Joint Independent Subspace Analysis: A Quasi-Newton Algorithm

Dana Lahat, Christian Jutten

Joint Independent Subspace Analysis: A Quasi-Newton Algorithm

Dana Lahat and Christian Jutten

GIPSA-Lab, UMR CNRS 5216, Grenoble Campus, BP46,
38402 Saint-Martin-d'Hères, France*

Abstract. In this paper, we present a quasi-Newton (QN) algorithm for joint independent subspace analysis (JISA). JISA is a recently proposed generalization of independent vector analysis (IVA). JISA extends classical blind source separation (BSS) to jointly resolve several BSS problems by exploiting statistical dependence between latent sources across mixtures, as well as relaxing the assumption of statistical independence within each mixture. Algebraically, JISA based on second-order statistics amounts to coupled block diagonalization of a set of covariance and cross-covariance matrices, as well as block diagonalization of a single permuted covariance matrix. The proposed QN algorithm achieves asymptotically the minimal mean square error (MMSE) in the separation of multidimensional Gaussian components. Numerical experiments demonstrate convergence and source separation properties of the proposed algorithm.

Keywords: Blind source separation, independent vector analysis, independent subspace analysis, joint block diagonalization

1 Introduction

In this paper, we present a new algorithm for joint independent subspace analysis (JISA) [1]. JISA is a blind source separation (BSS) framework inspired by two recently-proposed extensions to BSS that until recently have been dealt with only separately: (1) relaxing the constraint that latent sources within a set of measurements must be statistically independent, sometimes termed independent subspace analysis (ISA) [2–4], and (2) solving several classical BSS problems simultaneously by exploiting statistical dependencies between latent sources across sets of measurements, a model often known as independent vector analysis (IVA) [5]. JISA provides a new flexible way to exploit links between different datasets, and thus has the potential to be useful to data fusion. The JISA model, and a relative gradient (RG) algorithm that achieves optimal separation in terms of minimal mean square error (MMSE) in the presence of noiseless Gaussian data, were first presented in [1]. A gradient algorithm that performs JISA based on the multivariate Laplace distribution has recently been

* This work is supported by the project CHESS, 2012-ERC-AdG-320684. GIPSA-Lab is a partner of the LabEx PERSYVAL-Lab (ANR–11-LABX-0025).

proposed [6]. The main contribution of this paper is a Newton-based algorithm that achieves optimal separation for Gaussian noise-free data and converges in a much smaller number of iterations than its RG counterpart [1].

Consider T observations of K vectors $\mathbf{x}^{[k]}(t)$, modelled as

$$\mathbf{x}^{[k]}(t) = \mathbf{A}^{[k]}\mathbf{s}^{[k]}(t) \quad 1 \le t \le T\,,\, 1 \le k \le K\,, \tag{1}$$

where $\mathbf{A}^{[k]}$ are $M \times M$ invertible matrices that may be different $\forall k$, and $\mathbf{x}^{[k]}(t)$ and $\mathbf{s}^{[k]}(t)$ are $M \times 1$ vectors. Given the partition $\mathbf{s}^{[k]}(t) = [\mathbf{s}_1^{[k]\mathsf{T}}(t), \ldots, \mathbf{s}_N^{[k]\mathsf{T}}(t)]^\mathsf{T}$, where $\mathbf{s}_i^{[k]}(t)$ are $m_i \times 1$ vectors, $m_i \ge 1$, $\sum_{i=1}^N m_i = M$, $N \le M$, $\cdot^\mathsf{T}$ denotes transpose, and the probability density function (pdf) of each random process $\mathbf{s}_i^{[k]}(t)$ irreducible in the sense that it cannot be factorized into a product of non-trivial pdfs, then each mixture (1) represents a single ISA [2–4] problem. The model that we define as JISA corresponds to linking several such standalone ISA problems via the assumption that the elements of the $n_i \times 1$ vector $\mathbf{s}_i(t) = [\mathbf{s}_i^{[1]\mathsf{T}}(t), \ldots, \mathbf{s}_i^{[K]\mathsf{T}}(t)]^\mathsf{T}$, where $n_i = Km_i$, are statistically *dependent*, whereas the pairs $(\mathbf{s}_i(t), \mathbf{s}_j(t))$ are statistically *independent* $\forall i \ne j \in \{1, \ldots, N\}$. JISA can be regarded as generalizing IVA since in IVA, $m_i = 1$ $\forall i$ (which implies $N = M$).

