminate the need to define explicit priors.

2 Preliminary work

2.1 MGGD parameter estimation

2.1.1 Introduction

Effective estimation of the parameters of multivariate generalized Gaussian distribution (MGGD) is vital to many signal processing applications — recall that MGGD depends on two parameters, the scatter matrix (which is symmetric positive definite) and the shape parameter. Even though the MGGD has a simple parametric form, it provides a model sufficiently flexible for most applications, and hence it has been an attractive solution for many applications in signal processing such as in video coding [1], denoising [2] and biomedical signal processing [3].

Existing approaches to this problem, see [4],[5] and [6–10], attempt to estimate the scatter matrix for a given value of the shape parameter. However, their accuracy suffers notable limitations when the value of shape parameter becomes large, which makes them unsuitable for many applications. The main purpose of this section is to present an effective new method that yields accurate estimates of the scatter matrix for any value of the shape parameter. In [4],[5], method of moments (MoM) and maximum likelihood (ML) techniques, were explored. In [6], authors prove that the scatter matrix ML estimator exists and is unique for any value of the shape parameter that belongs to the interval $(0, 1)$, i.e., when the marginals have distributions ranging from very peaky and heavy tailed to Gaussian (for shape parameter 1). In addition, a fixed point algorithm (ML-FP) has been introduced for estimating the parameters of an MGGD. Simulation results reveal the unbiasedness and consistency of the ML estimators of the scatter matrix and the shape parameter of the distribution. However, since many applications require values of the shape parameter that are greater than one, i.e., flatter distributions as well, in this section we propose two ML algorithms for estimating the scatter matrix for any given value of the shape parameter.

The first algorithm is a variation of the Newton-Raphson optimization algorithm called maximum likelihood Fisher scoring (ML-FS) algorithm. The main difference between the classical

Newton-Raphson algorithm and ML-FS is that the negative inverse Hessian has been replaced by the Fisher information matrix. The second algorithm is a fixed point algorithm called Riemannian averaged fixed point (RA-FP). The main idea is that RA-FP implements successive Riemannian averages of fixed-point iterates, preventing them from diverging away from the true value of the scatter matrix.

2.1.2 Background

The probability density function of an MGGD is given by [11]

$$p(\mathbf{y};\boldsymbol{\Sigma},\beta,m) = \frac{\Gamma\left(\frac{K}{2}\right)}{\pi^{\frac{K}{2}}\Gamma\left(\frac{K}{2\beta}\right)2^{\frac{K}{2\beta}}}\frac{\beta}{m^{\frac{K}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\exp\left[-\frac{1}{2m^{\beta}}\left(\mathbf{y}^{\top}\boldsymbol{\Sigma}^{-1}\mathbf{y}\right)^{\beta}\right],$$

where K is the dimension of the probability space, $\mathbf{y} \in \mathbb{R}^K$ is a random vector, m is the scale parameter, $\beta > 0$ is the shape parameter that controls the peakedness and the spread of the distribution and $\boldsymbol{\Sigma}$ is a $K \times K$ symmetric positive scatter matrix. If $\beta = 1$, the MGGD is equivalent to the multivariate Gaussian and the matrix $\boldsymbol{\Sigma}$ becomes the covariance matrix. If $\beta < 1$ the distribution of the marginals is more peaky and has heavier tails, while $\beta > 1$ is less peaky and has lighter tails.

Let $(\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T)$ be a random sample of T observation vectors of dimension K, drawn from a zero mean MGGD with parameters β and $\boldsymbol{\Sigma}$, and m. The corresponding ML estimates $\hat{\beta}$, $\hat{\boldsymbol{\Sigma}}$ and $\hat{m}$ are found by solving the ML equations, described in the following. Assume first β is known. The ML estimate $\hat{\boldsymbol{\Sigma}}$ is the solution of the following equation [6]

$$\boldsymbol{\Sigma} = \sum_{i=1}^{N}\frac{p}{u_i + u_i^{1-\beta}\sum_{i\neq j}u_j^{\beta}}\mathbf{y}_i\mathbf{y}_i^{\top}, \tag{1}$$

for unknown $\mathbf{\Sigma}$, where $u_i = \mathbf{y}_i^\top \mathbf{\Sigma}^{-1} \mathbf{y}_i$. Once $\hat{\mathbf{\Sigma}}$ has been computed, $\hat{m}$ is immediately given by

$$\hat{m} = \left(\frac{1}{N} \sum_{i=1}^{N} \hat{u}_i^{\beta} \right)^{\frac{1}{\beta}}, \tag{2}$$

where $\hat{u}_i = \mathbf{y}_i^\top \hat{\mathbf{\Sigma}}^{-1} \mathbf{y}_i$.

In the general case where β is unknown, $\hat{\mathbf{\Sigma}}$ and $\hat{\beta}$ are found by solving equation (1), along with

$$\gamma(\beta) = \frac{pN}{2 \sum_{i=1}^{N} u_i^{\beta}} \sum_{i=1}^{N} \left[u_i^{\beta} \ln(u_i) \right] - \frac{pN}{2\beta} \left[\Psi\left(\frac{p}{2\beta} \right) + \ln 2 \right] - N - \frac{pN}{2\beta} \ln\left(\frac{\beta}{pN} \sum_{i=1}^{N} u_i^{\beta} \right) = 0, \tag{3}$$

whose solution is $\hat{\beta}$. Here, Ψ is the digamma function. Once the solutions $\hat{\mathbf{\Sigma}}$ and $\hat{\beta}$ of (1) and (3) have been found, $\hat{m}$ is computed directly from (2).

It is seen from the above that the main difficulty, in the computation of $\hat{\beta}$, $\hat{\mathbf{\Sigma}}$ and $\hat{m}$, lies in solving equation (1). There is no closed form solution for the ML estimation of these parameters. Hence, we use an iterative scheme to simultaneously estimate $\mathbf{\Sigma}$ and β using (1) and (3) respectively as discussed in the next section.

2.1.3 ML-FS algorithm

The likelihood function of $(\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T)$ where m has been replaced by its estimate in (2) is given by

$$\mathcal{L}(\mathbf{\Sigma}; \Theta) = \prod_{i=1}^{T} p(\mathbf{y}_i; \Theta) = \left[\frac{\beta \Gamma\left(\frac{K}{2}\right)}{\pi^{\frac{K}{2}} \Gamma\left(\frac{K}{2\beta}\right) 2^{\frac{K}{2\beta}}} \right]^T \left(\frac{KT}{\beta} \right)^{\frac{KT}{2\beta}} \exp\left(-\frac{KT}{2\beta} \right) \times \left[\frac{1}{|\mathbf{\Sigma}|} \left(\sum_{i=1}^{T} u_i^{\beta} \right)^{-\frac{K}{\beta}} \right]^{\frac{T}{2}}, \tag{4}$$

where Θ denotes the parameter space that contains the entries of the scatter matrix $\mathbf{\Sigma}$, m and β. By fixing β, the likelihood function depends only on the entries of $\mathbf{\Sigma}$. Defining the gradient of the

likelihood function similar to [6], we introduce the functional

$$\begin{aligned} F : \mathcal{S}_+^K &\to \mathbb{R}^+ \setminus \{0\} \\ \mathbf{\Sigma} &\mapsto |\mathbf{\Sigma}|^{-1} \left(\sum_{i=1}^{T} u_i^{\beta} \right)^{-\frac{K}{\beta}}, \end{aligned}$$

where $\mathcal{S}_+^K$ is the space of all real $K \times K$ symmetric and positive definite matrices. By omitting the constant term from (4), the gradient of the likelihood can then be written as

$$\nabla \mathcal{L}(\mathbf{\Sigma}; \Theta) = [F(\mathbf{\Sigma})]^{\frac{T-2}{2}} \nabla F(\mathbf{\Sigma}),$$

and the gradient of F at a point $\mathbf{\Sigma} \in \mathcal{S}_+^K$ is given by [6]

$$\nabla F(\mathbf{\Sigma}) = F(\mathbf{\Sigma}) \mathbf{\Sigma}^{-1} [f(\mathbf{\Sigma}) - \mathbf{\Sigma}] \mathbf{\Sigma}^{-1},$$

where

$$\begin{aligned} f : \mathcal{S}_+^K &\to \mathcal{S}_{++}^K \\ \mathbf{\Sigma} &\mapsto \sum_{i=1}^{T} \frac{K}{u_i + u_i^{1-\beta} \sum_{i \neq j} u_j^{\beta}} \mathbf{y}_i \mathbf{y}_i^{\top}. \end{aligned}$$

To numerically maximize the likelihood function we use a variation of the Newton-Raphson optimization algorithm called the ML-FS algorithm. Since the negative inverse Hessian has been replaced by the inverse of the Fisher information matrix, we need to calcultate its entries.

