Shared Independent Component Analysis for Multi-Subject Neuroimaging

Hugo Richard, Pierre Ablin, Bertrand Thirion, Alexandre Gramfort, Aapo Hyvärinen

Shared Independent Component Analysis for Multi-Subject Neuroimaging

Hugo Richard
Inria
Université Paris-Saclay
Palaiseau, France

Pierre Ablin
DMA
CNRS and ENS
Paris, France

Bertrand Thirion
Inria
Université Paris-Saclay
Palaiseau, France

Alexandre Gramfort
Inria
Université Paris-Saclay
Palaiseau, France

Aapo Hyvärinen
Department of Computer Science
University of Helsinki
Helsinki, Finland

Abstract

We consider shared response modeling, a multi-view learning problem where one wants to identify common components from multiple datasets or views. We introduce Shared Independent Component Analysis (ShICA) that models each view as a linear transform of shared independent components contaminated by additive Gaussian noise. We show that this model is identifiable if the components are either non-Gaussian or have enough diversity in noise variances. We then show that in some cases multi-set canonical correlation analysis can recover the correct unmixing matrices, but that even a small amount of sampling noise makes Multiset CCA fail. To solve this problem, we propose to use joint diagonalization after Multiset CCA, leading to a new approach called ShICA-J. We show via simulations that ShICA-J leads to improved results while being very fast to fit. While ShICA-J is based on second-order statistics, we further propose to leverage non-Gaussianity of the components using a maximum-likelihood method, ShICA-ML, that is both more accurate and more costly. Further, ShICA comes with a principled method for shared components estimation. Finally, we provide empirical evidence on fMRI and MEG datasets that ShICA yields more accurate estimation of the components than alternatives.

1 Introduction

In many data science problems, data are available through different views. Generally, the views represent different measurement modalities such as audio and video, or the same text that may be available in different languages. Our main interest here is neuroimaging where recordings are made from multiple subjects. In particular, it is of interest to find common patterns or responses that are shared between subjects when they receive the same stimulation or perform the same cognitive task [16, 51].

A popular line of work to perform such shared response modeling is group Independent Component Analysis (ICA) methods. The fastest methods [14, 58] are among the most popular, yet they are not grounded on principled probabilistic models for the multiview setting. More principled approaches exist [51, 27], but they do not model subject-specific deviations from the shared response. However, such deviations are expected in most neuroimaging settings, as the magnitude of the response may differ from subject to subject [46], as may any noise due to heartbeats, respiratory artefacts or head

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

movements [39]. Furthermore, most GroupICA methods are typically unable to separate components whose density is close to a Gaussian.

Independent vector analysis (IVA) [36, 5] is a powerful framework where components are independent within views but each component of a given view can depend on the corresponding component in other views. However, current implementations such as IVA-L [36], IVA-G [5], IVA-L-SOS [11], IVA-GGD [7] or IVA with Kotz distribution [6] estimate only the view-specific components, and do not model or extract a shared response which is the main focus in this work.

On the other hand, the shared response model [16] is a popular approach to perform shared response modeling, yet it imposes orthogonality constrains that are restrictive and not biologically plausible.

In this work we introduce Shared ICA (ShICA), where each view is modeled as a linear transform of shared independent components contaminated by additive Gaussian noise. ShICA allows the principled extraction of the shared components (or responses) in addition to view-specific components. Since it is based on a statistically sound noise model, it enables optimal inference (minimum mean square error, MMSE) of the shared responses.

Let us note that ShICA is no longer the method of choice when the concept of common response is either not useful or not applicable. Nevertheless, we believe that the ability to extract a common response is an important feature in most contexts because it highlights a stereotypical brain response to a stimulus. Moreover, finding commonality between subjects reduces often unwanted inter-subject variability.

The paper is organized as follows. We first analyse the theoretical properties of the ShICA model, before providing inference algorithms. We exhibit necessary and sufficient conditions for the ShICA model to be identifiable (previous work only shows local identifiability [7]), in the presence of Gaussian or non-Gaussian components. We then use Multiset CCA to fit the model when all the components are assumed to be Gaussian. We exhibit necessary and sufficient conditions for Multiset CCA to be able to recover the unmixing matrices (previous work only gives sufficient conditions [38]). In addition, we provide instances of the problem where Multiset CCA cannot recover the mixing matrices while the model is identifiable. We next point out a practical problem : even a small sampling noise can lead to large error in the estimation of unmixing matrices when Multiset CCA is used. To address this issue and recover the correct unmixing matrices, we propose to apply joint diagonalization to the result of Multiset CCA yielding a new method called ShICA-J. We further introduce ShICA-ML, a maximum likelihood estimator of ShICA that models non-Gaussian components using a Gaussian mixture model. While ShICA-ML yields more accurate components, ShICA-J is significantly faster and offers a great initialization to ShICA-ML. Experiments on fMRI and MEG data demonstrate that the method outperforms existing GroupICA and IVA methods.