In the rest of this paper, we focus on JISA using second-order statistics (SOS). In this case, further insights can be obtained by rewriting (1) as

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{2}$$

where $\mathbf{s}(t) = [\mathbf{s}^{[1]\mathsf{T}}(t), \ldots, \mathbf{s}^{[K]\mathsf{T}}(t)]^\mathsf{T}$ and $\mathbf{x}(t) = [\mathbf{x}^{[1]\mathsf{T}}(t), \ldots, \mathbf{x}^{[K]\mathsf{T}}(t)]^\mathsf{T}$ are $L \times 1$ vectors, $L = KM$, $\mathbf{A} = \oplus_{k=1}^K \mathbf{A}^{[k]} \in \mathcal{B}_\mathbf{k}$ is a matrix direct sum, $\mathbf{k} = M\mathbf{1}_K$ the block-pattern of $\mathbf{A}$, and $\mathcal{B}_\mathbf{b}$ denotes the subspace of all invertible block-diagonal matrices with block-pattern $\mathbf{b}$. With these notations, $\widetilde{\mathbf{s}}(t) = \mathbf{\Phi}\mathbf{s}(t)$, where $\widetilde{\mathbf{s}}(t) = [\mathbf{s}_1^\mathsf{T}(t), \ldots, \mathbf{s}_N^\mathsf{T}(t)]^\mathsf{T}$ and $\mathbf{\Phi}$ is the corresponding $L \times L$ permutation matrix. Assuming sample independence $\forall t \ne t'$, the model assumptions imply that $\widetilde{\mathbf{S}} \triangleq E\{\widetilde{\mathbf{s}}(t)\widetilde{\mathbf{s}}^\mathsf{T}(t)\} = \begin{bmatrix} \mathbf{S}_{11} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{S}_{NN} \end{bmatrix} = \oplus_{i=1}^N \mathbf{S}_{ii} \in \mathcal{B}_\mathbf{n}$, where $\widetilde{\mathbf{S}}$ is an $L \times L$ block-diagonal matrix with block-pattern $\mathbf{n} = [n_1, \ldots, n_N]^\mathsf{T}$ and $\widetilde{\mathbf{S}} = \mathbf{\Phi}\mathbf{S}\mathbf{\Phi}^\mathsf{T} \in \mathcal{B}_\mathbf{n}$. The linear model (2) implies that $\mathbf{X} = \mathbf{A}\mathbf{S}\mathbf{A}^\mathsf{T}$ where $\mathbf{S} = E\{\mathbf{s}(t)\mathbf{s}^\mathsf{T}(t)\}$ and $\mathbf{X} = E\{\mathbf{x}(t)\mathbf{x}^\mathsf{T}(t)\}$. In the sequel, we assume that all $\mathbf{S}_{ii}$ are invertible and do not contain values fixed to zero. Typical structures of some of these matrices are depicted in Fig. 1.

Figure of merit: The above partition of $\mathbf{s}^{[k]}(t)$ induces a corresponding partition in the mixing matrices: $\mathbf{A}^{[k]} = [\mathbf{A}_1^{[k]} | \cdots | \mathbf{A}_N^{[k]}]$ with $\mathbf{A}_i^{[k]}$ the ith $M \times m_i$ column-block of $\mathbf{A}^{[k]}$. The multiplicative model (1) may now be rewritten as a sum of $N \le M$ *multidimensional components*: $\mathbf{x}^{[k]}(t) = \sum_{i=1}^N \mathbf{x}_i^{[k]}(t)$ where the ith $M \times 1$ component $\mathbf{x}_i^{[k]}(t)$ is defined as $\mathbf{x}_i^{[k]}(t) = \mathbf{A}_i^{[k]}\mathbf{s}_i^{[k]}(t)$. In a blind context, the component vector $\mathbf{x}_i^{[k]}(t)$ is better defined than the source vector $\mathbf{s}_i^{[k]}(t)$. Indeed, for any invertible $m_i \times m_i$ matrix $\mathbf{Z}_{ii}^{[k]}$, it is impossible to discriminate between the representation of a component $\mathbf{x}_i^{[k]}(t)$ by the pair $(\mathbf{A}_i^{[k]}, \mathbf{s}_i^{[k]}(t))$ and $(\mathbf{A}_i^{[k]}\mathbf{Z}_{ii}^{-[k]}, \mathbf{Z}_{ii}^{[k]}\mathbf{s}_i^{[k]}(t))$, where $\mathbf{Z}_{ii}^{-[k]}$ denotes $(\mathbf{Z}_{ii}^{[k]})^{-1}$. This means that only the

column space of $\mathbf{A}_i^{[k]}$, $\mathrm{span}(\mathbf{A}_i^{[k]})$, can be blindly identified. Therefore, JISA is in fact a (joint) subspace estimation problem. Given $\mathbf{m} = [m_1, \ldots, m_N]$ and the set of observations $\mathcal{X} = \{\mathbf{x}^{[k]}(t)\}_{k=1,t=1}^{K,T}$, the problem that we define as JISA can thus be stated as estimating $\mathcal{A} = \{\mathbf{A}^{[k]}\}_{k=1}^{K}$ such that the components $\mathbf{x}_1(t), \ldots, \mathbf{x}_N(t)$, where $\mathbf{x}_i(t) = [\mathbf{x}_i^{[1]\mathsf{T}}(t), \ldots, \mathbf{x}_i^{[K]\mathsf{T}}(t)]^{\mathsf{T}}$, are as independent as possible. In the sequel, we set up a simple statistical model that, via its likelihood function, yields a quantitative measure of independence. Accordingly, we define the figure of merit as the mean square error (MSE) in the estimation of $\mathbf{x}_i(t)$,