To calculate the entries of the Fisher information matrix, we first define the manifold $\mathcal{M}$ of a zero-mean MGGD as well as an associated metric. The MGGD manifold is parameterized by β and the matrix $\mathbf{\Sigma}$. In our particular case since β is fixed, $\mathcal{M}$ is parameterized only by the entries of $\mathbf{\Sigma} \in \mathcal{S}_+^K$, so $\mathcal{M}$ is isomorphic to $\mathbb{R}^n$ where $n = \frac{K(K+1)}{2}$. Each point that lies on $\mathcal{M}$ is a probability density function. To measure the distance between two pdfs we need to calculate the length of the curve that connects those two points and has minimum length. This curve is called a geodesic path

and is determined through the elements of the Fisher information matrix. Thus, if $\Theta = (\theta_1, \theta_2, ..., \theta_n)$ denotes the parameter space of $\mathcal{M}$, the Fisher metric is defined by the matrix elements

$$\mathbf{G}_{ij}(\theta) = -E\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln p(\mathbf{y};\Theta)\right\}, \quad i, j = 1, ...n,$$

where $\mathbf{G}$ is an $n \times n$ matrix. Trying to define geodesic paths within an MGGD manifold, [4] proposed a simpler way for the elements of the metric defined on a K-dimensional sub-manifold. Thus the entries of the Fisher information matrix are defined by

$$\mathbf{G}_{ii}(\beta) = \frac{1}{4}\left(\frac{3K + 6\beta}{K + 2} - 1\right), \tag{5}$$

and

$$\mathbf{G}_{ij}(\beta) = \frac{1}{4}\left(\frac{K + 2\beta}{K + 2} - 1\right), \; i \neq j, \tag{6}$$

for $i, j = 1, ..., K$. From (5) and (6), we see that Fisher information matrix depends only on the fixed value of β and the dimension K. This results in the reduction of the computational cost since the update of the inverse of the Fisher information matrix is not required. Fisher scoring iteration is defined as

$$\mathbf{\Sigma}^{(k+1)} = \mathbf{\Sigma}^{(k)} + \mathbf{G}^{-1}\nabla\mathcal{L}(\mathbf{\Sigma};\Theta). \tag{7}$$

MoM provides an effective and efficient initialization for the ML-FS algorithm with only a few steps of the algorithm being sufficient to obtain good estimates.

A pseudo-code description of the ML-FS algorithm is given in Algorithm 2 below. The main part of this algorithm is the loop described in lines 4-11. The algorithm exits this loop when $D(k) <$ tol, where $D(k)$ is the relative difference between two successive estimates, and tol is a tolerance bound, chosen by the user. The loop is also terminated whenever the number of iterations exceeds a pre-defined upper bound $\mathrm{N_{max}}$.

Algorithm 1 ML-FS

Input: $\mathbf{X} \in \mathbb{R}^{K\times T}$, optionally β
Initialize $\mathbf{\Sigma}$ using MoM
If β is not given initialize both $\mathbf{\Sigma}$ and β using MoM
while ($D(k) >$ tol) and ($k < \mathrm{N}_{\max}$) **do**
    Estimate $\mathbf{\Sigma}$ using one iteration of (7)
    if β is not given **then**
        Estimate β by applying Newton-Raphson into (3)
    else
        Go to step 5
    end
end
Using $\mathbf{\Sigma}$ and β, estimate m with (2)
Output: $\mathbf{\Sigma}$, β, m

2.1.4 RA-FP algorithm

With regard to ML estimators, it has become clear, from [6–10], that they can be computed using FP algorithms. As in [6][10], it is possible to formulate (1) as a fixed point equation. Considering the function (5) we observe that is just the right hand side of equation (1). Therefore, this equation can be written

$$\mathbf{\Sigma} = f(\mathbf{\Sigma}) \tag{8}$$

which is indeed a fixed point equation. In other words, the ML estimate $\hat{\mathbf{\Sigma}}$ is the solution of the fixed point equation (8) associated with the function f defined in (5). It is well-known that the solution of a fixed point equation, such as (8) may be attempted using an FP algorithm, which gives successive fixed point iterates

$$\mathbf{\Sigma}_{k+1} = f(\mathbf{\Sigma}_k) \qquad k = 0, 1, 2, \ldots \tag{9}$$

Indeed, this algorithm was used in [6][10]. Concretely, it consists in repeating (9) until the iterates $\mathbf{\Sigma}_k$ stabilize. That is, until there no sensible difference between $\mathbf{\Sigma}_k$ and $\mathbf{\Sigma}_{k+1}$.

The convergence of the FP algorithm (9) depends on the function f being contractive. In the

present context, numerical experiments show the function f, (which depends on β as can be seen in (5)), is not contractive when $\beta \geq 2$. The proposed RA-FP algorithm, overcomes this difficulty. It has been shown in [6][10], that the FP algorithm (9) gives accurate estimates of $\mathbf{\Sigma}$ when $\beta < 2$. The main contribution of the present paper is to describe the new RA-FP algorithm, which is a generalization of the FP algorithm (9), and is capable of producing accurate estimates of $\mathbf{\Sigma}$ when $\beta \geq 2$.

The RA-FP algorithm uses the Riemannian geometry of the space $\mathcal{S}_+^K$. Precisely, it implements successive Riemannian averages of fixed point iterates. The definition of the Riemannian average of $\mathbf{P}, \mathbf{Q} \in \mathcal{S}_+^p$ is the following. For $t \in [0, 1]$, the Riemannian average with ratio t of $\mathbf{P}$ and $\mathbf{Q}$ is $\mathbf{P}\#_t\mathbf{Q}$, given as in [10]

$$\mathbf{P}\#_t\mathbf{Q} = \mathbf{P}^{1/2} \left(\mathbf{P}^{-1/2}\mathbf{Q}\mathbf{P}^{-1/2}\right)^t \mathbf{P}^{1/2}, \tag{10}$$

where, on the right hand side, the exponent $(\ldots)^t$ denotes elevation of a symmetric matrix to the power t. Note that

$$\mathbf{P}\#_0\mathbf{Q} = \mathbf{P} \qquad \mathbf{P}\#_1\mathbf{Q} = \mathbf{Q} \tag{11}$$

The RA-FP algorithm is defined as follows. When $\mathbf{\Sigma}_k$ is given, instead of of defining $\mathbf{\Sigma}_{k+1}$ by (9), let

$$\mathbf{\Sigma}_{k+1} = \mathbf{\Sigma}_k \#_{t_k} f(\mathbf{\Sigma}_k), \tag{12}$$

where $t_k \in [0, 1]$. The RA-FP algorithm (12) is indeed a generalization of the FP algorithm (9), since putting $t_k = 1$ in (12) yields (9), as can be seen from (11). In our work, we set

$$t_k = \frac{1}{k+1} \tag{13}$$

Similarly to ML-FS a pseudo-code description of the RA-FP algorithm is given in Algorithm 2 below.

Algorithm 2 RA-FP

Input: $\mathbf{X} \in \mathbb{R}^{K \times T}$, optionally β
Initialize $\mathbf{\Sigma}$ using either MoM or $\mathbf{\Sigma} = \mathbf{I}_p$
If β is not given initialize both $\mathbf{\Sigma}$ and β using MoM
while ($D(k) >$ tol) and ($k < \mathrm{N}_{\max}$) **do**
    Estimate $\mathbf{\Sigma}$ using one iteration of (12)
    if β is not given **then**
        Estimate β by applying Newton-Raphson into (3)
    else
        Go to step 5
    end
end
Using $\mathbf{\Sigma}$ and β, estimate m with (2)
Output: $\mathbf{\Sigma}$, β, m

In the next section, we provide a rigorous mathematical proof of the convergence of the RA-FP algorithm.

2.1.5 Convergence of RA-FP

The proof essentially relies on the Riemannian geometry of the space $\mathcal{S}_+^K$, the space of symmetric positive definite, $K \times K$ real matrices [12][13]. The main geometric property to be used is the *strong convexity of Rao's distance* [14], which is now explained.