2 Shared ICA (ShICA): an identifiable multi-view model

Notation We write vectors in bold letter $\mathbf{v}$ and scalars in lower case a. Upper case letters M are used to denote matrices. We denote $|M|$ the absolute value of the determinant of M. $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ means that $\mathbf{x} \in \mathbb{R}^k$ follows a multivariate normal distribution of mean $\boldsymbol{\mu} \in \mathbb{R}^k$ and covariance $\Sigma \in \mathbb{R}^{k \times k}$. The j, j entry of a diagonal matrix Σ_i is denoted Σ_{ij}, the j entry of $\mathbf{y}_i$ is denoted y_{ij}. Lastly, δ is the Kronecker delta.

Model Definition In the following, $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{R}^p$ denote the m observed random vectors obtained from the m different views. We posit the following generative model, called Shared ICA (ShICA): for $i = 1 \dots m$

$$\mathbf{x}_i = A_i(\mathbf{s} + \mathbf{n}_i) \tag{1}$$

where $\mathbf{s} \in \mathbb{R}^p$ contains the latent variables called *shared components*, $A_1, \dots, A_m \in \mathbb{R}^{p \times p}$ are the invertible mixing matrices, and $\mathbf{n}_i \in \mathbb{R}^p$ are *individual noises*. The individual noises model both the deviations of a view from the mean —i.e. individual differences— and measurement noise. Importantly, we explicitly model both the shared components and the individual differences in a probabilistic framework to enable an optimal inference of the parameters and the responses.

We assume that the shared components are statistically independent, and that the individual noises are Gaussian and independent from the shared components: $p(\mathbf{s}) = \prod_{j=1}^{p} p(s_j)$ and $\mathbf{n}_i \sim \mathcal{N}(0, \Sigma_i)$, where the matrices Σ_i are assumed diagonal and positive. Without loss of generality, components

are assumed to have unit variance $\mathbb{E}[\mathbf{s}\mathbf{s}^\top] = I_p$. We further assume that there are at least 3 views: $m \geq 3$.

In contrast to almost all existing works, we assume that some components (possibly all of them) may be Gaussian, and denote $\mathcal{G}$ the set of Gaussian components: $\mathbf{s}_j \sim \mathcal{N}(0,1)$ for $j \in \mathcal{G}$. The other components are non-Gaussian: for $j \notin \mathcal{G}$, $\mathbf{s}_j$ is non-Gaussian.

Identifiability The parameters of the model are $\Theta = (A_1, \ldots, A_m, \Sigma_1, \ldots, \Sigma_m)$. We are interested in the identifiability of this model: given observations $\mathbf{x}_1, \ldots, \mathbf{x}_m$ generated with parameters Θ, are there some other Θ' that may generate the same observations? Let us consider the following assumption that requires that the individual noises for Gaussian components are sufficiently diverse:

Assumption 1 (Noise diversity in Gaussian components)**.** *For all $j, j' \in \mathcal{G}, j \neq j'$, the sequences $(\Sigma_{ij})_{i=1\ldots m}$ and $(\Sigma_{ij'})_{i=1\ldots m}$ are different where Σ_{ij} is the j,j entry of Σ_i*

It is readily seen that there is one trivial set of indeterminacies in the problem: if $P \in \mathbb{R}^{p\times p}$ is a sign and permutation matrix (i.e. a matrix which has one ± 1 coefficient on each row and column, and 0's elsewhere) the parameters $(A_1P, \ldots, A_mP, P^\top\Sigma_1 P, \ldots, P^\top\Sigma_m P)$ also generate $\mathbf{x}_1, \ldots, \mathbf{x}_m$. The following theorem shows that under the above assumption, these are the only indeterminacies of the problem.

Theorem 1 (Identifiability)**.** *We make Assumption 1. We let $\Theta' = (A'_1, \ldots, A'_m, \Sigma'_1, \ldots, \Sigma'_m)$ another set of parameters, and assume that they also generate $\mathbf{x}_1, \ldots, \mathbf{x}_m$. Then, there exists a sign and permutation matrix P such that for all i, $A'_i = A_iP$, and $\Sigma'_i = P^\top\Sigma_i P$.*

The proof is in Appendix A.1. Identifiability in the Gaussian case is a consequence of the identifiability results in [59] and in the general case, local identifiability results can be derived from the work of [7]. However local identifiability only shows that for a given set of parameters there exists a neighborhood in which no other set of parameters can generate the same observations [52]. In contrast, the proof of Theorem 1 shows global identifiability.