$$\widehat{\mathrm{MSE}}_i = \frac{1}{\sigma_i^2}\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathbf{x}}_i(t) - \mathbf{x}_i(t)\|^2\,, \tag{3}$$

where $\sigma_i^2 = E\{\|\mathbf{x}_i(t)\|^2\}$. For Gaussian data, estimates of $\mathbf{x}_i(t)$ obtained from matrices that are maximum likelihood (ML) estimates of $\mathcal{A}$ achieve asymptotically (i.e., $T \to \infty$) the MMSE [7].

Likelihood and contrast function: In the following, we consider a Gaussian model in which $\mathbf{s}_i(t) \sim \mathcal{N}(\mathbf{0}_{n_i\times 1}, \mathbf{S}_{ii})$ are mutually independent samples $\forall t \neq t'$. The log-likelihood for the model just described is [1] $\log p(\mathcal{X}; \mathcal{A}, \mathbf{S}) = -TD(\mathbf{\Phi}\mathbf{A}^{-1}\overline{\mathbf{X}}\mathbf{A}^{-\mathsf{T}}\mathbf{\Phi}^{\mathsf{T}}, \widetilde{\mathbf{S}}) - \kappa$, where $\mathcal{A} = \{\mathbf{A}^{[k]}\}_{k=1}^{K}$, and $\overline{\mathbf{X}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}(t)\mathbf{x}^{\mathsf{T}}(t)$ is the empirical counterpart of $\mathbf{X}$. The term $\kappa = \frac{T}{2}(\log\det(2\pi\overline{\mathbf{X}}) + L)$ is irrelevant to the maximization of the likelihood since it depends only on the data and not on the parameters. The scalar $D(\mathbf{R}_1, \mathbf{R}_2) = \frac{1}{2}(\mathrm{tr}\{\mathbf{R}_1\mathbf{R}_2^{-1}\} - \log\det(\mathbf{R}_1\mathbf{R}_2^{-1}) - M)$, defined for any two $M \times M$ symmetric positive-definite matrices $\mathbf{R}_1$ and $\mathbf{R}_2$, is the Kullback-Leibler divergence (KLD) between the distributions $\mathcal{N}(\mathbf{0}, \mathbf{R}_1)$ and $\mathcal{N}(\mathbf{0}, \mathbf{R}_2)$ [8]. Given the block-diagonal structure of $\widetilde{\mathbf{S}}$, its ML estimate is [1] $\widehat{\widetilde{\mathbf{S}}}^{\mathrm{ML}} = \mathrm{bdiag}_{\mathbf{n}}\{\mathbf{\Phi}\mathbf{A}^{-1}\overline{\mathbf{X}}\mathbf{A}^{-\mathsf{T}}\mathbf{\Phi}^{\mathsf{T}}\}$. We can now write $\max_{\mathbf{S}}\log p(\mathcal{X}; \mathcal{A}, \mathbf{S}) = -TC(\mathbf{A}) + \kappa$, where in the latter we have defined the *contrast function*

$$C(\mathbf{A}) = D(\mathbf{\Phi}\mathbf{A}^{-1}\overline{\mathbf{X}}\mathbf{A}^{-\mathsf{T}}\mathbf{\Phi}^{\mathsf{T}}, \mathrm{bdiag}_{\mathbf{n}}\{\mathbf{\Phi}\mathbf{A}^{-1}\overline{\mathbf{X}}\mathbf{A}^{-\mathsf{T}}\mathbf{\Phi}^{\mathsf{T}}\})\,. \tag{4}$$

It holds that $D(\mathbf{R}, \mathrm{bdiag}_{\mathbf{b}}\{\mathbf{R}\}) \geq 0$ with equality if and only if (iff) $\mathbf{R} \in \mathcal{B}_{\mathbf{b}}$. Hence, for any positive-definite matrix $\mathbf{R}$, $D(\mathbf{R}, \mathrm{bdiag}_{\mathbf{b}}\{\mathbf{R}\})$ is a measure of the block-diagonality of $\mathbf{R}$. Therefore, minimizing[1] the contrast function (4) amounts to *(approximate) block diagonalization* of $\overline{\mathbf{X}}$ by a permuted block-diagonal matrix $\mathbf{\Phi}\mathbf{A}^{-1}$. The RG of (4), $\nabla C(\mathbf{A}) = \mathrm{bdiag}_{\mathbf{k}}\{\mathbf{\Phi}^{\mathsf{T}}\,\mathrm{bdiag}_{\mathbf{n}}^{-1}\{\mathbf{\Phi}\mathbf{A}^{-1}\overline{\mathbf{X}}\mathbf{A}^{-\mathsf{T}}\mathbf{\Phi}^{\mathsf{T}}\}\mathbf{\Phi}\mathbf{A}^{-1}\overline{\mathbf{X}}\mathbf{A}^{-\mathsf{T}}\} - \mathbf{I}$, where $\mathrm{bdiag}_{\mathbf{m}}^{-1}\{\cdot\}$ stands for $(\mathrm{bdiag}_{\mathbf{m}}\{\cdot\})^{-1}$, was derived in [1]. Matrices that satisfy $\nabla C(\mathbf{A}) = \mathbf{0}$ are ML estimates of $\mathcal{A}$. This is the basis for the RG algorithm [1].