To begin, the length of a differentiable curve $c : [0, 1] \to \mathcal{S}_+^K$ is defined as [12]

$$L(c) = \int_0^1 \|c^{-1}(t)\dot{c}(t)\|_F \times dt \tag{14}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Let $\mathbf{P}$ and $\mathbf{Q}$ be two points in $\mathcal{S}_+^K$. A curve c is said to connect $\mathbf{P}$ and $\mathbf{Q}$ if $c(0) = \mathbf{P}$ and $c(1) = \mathbf{Q}$. Among all curves connecting $\mathbf{P}$ and $\mathbf{Q}$, there exists a unique curve γ, whose length is minimum, (recall length is defined by (14)). This curve γ is called the *geodesic* connecting $\mathbf{P}$ and $\mathbf{Q}$. Its equation, in the notation of (10), is [13][10]

$$\gamma(t) = \mathbf{P}\#_t\mathbf{Q} \tag{15}$$

In particular, this exhibits the geometric meaning of the Riemannian average of $\mathbf{P}$ and $\mathbf{Q}$. The Riemannian average with ratio t of $\mathbf{P}$ and $\mathbf{Q}$ is the point $\gamma(t)$ lying on the geodesic γ connecting $\mathbf{P}$ and $\mathbf{Q}$.

Rao's distance between $\mathbf{P}$ and $\mathbf{Q}$, denoted $d(\mathbf{P}, \mathbf{Q})$ is the length of the geodesic curve γ, defined by (15). Using (14), it can be found analytically [12],

$$d(\mathbf{P}, \mathbf{Q}) = \| \log(\mathbf{P}^{-1/2}\mathbf{Q}\mathbf{P}^{-1/2}) \|_F \tag{16}$$

The main property of Rao's distance, used in the proof of convergence of the RA-FP algorithm is its strong convexity [14]. This is defined as follows. Let $\mathbf{R}, \mathbf{P}, \mathbf{Q} \in \mathcal{S}_+^K$ and $\gamma : [0, 1] \to \mathcal{S}_+^K$ the geodesic connecting $\mathbf{P}$ and $\mathbf{Q}$, given by (15). Then,

$$\begin{aligned} d^2(\mathbf{R}, \gamma(t)) \leq t\, d^2(\mathbf{R}, \mathbf{Q}) + (1 - t)\, d^2(\mathbf{R}, \mathbf{P}) \\ -t(1 - t) d^2(\mathbf{P}, \mathbf{Q}) \end{aligned} \tag{17}$$

This inequality simply means the function $t \mapsto d^2(\mathbf{R}, \gamma(t))$, which is a real-valued function of the real variable t, is a strongly convex function.

Consider now, once more, the fixed point equation (8). The FP algorithm (9), produces iterates $\mathbf{\Sigma}_k$ which converge to the unique fixed point $\hat{\mathbf{\Sigma}}$ of the function f, whenever f is contractive. That is, whenever [15]

$$d(f(\mathbf{P}), f(\mathbf{Q})) \leq \lambda \times d(\mathbf{P}, \mathbf{Q}) \qquad \lambda < 1 \tag{18}$$

for all $\mathbf{P}, \mathbf{Q} \in \mathcal{S}_+^K$. On the other hand, the FP algorithm (9) has no guarantee of convergence when $\lambda = 1$, in which case f is said to be non-expansive. Precisely, in this case [15],

$$d(f(\mathbf{P}), f(\mathbf{Q})) \leq d(\mathbf{P}, \mathbf{Q}) \tag{19}$$

for all $\mathbf{P}, \mathbf{Q} \in \mathcal{S}_+^K$. For the function f, as defined in (5), numerical experiments have shown that, in a neighborhood of $\hat{\mathbf{\Sigma}}$, this function is contractive when $\beta < 2$, but only non expansive, when $\beta \geq 2$. In this case, the FP algorithm (9) fails to converge to $\hat{\mathbf{\Sigma}}$, while the RA-FP algorithm (12) converges systematically.

The mathematical explanation of this convergence is given in the following proposition.

Proposition 1. *Let $f : \mathcal{S}_+^K \to \mathcal{S}_+^K$ be a function, which has a fixed point $\hat{\mathbf{\Sigma}}$. Assume there exists a neighborhood U of $\hat{\mathbf{\Sigma}}$, such that $\hat{\mathbf{\Sigma}}$ is the unique fixed point of f in U. Assume also f is non-expansive in U. That is, for $\mathbf{P}, \mathbf{Q} \in U$, inequality (19) holds. If $\mathbf{\Sigma}_0 \in U$ and, $\mathbf{\Sigma}_{k+1}$ is defined by the RA-FP algorithm (12), for $k = 0, 1, 2, \ldots$, then the sequence $\{\mathbf{\Sigma}_k\}$ remains in U and converges to $\hat{\mathbf{\Sigma}}$, as $k \to \infty$.*

Proof : Assume $\mathbf{\Sigma}_k \in U$. Since $\hat{\mathbf{\Sigma}} \in U$, It follows from (19),

$$d(f(\hat{\mathbf{\Sigma}}), f(\mathbf{\Sigma}_k)) \leq d(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}_k)$$

But $\hat{\mathbf{\Sigma}}$ is a fixed point of f, so $f(\hat{\mathbf{\Sigma}}) = \hat{\mathbf{\Sigma}}$. Replacing in the above inequality, it follows that

$$d(\hat{\mathbf{\Sigma}}, f(\mathbf{\Sigma}_k)) \leq d(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}_k) \tag{20}$$

Now, apply the strong convexity property (17), with $\mathbf{R} = \hat{\mathbf{\Sigma}}$, $\mathbf{P} = \mathbf{\Sigma}_k$, $\mathbf{Q} = f(\mathbf{\Sigma}_k)$, and $t = t_k$. Using (12) and (15), this gives

$$\begin{aligned} d^2(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}_{k+1}) \leq t_k\, d^2(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}_k) + (1 - t_k)\, d^2(\hat{\mathbf{\Sigma}}, f(\mathbf{\Sigma}_k)) \\ -t_k(1 - t_k) d^2(\mathbf{\Sigma}_k, f(\mathbf{\Sigma}_k)) \end{aligned}$$

Replacing (20) in this last ienquality, it folows after a short calculation

$$d^2(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}_k) - d^2(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}_{k+1}) \geq t_k(1 - t_k) d^2(\mathbf{\Sigma}_k, f(\mathbf{\Sigma}_k)) \tag{21}$$

This shows that $d(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}_{k+1}) \leq d(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}_k)$. So if $\mathbf{\Sigma}_k$ belongs to U, so does $\mathbf{\Sigma}_{k+1}$. Thus, if $\mathbf{\Sigma}_0 \in U$, then the sequence $\{\mathbf{\Sigma}_k\}$ remains in U. To prove this sequence converges to $\hat{\mathbf{\Sigma}}$, sum (21) over $k = 0, \ldots, n-1$. This gives,

$$d^2(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}_0) - d^2(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}_n) \geq \sum_{k=0}^{n-1} t_k(1-t_k) d^2(\mathbf{\Sigma}_k, f(\mathbf{\Sigma}_k)) \tag{22}$$

The right hand side of this inequality is bounded above by $d^2(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}_0)$, which does not depend on n. Therefore,

$$\sum_{k=0}^{\infty} t_k(1-t_k) d^2(\mathbf{\Sigma}_k, f(\mathbf{\Sigma}_k)) < +\infty \tag{23}$$

To complete the proof, take the neighborhood U of $\hat{\mathbf{\Sigma}}$ to be compact. This can be done without any loss of generality.

The sequence $\mathbf{\Sigma}_k$ converges to $\hat{\mathbf{\Sigma}}$ if and only if $d(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}_k) \to 0$. It is now shown that assuming this is not true would lead to a contradiction.

By (21), the sequence of distances $d(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}_k)$ is decreasing. Therefore, if it does not converge to 0, there exists a positive number δ such that $d(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}_k) \geq \delta$ for all k.

Let C be the set of matrices $\mathbf{\Sigma}$ such that $d(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}) \geq \delta$. This is a closed set. Therefore, the set $U \cap C$ is compact. Note the function $\mathbf{\Sigma} \mapsto d(\mathbf{\Sigma}, f(\mathbf{\Sigma}))$ is continuous. Therefore, this function reaches its minimum, say c, over $U \cap C$. Since $U \cap C$ does not contain any fixed points of f, it follows that $c > 0$.