2 Derivation of the Approximate Hessian

The derivation of the Hessian is based on a relative perturbation of each $\mathbf{A}^{[k]}$ by $\widehat{\mathbf{A}}^{[k]} = \mathbf{A}^{[k]}(\mathbf{I}_M + \boldsymbol{\mathcal{E}}^{[k]})^{-1}\boldsymbol{\Lambda}^{[k]}$, where the $M \times M$ matrix $\boldsymbol{\mathcal{E}}^{[k]}$ reflects the *relative*

[1] We assume that an optimum exists.

change in $\mathbf{A}^{[k]}$, up to a scale ambiguity that is represented by the arbitrary invertible matrix $\boldsymbol{\Lambda}^{[k]} \in \mathcal{B}_{\mathbf{m}}$. It can be shown [7] that the MSE (3) is invariant to $\boldsymbol{\Lambda}^{[k]}$. The first-order expansion of the equations that satisfy $\nabla C(\mathbf{A}) = \mathbf{0}$ can be rewritten [7], for each pair $i \neq j$, as $\mathbf{e} = -\boldsymbol{\mathcal{H}}^{-1}\mathbf{g} + \Omega(\frac{1}{T})$, where $\mathbf{e}$ and $\mathbf{g}$ are $2Km_im_j \times 1$ vectors, $\mathbf{e} = [\mathbf{e}_{ij}^\intercal \ \mathbf{e}_{ji}^\intercal]^\intercal$, $\mathbf{e}_{ij} = [\mathrm{vec}^\intercal\{\boldsymbol{\mathcal{E}}_{ij}^{[1]}\} \ \cdots \ \mathrm{vec}^\intercal\{\boldsymbol{\mathcal{E}}_{ij}^{[K]}\}]^\intercal$, $\mathbf{g} = [\mathbf{g}_{ij}^\intercal \ \mathbf{g}_{ji}^\intercal]^\intercal$, $\mathbf{g}_{ij} = [\mathrm{vec}^\intercal\{[\mathbf{S}_{ii}^{-1}\overline{\mathbf{S}}_{ij}]_{11}\} \ \cdots \ \mathrm{vec}^\intercal\{[\mathbf{S}_{ii}^{-1}\overline{\mathbf{S}}_{ij}]_{KK}\}]^\intercal$, $\boldsymbol{\mathcal{H}} = \begin{bmatrix} \mathbf{S}_{jj} \odot \mathbf{S}_{ii}^{-1} & \mathbf{I}_K \otimes \boldsymbol{\mathcal{T}}_{m_j,m_i} \\ \mathbf{I}_K \otimes \boldsymbol{\mathcal{T}}_{m_i,m_j} & \mathbf{S}_{ii} \odot \mathbf{S}_{jj}^{-1} \end{bmatrix}$ is a $2Km_im_j \times 2Km_im_j$ matrix that we assume invertible, $\otimes$ is the Kronecker product, and $\mathbf{S}_{jj} \odot \mathbf{S}_{ii}^{-1} = \begin{bmatrix} \mathbf{S}_{jj}^{[1,1]} \otimes [\mathbf{S}_{ii}^{-1}]_{11} & \cdots & \mathbf{S}_{jj}^{[1,K]} \otimes [\mathbf{S}_{ii}^{-1}]_{11} \\ \vdots & & \vdots \\ \mathbf{S}_{jj}^{[K,1]} \otimes [\mathbf{S}_{ii}^{-1}]_{K1} & \cdots & \mathbf{S}_{jj}^{[K,K]} \otimes [\mathbf{S}_{ii}^{-1}]_{KK} \end{bmatrix}$ is a $Km_im_j \times Km_im_j$ matrix whose (k,l)th block according to the partition $m_im_j\mathbf{1}_K$ is $\mathbf{S}_{jj}^{[k,l]} \otimes [\mathbf{S}_{ii}^{-1}]_{kl}$. The commutation matrix $\boldsymbol{\mathcal{T}}_{P,Q} \in \mathbb{R}^{PQ \times PQ}$ is such that $\mathrm{vec}\{\mathbf{M}^\intercal\} = \boldsymbol{\mathcal{T}}_{P,Q}\mathrm{vec}\{\mathbf{M}\}$ for any $\mathbf{M} \in \mathbb{R}^{P \times Q}$. $\Omega(f)$ stands for zero-mean stochastic terms whose standard deviation is proportional to f, or to higher powers thereof.

3 Algorithm

The approximation of the Hessian gives rise to a quasi-Newton (QN) algorithm, which is given in pseudocode in Algorithm 1. In line 5 of Algorithm 1 we

Algorithm 1 An iterative Newton-based algorithm for JISA

1: **function** JISA($\overline{\mathbf{X}}$, $\boldsymbol{\Phi}$, $\mathbf{A}_{\text{init}}$, $\mathbf{m}$, threshold, K)
2: $\quad\mathbf{A} \leftarrow \mathbf{A}_{\text{init}}$, $\mathbf{R} \leftarrow \overline{\mathbf{X}}$ ▷ initialization
3: $\quad$**while** $\|\nabla C\| >$ threshold **do**
4: $\quad\quad$**for** i=2:N, j=1:i-1 **do** ▷ Sweep on $(i, j \neq i)$
5: $\quad\quad\quad\mathbf{g} \leftarrow \begin{bmatrix} \mathrm{vecbd}_{m_i\mathbf{1}_K \times m_j\mathbf{1}_K}\{\mathbf{R}_{ii}^{-1}\mathbf{R}_{ij}\} \\ \mathrm{vecbd}_{m_j\mathbf{1}_K \times m_i\mathbf{1}_K}\{\mathbf{R}_{jj}^{-1}\mathbf{R}_{ji}\} \end{bmatrix}$
6: $\quad\quad\quad\boldsymbol{\mathcal{H}} \leftarrow \begin{bmatrix} \mathbf{R}_{jj} \odot \mathbf{R}_{ii}^{-1} & \mathbf{I}_K \otimes \boldsymbol{\mathcal{T}}_{m_j,m_i} \\ \mathbf{I}_K \otimes \boldsymbol{\mathcal{T}}_{m_i,m_j} & \mathbf{R}_{ii} \odot \mathbf{R}_{jj}^{-1} \end{bmatrix}$
7: $\quad\quad\quad$Evaluate $\boldsymbol{\mathcal{E}}_{ij}^{[k]}, \boldsymbol{\mathcal{E}}_{ji}^{[k]}$, k=1,...,K
8: $\quad\quad\quad$Set $\{\boldsymbol{\mathcal{E}}_{ij}^{[k]}, \boldsymbol{\mathcal{E}}_{ji}^{[k]}\}_{k=1}^{K}$ in $\boldsymbol{\mathcal{E}} = \oplus_{k=1}^{K}\boldsymbol{\mathcal{E}}^{[k]}$
9: $\quad\quad\quad\mathbf{T} \leftarrow \mathbf{I} + \boldsymbol{\mathcal{E}}$ ▷ $\boldsymbol{\mathcal{E}}_{ii}^{[k]} = 0$
10: $\quad\quad\quad\mathbf{R} \leftarrow \mathbf{T}^{-1}\mathbf{R}\mathbf{T}^{-\intercal}$
11: $\quad\quad\quad\mathbf{A} \leftarrow \mathbf{A}\mathbf{T}$ ▷ For output only
12: $\quad\quad$**end for**
13: $\quad\quad\nabla C \leftarrow \mathrm{bdiag}_{\mathbf{k}}\{\boldsymbol{\Phi}^\intercal\,\mathrm{bdiag}_{\mathbf{n}}^{-1}\{\boldsymbol{\Phi}\mathbf{A}^{-1}\mathbf{R}\mathbf{A}^{-\intercal}\boldsymbol{\Phi}^\intercal\}\boldsymbol{\Phi}\mathbf{A}^{-1}\mathbf{R}\mathbf{A}^{-\intercal}\boldsymbol{\Phi}^\intercal\boldsymbol{\Phi}\} - \mathbf{I}$
14: $\quad$**end while**
15: $\quad$**return A**
16: **end function**

introduce the operator $\mathrm{vecbd}_{\boldsymbol{\alpha}\times\boldsymbol{\beta}}\{\mathbf{X}\} \triangleq \left[\mathrm{vec}^\intercal\{\mathbf{X}_{11}\} \ldots \mathrm{vec}^\intercal\{\mathbf{X}_{KK}\}\right]^\intercal$,where

$\mathrm{vecbd}_{\boldsymbol{\alpha}\times\boldsymbol{\beta}}\{\mathbf{X}\}$ is a vector that consists only of the (vectorized) entries of the main-diagonal blocks of matrix $\mathbf{X}$. Matrices $\mathbf{X}_{kk}$ are the blocks on the main diagonal of $\mathbf{X}$ where the rows of $\mathbf{X}$ are partitioned according to $\boldsymbol{\alpha}$ and the columns by $\boldsymbol{\beta}$. The difference from the RG algorithm is in lines 5–8, see [1, Algorithm 2].