It has been proved that $\mathbf{\Sigma}_k \in U$ for all k, and that, assuming $\mathbf{\Sigma}_k$ does not converge to $\hat{\mathbf{\Sigma}}$, $\mathbf{\Sigma}_k \in C$ for all k. In this case, $\mathbf{\Sigma}_k \in U \cap C$ for all k. This implies $d(\mathbf{\Sigma}_k, f(\mathbf{\Sigma}_k)) \geq c$ for all k. Replacing in the right hand side of (23),

$$\sum_{k=0}^{\infty} t_k(1-t_k) d^2(\mathbf{\Sigma}_k, f(\mathbf{\Sigma}_k)) \geq c^2 \sum_{k=0}^{\infty} t_k(1-t_k)$$

Since $t_k = \frac{1}{1+k}$, this sum is infinite, which contradicts (23).

Since the assumption that $d(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}_k)$ does not converge to zero has lead to a contradiction, it follows that $d(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}_k) \to 0$, which means that $\mathbf{\Sigma}_k$ converges to $\hat{\mathbf{\Sigma}}$. ■

For the function f defined in (5), the maximum likelihood estimate $\hat{\mathbf{\Sigma}}$ is a fixed point, as discussed after (8). Moreover, numerical experiments have shown the conditions of Proposition 1 apply to this function f. Proposition 1 therefore asserts that the RA-FP algorithm (12) converges to the maximum likelihood estimate $\hat{\mathbf{\Sigma}}$. This is in accord with the numerical results provided in the next section.

2.1.6 Experimental results

To quantify the performance of ML-FS and RA-FP, we generate data according to [6], [16] with $\mathbf{\Sigma}$ defined by

$$\mathbf{\Sigma}(i, j) = \sigma^{|i-j|}, \ i, j = \{1, 2, ...p - 1\}, \tag{24}$$

where σ belongs to the interval $[0, 1)$ and controls the correlations between the entries of the data. All results are averaged over 500 runs. For these types of experiments we used $p = 3$, $N = 10000$, and σ has been uniformly selected from the range $(0.4, 0.6)$.

Fig. 1 shows the difference between the estimated and the original matrix $\mathbf{\Sigma}$ as a function of shape parameter and for known value of β. The difference is measured by the Frobenius norm. It can be observed that for $\beta < 1$, RA-FP and ML-FP provide better results than the MoM and ML-FS, while for $\beta \geq 2$ RA-FP performs the best among these four algorithms. Fig. 2 displays the Frobenius norm between the estimated and the original scatter matrix, when $\mathbf{\Sigma}$ and β have been jointly estimated. When $\beta > 4$ RA-FP performs the best and for $\beta < 1$ the ML techniques perform better than MoM.

Fig.3 shows the comparison between the variances of the β estimators generated from the MoM, ML-FP, ML-FS, and RA-FP estimators as well as the Cramer-Rao lower bound (CRLB), as a function of different values of β. CRLB can be obtained by inverting the Fisher information

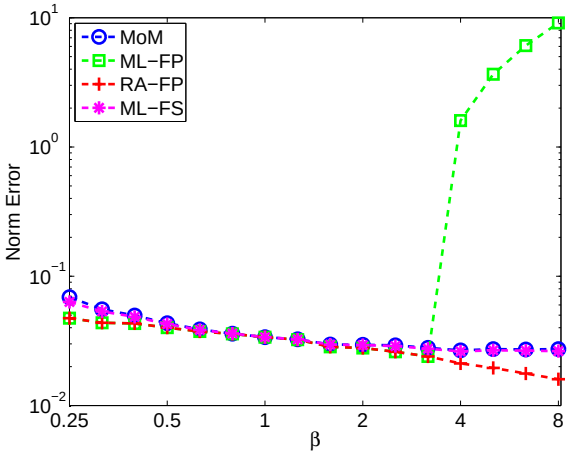

Figure 1: Scatter matrix estimation performance for different values of shape parameter, for $N = 10000$, $\sigma \in (0.4, 0.6)$.

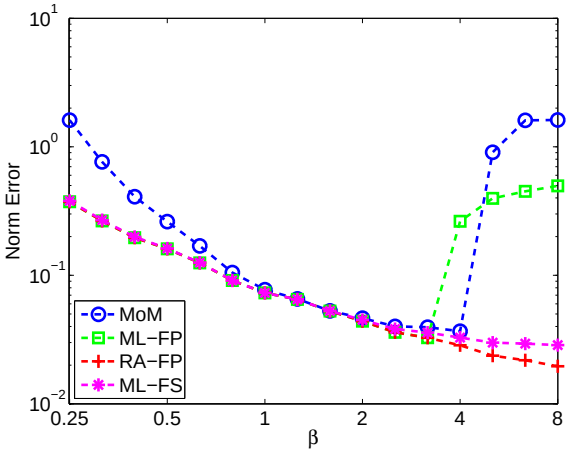

Figure 2: Scatter matrix estimation performance for different values of shape parameter when $\mathbf{\Sigma}$ and β have been jointly estimated. $N = 10000$, $\sigma \in (0.4, 0.6)$.

matrix derived by [4]. Numerical experiments [6], show the unbiasedness of the estimator β, so by Fig.3, we observe that the performance of the ML-FS and RA-FP (overlap) is very close to the CRLB, illustrating the MLE efficiency. On the other hand, the other two methods have issues when β moves away from one.

Fig. 4 shows the difference between the estimated and the true $\mathbf{\Sigma}$ as a function of β for various initial values. As observed for any value of $\beta \in (0.25, 8)$ algorithm remains invariant to the initial point. This can result in the reduction of the computational cost of the algorithm, since addition

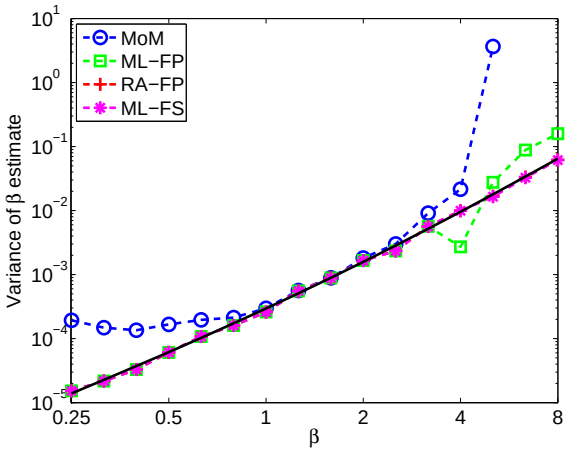

Figure 3: Variance of β estimate for different values of β. $T = 10000$, $\sigma \in (0.4, 0.6)$.

computations produced by the MoM can be avoided. Finally, Fig. 5 shows the number of iterations

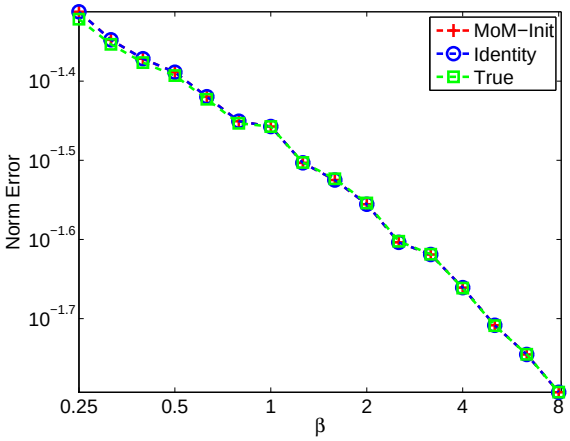

Figure 4: Scatter matrix estimation performance for different values of shape parameter using different initial points of $\mathbf{\Sigma}$. $N = 10000$, $\sigma \in (0.4, 0.6)$.

for RA-FP to successfully converge. When $\beta = 1$ the MGGD reduces to the Gaussian distribution, where the ML estimator is the covariance matrix. Hence, only one iteration of RA-FP is needed. In addition, it can observed that the number of iterations depend on the value of β and decreases as β becomes large.

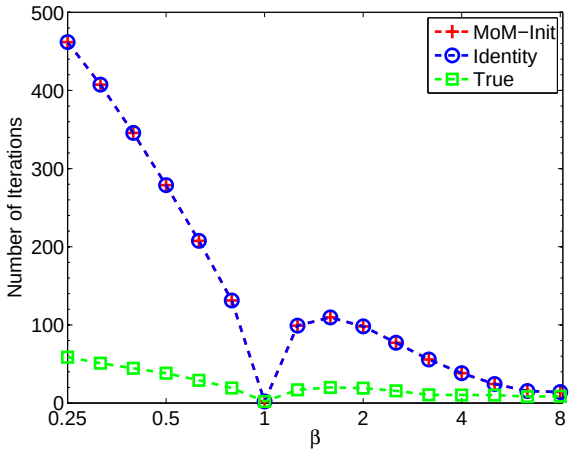

Figure 5: Number of iterations needed for RA-FP to converge as a function of β. $N = 10000$, $\sigma \in (0.4, 0.6)$.