4 Numerical Validation

In this section, we explore some numerical properties of the QN algorithm and validate its proper functioning. In all the following numerical experiments, the real positive definite matrices $\mathbf{S}_{ii}$ are generated as $\mathbf{S}_{ii} = \mathrm{diag}^{-\frac{1}{2}}\{\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\intercal\}\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\intercal\,\mathrm{diag}^{-\frac{1}{2}}\{\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\intercal\}$, where $\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\intercal$ is the singular value decomposition (SVD) of a $Km_i \times Km_i$ matrix whose independent and identically distributed (i.i.d.) entries $\sim\mathcal{N}(0,1)$. The corresponding samples are created by right-multiplying the transpose of the Cholesky factorization of $\widetilde{\mathbf{S}}_{ii}$ with $Km_i\times T$ i.i.d. $\sim\mathcal{N}(0,1)$ numbers. $\mathbf{A}^{[k]}$ is arbitrary and thus, for simplicity, fixed to $\mathbf{I}$. In order to allow varying degrees of initialization, $\mathbf{A}^{[k]}_{\mathrm{init}} = p\boldsymbol{\Upsilon} + (1-p)\mathbf{I}$, $0\leq p\leq 1$, where $p=1$ implies fully random initialization. The entries of $\boldsymbol{\Upsilon}$ are $\sim\mathcal{N}(0,1)$ i.i.d. and drawn anew for each new $\mathbf{A}^{[k]}_{\mathrm{init}}$. The stopping threshold is set to 10^{-6}, and $T=10^4$. In order to emphasize the difference between JISA and IVA, at each Monte Carlo (MC) trial, the algorithm is run twice on the same data, in two modes: in the first mode, the input parameter $\mathbf{m}$ (Algorithm 1 line 1) reflects the true data properties. In the second mode, the input $\mathbf{m}$ is set to $\mathbf{1}_{M\times 1}$. The latter amounts to assuming that each $Km_i \times Km_i$ (irreducible, by definition) block on the diagonal of $\widetilde{\mathbf{S}}$ *can* be further reduced into m_i smaller blocks of dimension $K\times K$, i.e., ignoring the true subspace structure of the data. We denote this approach "mismodeling" [9].

4.1 Sensitivity to Initizalization and Number of Iterations

In order to focus here on the initialization, we avoid finite sample size errors by using data that can be exactly block-diagonalized. Fig. 1 depicts typical such data, as well as outputs of QN, in mismodeling and correct model scenarios, for two types of initialization: mild ($p=0.2$, Fig. 1(d)–1(e)), and fully random ($p=1$, Fig. 1(f)–1(i)). A key observation is that JISA is sensitive to initialization: compare Fig. 1(e) with Fig. 1(g). In the latter, a fully random initialization results in an inability to recover the block structure. On the other hand, due to the convexity of the mismodeled algorithm (see [10]), it minimizes properly the *mismodeled* contrast function $\forall p$. However, minimizing the mismodeled contrast function does not imply separation: there is still need to cluster the blocks, as shows Fig. 1(h). In Fig. 1(i), we cluster by simple enumeration on all $M!$ possibilities. This is definitely not a viable approach. Further discussion of this topic is beyond the scope of this paper.

We now compare the QN and RG algorithms in terms of number of iterations. Both RG and QN minimize the same contrast function (4) and thus achieve the