2.1.7 Discussion

We present two new ML algorithms, ML-FS and RA-FP, which accurately estimate $\mathbf{\Sigma}$ for any positive value of β. ML-FS is a variation of Newton-Raphson algorithm where the inverse Hessian has been replased by the inverse of the Fisher information matrix. RA-FP is based on the implementation of successful Riemannian averages of the FP iterates, in order to prevent them from diverging from the true value. Numerical results show that for any value of shape parameter, the sequence of matrices produced by ML-FS or RA-FP converges to the true value. Moreover, numerical experiments show that for any and for any initial symmetric positive definite matrix RA-FP converges to the true value of the matrix $\mathbf{\Sigma}$.

2.2 Application to independent vector analysis

2.2.1 Introduction

Independent vector analysis (IVA) is a generalization of independent component analysis (ICA) that makes full use of the statistical dependence across multiple datasets to achieve source separation, and can take both second and higher-order statistics into account. MGGD provides an effective model for IVA as well as for modeling the latent multivariate variables–sources–and the

performance of the IVA algorithm highly depends on the estimation of the source parameters.

IVA achieves better performance than performing ICA separately on each dataset by exploiting the dependence across datasets. Among the solutions to IVA, IVA-Laplacian (IVA-L) [17, 18] does not exploit the linear dependencies expressed in the second-order statistics but only takes higher-order statistics into account and assumes a Laplacian distribution. Conversely, IVA-Gaussian (IVA-G) [19, 20] exploits linear dependencies but does not take higher-order statistics into account. Finally, IVA-GGD [21] is a more general IVA implementation where both second and higher order statistics are taken into account. IVA-GGD uses a fixed set of shape parameter values and only estimates the scatter matrix. In addition, IVA-GGD uses the MoM for the estimation of the scatter matrix, which does not possess the large sample optimality property. In our work, we use ML-FS and RA-FP as main estimation techniques to precisely estimate all the parameters of the MGGD sources simultaneously. Therefore, the corresponding IVA score and cost functions can be readily calculated, providing efficient performance for IVA.

In this section, we present a new IVA algorithm, IVA with adaptive MGGD (IVA-A-GGD), that estimates shape parameter and scatter matrix jointly and takes into account both second and higher-order statistics.

2.2.2 IVA

Let each dataset $\mathbf{x}^{[k]}$, $k = 1, ..., K$ be a linear mixture of N statistically independent sources

$$\mathbf{x}^{[k]} = \mathbf{A}^{[k]}\mathbf{s}^{[k]}, \quad k = 1, ..., K,$$

where $\mathbf{A}^{[k]} \in \mathbb{R}^{N\times N},\ k = 1, ..., K$ are invertible mixing matrices and $\mathbf{s}^{[k]} = [s_1^{[k]}, ..., s_N^{[k]}]^\top$ is the vector of latent sources for the kth dataset. The nth source component vector (SCV) $\mathbf{s}_n = [s_n^{[1]}, ..., s_n^{[k]}]^\top$, can be defined by concatenating the nth source from each of the K data sets. An illustration of SCV is shown in Figure 6.

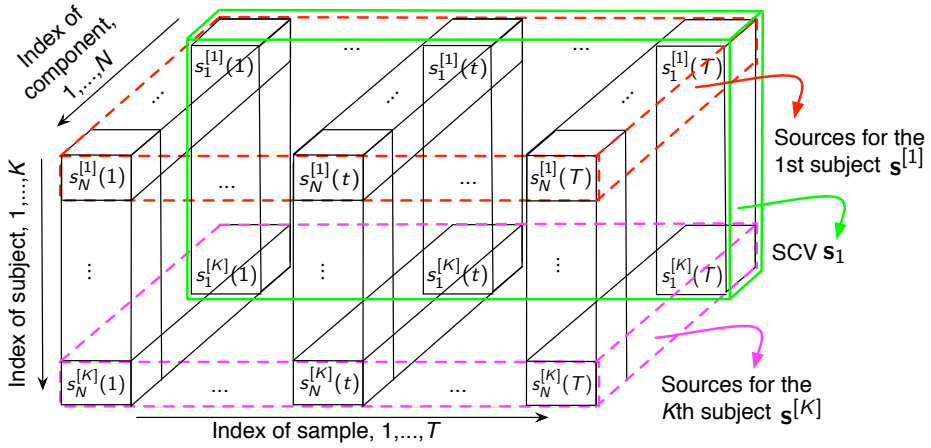

Figure 6: Illustration of SCV and source vectors.

The goal in IVA is to estimate K demixing matrices to yield source estimates $\mathbf{y}^{[k]} = \mathbf{W}^{[k]}\mathbf{x}^{[k]}$, such that each SCV is maximally independent of all other SCVs. This is achieved by minimizing the mutual information cost function

$$\mathcal{I}_{\text{IVA}} = \sum_{n=1}^{N} \mathcal{H}[\mathbf{y}_n] - \sum_{k=1}^{K} \log \left| \det \left(\mathbf{W}^{[k]} \right) \right| - C, \tag{25}$$

where $\mathcal{H}[\mathbf{y}_n]$ denotes the entropy of the nth SCV and C is the constant term $\mathcal{H}[\mathbf{x}^{[1]}, ..., \mathbf{x}^{[K]}]$. Different algorithms for minimizing the IVA cost function have been described in [19, 22]. The gradient of the cost function in (25) is given by

$$\frac{\partial \mathcal{I}_{\text{IVA}}}{\partial \mathbf{W}^{[k]}} = -\sum_{n=1}^{N} E\left\{ \frac{\partial \log p(\mathbf{y}_n)}{\partial y_n^{[k]}} \frac{\partial y_n^{[k]}}{\partial \mathbf{W}^{[k]}} \right\} - \left(\mathbf{W}^{[k]} \right)^{-\top}. \tag{26}$$

As observed in (26), the probability density function or its approximation for each estimated SCV plays an important role on the development of IVA algorithms.

2.2.3 IVA-A-GGD

Using the ML-FS and RA-FP algorithms as the main parameter estimation techniques, we propose a new IVA algorithm with adaptive MGGD prior to estimate the shape parameter and scatter matrix jointly and exploit both second and higher-order statistics. Instead of minimizing (25) with respect to $\mathbf{W}^{[\mathbf{k}]}$ we use a decoupling procedure [23, 24] to minimize (25) with respect to each of the row vectors $\mathbf{w}_i^{[k]}$, $i = 1, ..., N$. Therefore by following [21], we can rewrite the cost function as

$$\mathcal{I}_{\text{IVA}} = \mathcal{H}[\mathbf{y}_n] - \log \left| \left(\mathbf{h}_n^{[k]}\right)^{\top} \mathbf{w}_n^{[k]} \right| - C_n^{[k]},$$

where $\mathbf{h}_n^{[k]}$ is the unit length vector, with the property that is perpendicular to all row vectors of the matrix $\mathbf{W}^{[k]}$ except of the vector $\mathbf{w}_n^{[k]}$. The differential entropy for an MGGD $\mathbf{y}_n$ is given by

$$\mathcal{H}[\mathbf{y}_n] = -E\left\{\log p(\mathbf{y}_n)\right\} = -\log\left(\frac{\beta\Gamma\left(\frac{K}{2}\right)}{\pi^{\frac{K}{2}}\Gamma\left(\frac{K}{2\beta}\right)2^{\frac{K}{2\beta}}}\right) + \frac{K}{2}\log m + \frac{1}{2}\log|\mathbf{\Sigma}| + \frac{1}{2m^{\beta}}E\left\{\left(\mathbf{y}_n^{\top}\mathbf{\Sigma}^{-1}\mathbf{y}_n\right)^{\beta}\right\},$$

and the gradient update rule by

$$\frac{\partial \mathcal{I}_{\text{IVA}}}{\partial \mathbf{w}_n^{[k]}} = E\left\{\phi_n^{[k]}(\mathbf{y}_n)\mathbf{x}^{[k]}\right\} - \frac{\mathbf{h}_n^{[k]}}{\left(\mathbf{h}_n^{[k]}\right)^{\top}\mathbf{w}_n^{[k]}},$$

where $\phi_n^{[k]}$ is the kth element of the multivariate MGGD score function

$$\phi(\mathbf{y}_n) = \frac{\beta}{m^{\beta}}\left(\mathbf{y}_n^{\top}\mathbf{\Sigma}^{-1}\mathbf{y}_n\right)^{\beta-1}\mathbf{\Sigma}^{-1}\mathbf{y}_n.$$

2.2.4 Experimental results

To show the effectiveness of the IVA-A-GGD algorithm, we compare its performance with a number of widely used JBSS algorithms in terms of the joint inter-symbol-interference (ISI), defined in [19, 25]. In this set of expirements, we generate MGGD sources and consider two different cases

for the shape parameter β. For the first case we generate ten MGGD sourses $N = 10$, $K = 3$, and β, σ have been uniformly selected from the range $(0.25, 4)$ and $(0.4, 0.6)$ respectively. For the second case, we have $\beta \in (4, 8)$.

The performances are compared in terms of the intersymbol-interference (ISI). ISI [26, 27] is a widely used performance metric, which does not require an ordering of the sources in order to assess performance and is given by

$$\mathrm{ISI}(\mathbf{G}) = \frac{\sum_{n=1}^{N}\left(\sum_{m=1}^{N}\frac{|g_{n,m}|}{\max_p |g_{n,p}|} - 1\right) + \sum_{m=1}^{N}\left(\sum_{n=1}^{N}\frac{|g_{n,m}|}{\max_p |g_{p,m}|} - 1\right)}{2N(N-1)},$$

where $\mathbf{G} = \mathbf{WA}$ is the global demixing-mixing matrix. In this work, we use average and joint ISI [28, 29] to measure the performance of the IVA algorithms. The average ISI is given by

$$\mathrm{ISI}_{\mathrm{AVG}}(\mathbf{G}^{[1]}, \ldots, \mathbf{G}^{[K]}) = \frac{1}{K}\sum_{k=1}^{K}\mathrm{ISI}(\mathbf{G}^{[k]}),$$

where $\mathbf{G}^{[k]}$ for $k = 1, \ldots, K$ is the global demixing-mixing matrix for the kth data set. The joint ISI is given by

$$\mathrm{ISI}_{\mathrm{JNT}}(\mathbf{G}^{[1]}, \ldots, \mathbf{G}^{[K]}) = \mathrm{ISI}\left(\frac{1}{K}\sum_{k=1}^{K}(\mathbf{G}^{[k]})\right) = \mathrm{ISI}\left(\sum_{k=1}^{K}(\mathbf{G}^{[k]})\right).$$

We note that joint ISI takes the source alignment errors into account while average ISI does not.

From Fig.7 and Fig.8, we observe that IVA-A-GGD performs the best in terms of the joint ISI as function of sample sizes for any value of β, showing the effectiveness of IVA-A-GGD.

2.2.5 Discussion

Based on ML-FP and RA-FP, we introduce a new IVA algorithm, IVA-A-GGD that estimates shape parameter and scatter matrix jointly and exploits both second and higher-order statistics. Simulation results reveal the effectiveness of ML-FS algorithm as well as the desirable performance of

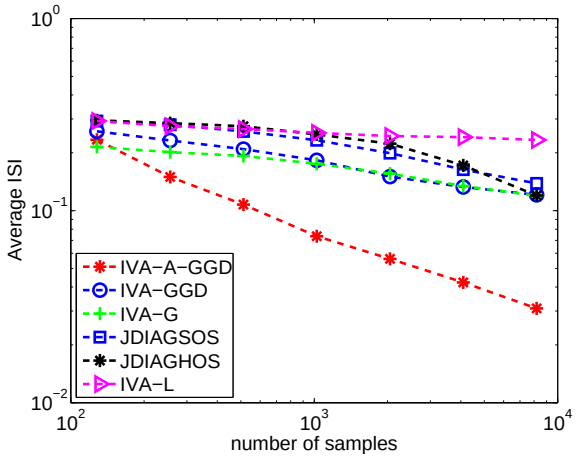

Figure 7: Performance of IVA-A-GGD for different number of sample size and $\beta \in (0.25, 4)$.

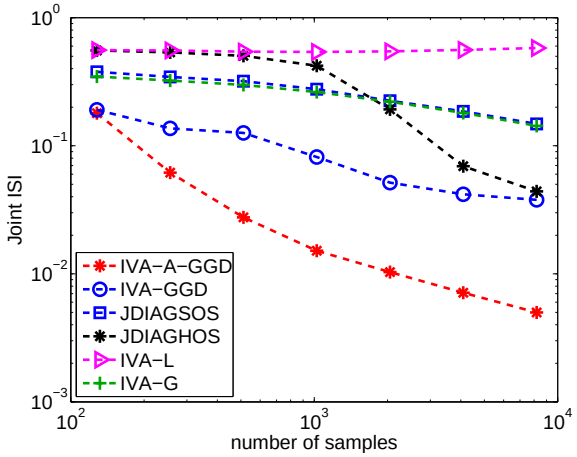

Figure 8: Performance of IVA-A-GGD for different number of sample size and $\beta \in (4, 8)$.

IVA-A-GGD when compared with existing competing algorithms.

3 Proposed work

3.1 Multivariate density estimation by entropy maximization with kernels

The estimation of a multivariate probability density function (PDF) is a common problem in a wide variety of fields ranging from physics to statistics. Within the field of machine learning, many estimation, detection, and classification problems require knowledge of the data's PDF either

explicitly or implicitly, see e.g., [30]. Thus, effective characterization of the density is vital to the success of these machine learning approaches.

There are a number of density estimation methods. The non-parametric multidimensional methods, such as histogram estimation, k-nearest neighbors (KNN), and kernel density estimation (KDE) [31, 32], can provide flexible density matching. However, these approaches are mostly computationally demanding, especially when sample size is large, and direct histogram estimation provides a discontinuous density function. In addition, the performance of non-parametric methods highly depend on the choice of parameters, such as the number of bins, number of samples in a neighborhood, and the bandwidth for histogram, KNN, and KDE, respectively. Many of the other methods assume a parametric model, such as the MGGD for the density. These parametric methods provide a simple form for the PDF and are computationally efficient, however the parametric form is usually very limited and provides satisfactory performance only when the assumed model is a good match to the true PDF. For example, the MGGD limits the PDFs to these that are symmetric and unimodal, and controls the shape through a single parameter. Semi-parametric methods combine the flexibility with the relatively simple density form. Multidimensiona Gaussian mixture model (MGMM) [32] has been widely used for semi-parametric density estimation. However, MGMM is typically time consuming and since the kernel function is limited to only a single type, the tradeoff between flexibility and generalization ability needs to be carefully weighted when selecting the number of mixtures and their parameters.

In this section, we propose a multivariate density estimator with smooth characteristics based on the maximum entropy principle. We propose to jointly use global and local measuring functions to provide flexible PDFs with as simple a form as possible. The global measuring functions provide constraints on the statistics of the PDF, such as the mean, variance, and higher-order statistics (HOS). However, methods using only global measuring functions [33–35] ignore local information, i.e., characteristics in a given interval, of the PDF, additionally the use of higher-order polynomials can introduce stability issues. We use Gaussian kernels, which do not have stability issues

due to their localized characteristics, as local measuring functions to provide local information. We call the new estimator multivariate entropy maximization with kernels (M-EMK) and consider it as an extension of the univariate case [36]