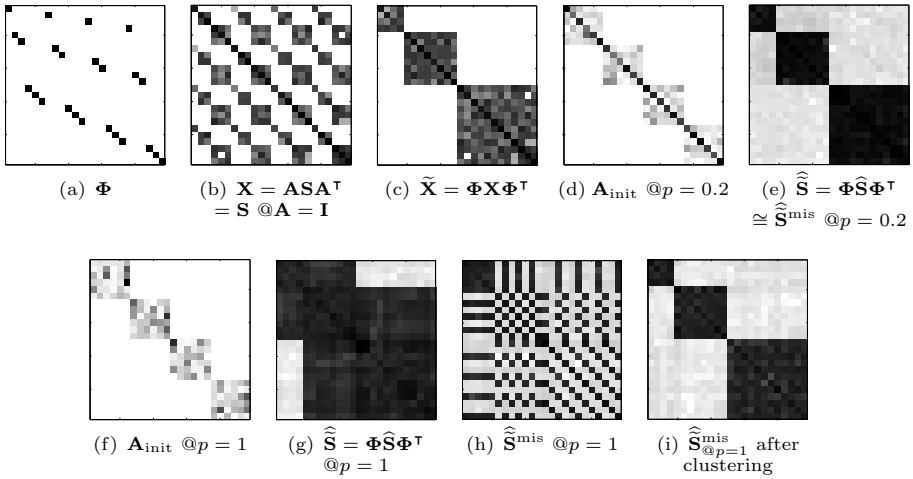

Fig. 1. Typical matrices and output of the QN algorithm, on error-free data ($\mathbf{X}$ is input to the algorithm). $\mathbf{A} = \mathbf{I}$, $\mathbf{m} = [1, 2, 3]^\intercal$, $K = 4$. In (d)–(e), $p = 0.2$. In (f)–(i), $p = 1$. In (e), the blocks are reconstructed properly both for JISA and mismodeling. Fig. (g) is a typical case where a fully random initialization prohibits the proper reconstruction of the blocks in JISA. Fig. (h)–(i) reflect the output in a mismodeling scenario before and after correcting the global permutation, respectively. False colours: in (b,c,e–i) we depict $\log 10|\cdot|$ in order to enhance small numerical features. White=zero.

same MSE, up to numerical precision. In the experiments that follow, only $\mathbf{A}_{\text{init}}$ varies at each of the 100 MC trials, while $\mathbf{A}$, $\mathbf{S}$ and $\overline{\mathbf{S}}$ (the empirical counterpart of $\mathbf{S}$) remain fixed. In this example, $\mathbf{m} = [1, 2, 4]^\intercal$, $K = 4$. Fig. 2 validates that indeed, the QN algorithm improves over RG in terms of number of iterations. In addition, these results reflect the fact that in mismodeling, the algorithm is trying to block-diagonalize $\widetilde{\mathbf{S}}$ into smaller blocks than is actually possible and thus doing unnecessary work. These results conform with previous ones [1, 11]. In this scenario, we set $p = 0.15$. This value guaranteed proper convergence to the correct minimum of the contrast function in all trials. For larger values of p, the number of iterations in the RG becomes prohibitive. Hence, in this respect, the advantage of the QN approach over RG is clear.

4.2 Component Separation

The component separation quality of the QN algorithm, quantified by its MSE, is illustrated in Fig. 3. In the following experiment, we run mutliple trials for fixed $\mathbf{S}$, $\mathbf{A}$, and $\mathbf{A}_{\text{init}}$. Only $\overline{\mathbf{S}}$ varies. For each trial we evaluate the normalized empirical MSE (3). As in the previous experiments, we compare JISA with its mismodeling counterpart. We set $\mathbf{m} = [6, 5, 1]^\intercal$, $K = 5$. Here, $M = 12$ which is prohibitive for enumeration (recall Fig. 1(h)–1(i)) and thus we use $p = 0.2$ in order to have a good chance that the output is automatically clustered properly into the correct N blocks, before evaluating the MSE (3). In fact, the convexity of the "mismodeled" variant, mentioned in Sec. 4.1, no longer holds as m_i largely

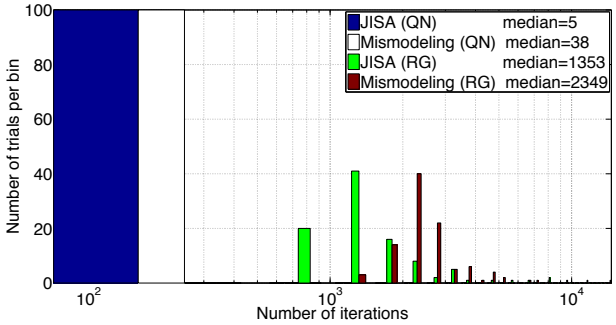

Fig. 2. Histogram of number of iterations in QN and RG, on the same data. Init with mild perturbation ($p = 0.15$): all runs converged properly. Logarithmic X-axis.

diverges from 1. In order to overcome this *for the error-analysis validation purpose only*, we choose a more strict initialization strategy in which in the first attempt $\mathbf{A}_{\text{init}}$ is taken from the output of the JISA run, and if the empirical MSE indicates large errors, new $\mathbf{A}_{\text{init}}$ are generated according to the original procedure until no permutation issues are detected. Fig. 3 illustrates our results. Subplot i corresponds to component i. The mean and standard deviation (std) of $\widehat{\text{MSE}}_i$ are written above the corresponding sublot, together with the theoretically predicted MSE for the JISA scenario [7]. The histograms represent 200 MC trials. The averaged MSE and its theoretical prediction for JISA are marked on the histograms. Fig. 3 shows good fit between the predicted and empirical values. It also shows improved MSE from using the correct multidimensional model, including for the component with $m_i = 1$, as expected.