3.1.1 Maximum joint entropy densities

The maximum entropy density, subject to known constraints, can be written as the following optimization problem [37]:

$$\begin{aligned} \max_{p(\mathbf{x})} \mathcal{H}(\mathbf{x}) &= -\int_{\Omega(\mathbf{x})} p(\mathbf{x}) \log p(\mathbf{x})\, d\mathbf{x} \\ \text{s.t.} \int_{\Omega(\mathbf{x})} r_i(\mathbf{x}) p(\mathbf{x})\, d\mathbf{x} &= \alpha_i, \text{ for } i = 0, \ldots, M, \end{aligned} \tag{27}$$

where $p(\mathbf{x}) \geq 0$, with equality outside the support set $\Omega(\mathbf{x})$, $r_i(\mathbf{x})$ and $\alpha_i = \sum_{t=1}^{T} r_i(\mathbf{x}_t)/T$ for $i = 0, \ldots, M$ are the measuring functions and their corresponding sample averages, respectively. From now on, we simplify the notation for clarity if it can be inferred from context. From (27), p is a density on support set $\Omega(\mathbf{x})$ meeting $M + 1$ constraints $\int r_i p = \alpha_i$. We note that the first constraint must be $\int p = 1$, equivalently $r_0 = 1$ and $\alpha_0 = 1$, to ensure that p is a valid PDF. The optimization problem in (27) can be written as a Lagrangian function:

$$\mathcal{L}(p) = -\int p \log p + \sum_{i=0}^{M} \lambda_i \int (r_i - \alpha_i)\, p, \tag{28}$$

where λ_i, $i = 0, \ldots, M$, are the Lagrangian multipliers. Through the use of functional variation, we can "differentiate" (28) with respect to p. By setting $\partial \mathcal{L}(p)/\partial p = 0$, we obtain the equation of maximum entropy distribution,

$$p(\mathbf{x}) = \exp\left\{-1 + \sum_{i=0}^{M} \lambda_i r_i(\mathbf{x})\right\}, \tag{29}$$

where Lagrangian multipliers are chosen such that p satisfies the constraints in (27). By substituting (29) into the constraints in (27), we generate a nonlinear system of $M+1$ equations for the $M+1$ Lagrange multipliers. We can solve the problem by a Newton iteration method,

$$\boldsymbol{\lambda}^{(n+1)} = \boldsymbol{\lambda}^{(n)} - \mathbf{J}^{-1} E_{p^{(n)}} \{\mathbf{r} - \boldsymbol{\alpha}\}, \tag{30}$$

where $p^{(n)}$ is the estimated PDF for the nth iteration, and $\mathbf{r} = [r_0, \ldots, r_M]^\top$, $\boldsymbol{\lambda} = [\lambda_0, \ldots, \lambda_M]^\top \in \mathbb{R}^{M+1}$, and $\boldsymbol{\alpha} = [\alpha_0, \ldots, \alpha_M]^\top$ denote the vector of measuring functions, Lagrange multipliers, and sample averages, respectively. The ith entry of $E_{p^{(n)}} \{\mathbf{r} - \boldsymbol{\alpha}\}$ is given by

$$E_{p^{(n)}} \{r_i - \alpha_i\} = \int (r_i - \alpha_i)\, p^{(n)}, \tag{31}$$

and the (i, j)th entry of $\mathbf{J}$ is given by

$$\mathbf{J}_{ij} = \frac{\partial \int (r_i - \alpha_i)\, p^{(n)}}{\partial \lambda_j} = \int r_i r_j p^{(n)}. \tag{32}$$

The main challenges in the multivariate case that one can face are the following:

- Choice of multivariate measuring functions, to ensure that we provide a flexible density estimation while keeping the complexity low.
- Choice of the number of constraints.
- Multi dimensional integration.

3.2 IVA-MEMK Algorithm

Based on the multivariate density estimator M-EMK, we propose to develop an IVA algorithm, IVA by multivariate entropy maximization with kernels IVA-EMK that uses M-EMK to successfully match multivariate sources from a wide range of distributions.

Recall from previous section that the decoupled cost function can be written as

$$\mathcal{I}_{\text{IVA}} = \mathcal{H}[\mathbf{y}_n] - \log\left|\left(\mathbf{h}_n^{[k]}\right)^{\top}\mathbf{w}_n^{[k]}\right| - C_n^{[k]},$$

and its gradient by

$$\frac{\partial \mathcal{I}_{\text{IVA}}}{\partial \mathbf{w}_n^{[k]}} = E\left\{\phi_n^{[k]}(\mathbf{y}_n)\mathbf{x}^{[k]}\right\} - \frac{\mathbf{h}_n^{[k]}}{\left(\mathbf{h}_n^{[k]}\right)^{\top}\mathbf{w}_n^{[k]}}.$$

By using (29) as the maximum entropy pdf the kth element of the multivariate score function can be written as

$$\phi(\mathbf{y}_n^{[k]}) = -\frac{\partial \log(p_n(\mathbf{y}_n))}{\partial y_n^{[k]}} = -\sum_{i=0}^{M}\lambda_i\frac{\partial r_i(\mathbf{y}_n)}{\partial y_n^{[k]}}.$$

Therefore, the gradient update rule can be written as

$$\frac{\partial \mathcal{I}_{\text{IVA}}}{\partial \mathbf{w}_n^{[k]}} = -\sum_{i=0}^{M}\lambda_i\frac{\partial r_i(\mathbf{y}_n)}{\partial y_n^{[k]}}E\left\{\mathbf{x}^{[k]}\right\} - \frac{\mathbf{h}_n^{[k]}}{\left(\mathbf{h}_n^{[k]}\right)^{\top}\mathbf{w}_n^{[k]}},$$

and the differential entropy

$$\begin{aligned}\mathcal{H}[\mathbf{y}_n] &= -\int p\log p \\ &= -\int\left(-1+\lambda_0+\sum_{i=1}^{M}\lambda_i r_i\right)p \\ &= 1-\lambda_0-\sum_{i=1}^{M}\lambda_i\alpha_i. \end{aligned} \tag{33}$$

This approach is quite different from the MGGD-based approach described in the preliminary work, where a density model is chosen and the parameters are estimated during the adaption of the demixing matrix

3.3 Constrained IVA

In many cases, there is important prior information about the data that is available. For instance, they might be certain properties for the components and their mixing either as properties such as sparsity, smoothness, and nonnegativity, or certain structures and relationships among datasets. Including such prior information in the analysis could provide significant improvement to the final extracted features.

As a motivation for the introduction of prior information, one can think the example of analysis of data sets with group structure, such as multiple fMRI, or fMRI and electroencephalography (EEG) data sets. Within these sets, it is to be expected that multiple similar components may appear, all of these dependent on a set of components in each of the other data sets. This is the case, for example, when studying the expressions of tasks on a group of subjects, where natural subgroups can be formed based on age, gender, or other subject-specific parameters. The multiple data sets may be based on different tasks or different recording modalities. Under these conditions, it is highly probable that subjects within a subgroup respond similarly to a task and thus to find multiple similar expressions that differentiate between the subgroups. Hence, the expressions could no longer be regarded as independent components but should be regrouped so as to form independent subspaces.

The IVA formulation offers a number of unique advantages. It makes use of the powerful assumption of independence and it has a simple but well-defined structure that can be fully exploited by the powerful decoupling trick. Therefore, based on the well structured IVA framework we propose to develop a powerful framework where incorporation of prior information and independent subspaces can be achieved jointly. There are several ways to incorporate prior information, including definition of structured sparse density models, relational constraints, and traditional constrained optimization framework.

3.3.1 Incorporation of prior knowledge through pdfs

One way to incorporate sparsity into the IVA formulation is by selecting the density model to favor such distributions. Therefore, the distribution of the marginals is more peaky and has heavier tails. This can be achieved by selecting the shape parameter β to be within the range $(0, 1)$, or by selecting skewed measuring functions and local kernels. One advantage with this approach is that we still have access to the maximum likelihood (ML) solution, hence we can preserve all the theoretical and practical advantages associated with the ML framework.

3.3.2 Incorporation of prior knowledge through explicit constraints

Another way to incorporate prior information is to use a constrained optimization framework and to directly add the constraints through Lagrange multipliers. Constraints can be expressed either on the components or on the mixing vectors. However, prior information is transferrable from the components or mixing matrix parameters to the unmixing matrix parameters, allowing for integration into our objective function.

As an example let $\mathcal{I}(\mathbf{W})$ be the IVA objective function and let γ be a prior with respect to $\mathbf{a}_i$. To simplify the notation we have dropped all irrelevant indices. For our constraint to yield a unique feasible solution in the component space, we have

$$\gamma^\top \mathbf{a}_i \geq \epsilon > \gamma^\top \mathbf{a}_j, \ \forall i \neq j.$$

If $\mathbf{W}^\top$ is an estimator of $\mathbf{A}^{-1}$ and letting ξ be the reference formulated with respect to $\mathbf{w}_i$ then

$$\gamma^\top \mathbf{A} = \xi^\top \mathbf{W} = \xi^\top \mathbf{A}^{-\top},$$

which implies $\gamma = (\mathbf{A}\mathbf{A}^\top)^{-1}\xi$, yielding

$$\xi^\top (\mathbf{A}\mathbf{A}^\top)^{-1} \mathbf{a}_i \geq \epsilon > \xi^\top (\mathbf{A}\mathbf{A}^\top)^{-1} \mathbf{a}_j, \ \forall i \neq j.$$

Since the components within an observation are to be uncorrelated, under our model we have

$$E\{\mathbf{x}\mathbf{x}^\top\} = \mathbf{A}E\{\mathbf{x}\mathbf{x}^\top\}\mathbf{A}^\top = \mathbf{A}\mathbf{A}^\top,$$

so the constraint that we impose on the ith column of $\mathbf{W}$ is

$$\xi = E\{\mathbf{x}\mathbf{x}^\top\}\gamma.$$

Examples include prior information about the task in fMRI analysis that can be included as constraints on the mixing matrix columns.

3.3.3 Incorporation of prior knowledge through relational constraints

Within IVA, we can define multiple relational constraints rather than explicit constraints. Multiple relational constraints can be introduced when the goal is the estimation of joint features from multimodality data. We propose to include such relational constraintstypically through maximization of the correlation of a set of mixing matrix profiles within IVA, through direct inclusion with a constrained optimization framework. Since the IVA cost function fully makes use of the dependence across datasets, such an approach promises to provide features that satisfy the relational constraints while making use of the diversity across the datasets.

4 Conclusion

In our preliminary work, we introduce two effective ML techniques for MGGD by using the Fisher scoring and a fixed point algorithm. Based on these techniques, we introduce a new IVA algorithm, IVA-A-GGD that estimates shape parameter and scatter matrix jointly and exploits both second and higher-order statistics. For our future work, we propose to extend IVA in two major directions: the design of a more general multivariate probability density function estimator and the investigation of

how prior knowledge can be incorporated into the IVA model. The multivariate density estimation technique is based on the maximum entropy principle. Based on this estimation technique, we propose a new IVA algorithm that will successfully match multivariate sources from a wide range of distributions. Finally, we propose to enrich IVA through incorporation of prior information yielding new perspectives for its model usefulness.

References

[1] M. Z. Coban and R. Mersereau, "Adaptive subband video coding using bivariate generalized gaussian distribution model," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 4. IEEE, 1996, pp. 1990–1993.

[2] J. Yang, Y. Wang, W. Xu, and Q. Dai, "Image and video denoising using adaptive dual-tree discrete wavelet packets," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 19, no. 5, pp. 642–655, 2009.

[3] T. Elguebaly and N. Bouguila, "Bayesian learning of generalized gaussian mixture models on biomedical images," in *Artificial Neural Networks in Pattern Recognition*. Springer, 2010, pp. 207–218.

[4] G. Verdoolaege and P. Scheunders, "On the geometry of multivariate generalized gaussian models," *Journal of Mathematical Imaging and Vision*, vol. 43, no. 3, pp. 180–193, 2012.

[5] ——, "Geodesics on the manifold of multivariate generalized gaussian distributions with an application to multicomponent texture discrimination," *International Journal of Computer Vision*, vol. 95, no. 3, pp. 265–286, 2011.

[6] F. Pascal, L. Bombrun, J.-Y. Tourneret, and Y. Berthoumieu, "Parameter estimation for multivariate generalized gaussian distributions," *Signal Processing, IEEE Transactions on*, vol. 61, no. 23, pp. 5960–5971, Dec 2013.

[7] T. Zhang, A. Wiesel, and M. S. Greco, "Multivariate generalized gaussian distribution: Convexity and graphical models," *Signal Processing, IEEE Transactions on*, vol. 61, no. 16, pp. 4141–4148, 2013.

[8] E. Ollila, D. Tyler, V. Koivunen, and H. Poor, "Complex elliptically symmetric distributions: Survey, new results and applications," *Signal Processing, IEEE Transactions on*, vol. 60, no. 11, pp. 5597–5625, Nov 2012.

[9] L. Bombrun, F. Pascal, J.-Y. Tourneret, and Y. Berthoumieu, "Performance of the maximum likelihood estimators for the parameters of multivariate generalized gaussian distributions," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012, pp. 3525–3528.

[10] S. Sra and R. Hosseini, "Geometric optimisation on positive definite matrices for elliptically contoured distributions," in *Advances in Neural Information Processing Systems*, 2013, pp. 2562–2570.

[11] S. Kotz, *Multivariate distributions at a cross road*. Springer, 1975.

[12] A. Terras, *Harmonic analysis on symmetric spaces and applications, Vol. II*. New York: Springer-Verlag, 1988.

[13] R. Bhatia, *Positive definite matrices*, 2007.

[14] K. Sturm, "Probability measures on metric spaces of nonpositive curvature," *Contemporary mathematics*, vol. 338, pp. 357–390, 2003.

[15] D. Smart, *Fixed point theorems*. Cambridge University Press, 1974.

[16] E. Gómez, M. Gomez-Viilegas, and J. Marin, "A multivariate generalization of the power exponential family of distributions," *Communications in Statistics-Theory and Methods*, vol. 27, no. 3, pp. 589–600, 1998.

[17] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ica to multivariate components," in *Independent Component Analysis and Blind Signal Separation*. Springer, 2006, pp. 165–172.

[18] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 70–79, 2007.

[19] M. Anderson, T. Adalı, and X.-L. Li, "Joint blind source separation with multivariate gaussian model: Algorithms and performance analysis," *Signal Processing, IEEE Transactions on*, vol. 60, no. 4, pp. 1672–1683, April 2012.

[20] J. Via, M. Anderson, X.-L. Li, and T. Adalı, "A maximum likelihood approach for independent vector analysis of gaussian data sets," in *Machine Learning for Signal Processing (MLSP), 2011 IEEE International Workshop on*, Sept 2011, pp. 1–6.

[21] M. Anderson, G.-S. Fu, R. Phlypo, and T. Adalı, "Independent vector analysis, the kotz distribution, and performance bounds," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 3243–3247.

[22] T. Adalı, M. Anderson, and G.-S. Fu, "Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging," *Signal Processing Magazine, IEEE*, vol. 31, no. 3, pp. 18–33, May 2014.

[23] X.-L. Li and X.-D. Zhang, "Nonorthogonal joint diagonalization free of degenerate solution," *Signal Processing, IEEE Transactions on*, vol. 55, no. 5, pp. 1803–1814, 2007.

[24] M. Anderson, X.-L. Li, P. Rodriguez, and T. Adali, "An effective decoupling method for matrix optimization and its application to the ica problem," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1885–1888.

[25] Y.-O. Li, T. Adalı, W. Wang, and V. D. Calhoun, "Joint blind source separation by multiset canonical correlation analysis," *Signal Processing, IEEE Transactions on*, vol. 57, no. 10, pp. 3918–3929, 2009.

[26] S. Choi, A. Cichocki, L. Zhang, and S. Amari, "Approximate maximum likelihood source separation using the natural gradient," in *Wireless Communications, 2001. (SPAWC '01). 2001 IEEE Third Workshop on Signal Processing Advances in*, 2001, pp. 235 –238.

[27] E. Moreau and O. Macchi, "A one stage self-adaptive algorithm for source separation," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, vol. iii, Apr 1994, pp. III/49–III/52 vol.3.

[28] Y.-O. Li, T. Adalı, W. Wang, and V. Calhoun, "Joint blind source separation by multiset canonical correlation analysis," *Signal Processing, IEEE Transactions on*, vol. 57, no. 10, pp. 3918–3929, Oct 2009.

[29] M. Anderson, T. Adalı, and X.-L. Li, "Joint blind source separation with multivariate Gaussian model: Algorithms and performance analysis," *Signal Processing, IEEE Transactions on*, vol. 60, no. 4, pp. 1672–1683, Apr. 2012.

[30] T. Adalı, M. Anderson, and G.-S. Fu, "Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging," *Signal Processing Magazine, IEEE*, vol. 31, no. 3, pp. 18–33, May 2014.

[31] A. J. Izenman, "Review papers: Recent developments in nonparametric density estimation," *Journal of the American Statistical Association*, vol. 86, no. 413, pp. 205–224, 1991.

[32] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[33] X.-L. Li and T. Adalı, "Independent component analysis by entropy bound minimization," *Signal Processing, IEEE Transactions on*, vol. 58, no. 10, pp. 5151–5164, Oct. 2010.

[34] B. Behmardi, R. Raich, and A. Hero, "Entropy estimation using the principle of maximum entropy," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 2008–2011.

[35] R. V. Abramov, "An improved algorithm for the multidimensional moment-constrained maximum entropy problem," *Journal of Computational Physics*, vol. 226, no. 1, pp. 621–644, 2007.

[36] G.-S. Fu, Z. Boukouvalas, and T. Adalı, "Density estimation by entropy maximization with kernels," *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2015.

[37] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York, NY, USA: Wiley-Interscience, 1991.