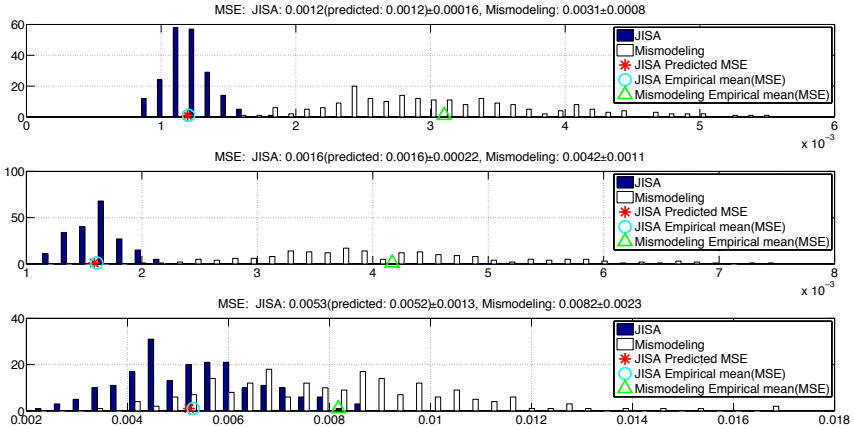

Fig. 3. Component separation. Histogram of the normalized empirical MSE for correct and mismodeling scenarios. Subplots correspond to components with dimensions 6, 5 and 1, respectively. $K = 5$, 200 trials.

Concluding remarks: In this paper, we introduced a new Newton-based algorithm for JISA that achieves asymptotically optimal performance for Gaussian noise-free data. Many other issues remain to be explored, such as its nu-

merical complexity, dependence of MSE on model parameters, efficient implementation, and behaviour in the presence of real-life data. We mention that generalizing [12], JISA can be regarded as a coupled block diagonalization problem, since $\mathbf{X}^{[k,l]} = \mathbf{A}^{[k]}\mathbf{S}^{[k,l]}\mathbf{A}^{[l]\mathsf{T}}$ $\forall k, l$, where $\mathbf{X}^{[k,l]} = E\{\mathbf{x}^{[k]}(t)\mathbf{x}^{[l]\mathsf{T}}(t)\}$ and $\mathbf{S}^{[k,l]} \in \mathcal{B}_{\mathbf{m}}$. Consequently, JISA falls within the framework of structured data fusion (SDF) [13] and thus can be solved, using a straightforward model-fit approach and a Euclidean norm, using Tensorlab [14]. Comparison with the QN algorithm is left for future work.

References

1. D. Lahat and C. Jutten, "Joint blind source separation of multidimensional components: Model and algorithm," in *Proc. EUSIPCO*, Lisbon, Portugal, Sep. 2014, pp. 1417–1421.
2. P. Comon, "Supervised classification, a probabilistic approach," in *Proc. ESANN*, Brussels, Belgium, Apr. 1995, pp. 111–128.
3. L. De Lathauwer, B. De Moor, and J. Vandewalle, "Fetal electrocardiogram extraction by source subspace separation," in *Proc. IEEE SP/ATHOS Workshop on HOS*, Girona, Spain, Jun. 1995, pp. 134–138.
4. J.-F. Cardoso, "Multidimensional independent component analysis," in *Proc. ICASSP*, vol. 4, Seattle, WA, May 1998, pp. 1941–1944.
5. 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*, ser. LNCS, vol. 3889. Springer Berlin Heidelberg, 2006, pp. 165–172.
6. R. F. Silva, S. Plis, T. Adalı, and V. D. Calhoun, "Multidataset independent subspace analysis extends independent vector analysis," in *Proc. ICIP*, Paris, France, Oct. 2014, pp. 2864–2868.
7. D. Lahat and C. Jutten, "Joint independent subspace analysis using second-order statistics," GIPSA-Lab, Technical report hal-01132297, Mar. 2015. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01132297
8. S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, Mar. 1951.
9. D. Lahat, J.-F. Cardoso, and H. Messer, "Blind separation of multidimensional components via subspace decomposition: Performance analysis," *IEEE Trans. Signal Process.*, vol. 62, no. 11, pp. 2894–2905, Jun. 2014.
10. M. Anderson, T. Adalı, and X.-L. Li, "Joint blind source separation with multivariate Gaussian model: Algorithms and performance analysis," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1672–1683, Apr. 2012.
11. D. Lahat, J.-F. Cardoso, and H. Messer, "Joint block diagonalization algorithms for optimal separation of multidimensional components," in *Latent Variable Analysis and Signal Separation*, ser. LNCS, F. Theis, A. Cichocki, A. Yeredor, and M. Zibulevsky, Eds., vol. 7191. Heidelberg: Springer, 2012, pp. 155–162.
12. X.-L. Li, T. Adalı, and M. Anderson, "Joint blind source separation by generalized joint diagonalization of cumulant matrices," *Signal Process.*, vol. 91, no. 10, pp. 2314–2322, Oct. 2011.
13. L. Sorber, M. Van Barel, and L. De Lathauwer, "Structured data fusion," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 4, Jun. 2015.
14. ——, "Tensorlab v2.0," Jan. 2014. [Online]. Available: http://www.tensorlab.net/