Theorem 1 shows that the task of recovering the parameters from the observations is a well-posed problem, under the sufficient condition of Assumption 1. We also note that Assumption 1 is necessary for identifiability. For instance, if j and j' are two Gaussian components such that $\Sigma_{ij} = \Sigma_{ij'}$ for all i, then a global rotation of the components j, j' yields the same covariance matrices. The current work assumes $m \geq 3$, in appendix B we give an identifiability result for $m = 2$, under stronger conditions.

3 Estimation of components with noise diversity via joint-diagonalization

We now consider the computational problem of efficient parameter inference. This section considers components with noise diversity, while the next section deals with non-Gaussian components.

3.1 Parameter estimation with Multiset CCA

If we assume that the components are all Gaussian, the covariance of the observations given by $C_{ij} = \mathbb{E}[\mathbf{x}_i\mathbf{x}_j^\top] = A_i(I_p + \delta_{ij}\Sigma_i)A_j^\top$ are sufficient statistics and methods using only second order information, like Multiset CCA, are candidates to estimate the parameters of the model. Consider the matrix $\mathcal{C} \in \mathbb{R}^{pm\times pm}$ containing $m \times m$ blocks of size $p \times p$ such that the block i, j is given by C_{ij}. Consider the matrix $\mathcal{D}$ identical to $\mathcal{C}$ excepts that the non-diagonal blocks are filled with zeros:

$$\mathcal{C} = \begin{bmatrix} C_{11} & \ldots & C_{1m} \\ \vdots & \ddots & \vdots \\ C_{m1} & \ldots & C_{mm} \end{bmatrix}, \quad \mathcal{D} = \begin{bmatrix} C_{11} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & C_{mm} \end{bmatrix}. \tag{2}$$

Generalized CCA consists of the following generalized eigenvalue problem:

$$\mathcal{C}\mathbf{u} = \lambda\mathcal{D}\mathbf{u}, \quad \lambda > 0, \quad \mathbf{u} \in \mathbb{R}^{pm} \quad . \tag{3}$$

Consider the matrix $U = [\mathbf{u}^1, \ldots, \mathbf{u}^p] \in \mathbb{R}^{mp\times p}$ formed by concatenating the p leading eigenvectors of the previous problem ranked in decreasing eigenvalue order. Then, consider U to be formed of m

blocks of size $p \times p$ stacked vertically and define $(W^i)^\top$ to be the i-th block. These m matrices are the output of Multiset CCA. We also denote $\lambda_1 \geq \cdots \geq \lambda_p$ the p leading eigenvalues of the problem.

An application of the results of [38] shows that Multiset CCA recovers the mixing matrices of ShICA under some assumptions.

Proposition 1 (Sufficient condition for solving ShICA via Multiset CCA [38])**.** *Let $r_{ijk} = (1 + \Sigma_{ik})^{-\frac{1}{2}}(1 + \Sigma_{jk})^{-\frac{1}{2}}$. Assume that $(r_{ijk})_k$ is non-increasing. Assume that the maximum eigenvalue ν_k of matrix $R^{(k)}$ of general element $(r_{ijk})_{ij}$ is such that $\nu_k = \lambda_k$. Assume that $\lambda_1 \dots \lambda_p$ are distinct. Then, there exists scale matrices Γ_i such that $W_i = \Gamma_i A_i^{-1}$ for all i.*

This proposition gives a sufficient condition for solving ShICA with Multiset CCA. It needs a particular structure for the noise covariances as well as specific ordering for the eigenvalues. The next theorem shows that we only need $\lambda_1 \dots \lambda_p$ to be distinct for Multiset CCA to solve ShICA:

Assumption 2 (Unique eigenvalues)**.** *$\lambda_1 \dots \lambda_p$ are distinct.*

Theorem 2. *We only make Assumption 2. Then, there exists a permutation matrix P and scale matrices Γ_i such that $W_i = P\Gamma_i A_i^{-1}$ for all i.*

The proof is in Appendix A.2. This theorem means that solving the generalized eigenvalue problem (3) allows to recover the mixing matrices up to a scaling and permutation: this form of generalized CCA recovers the parameters of the statistical model. Note that Assumption 2 is also a necessary condition. Indeed, if two eigenvalues are identical, the eigenvalue problem is not uniquely determined.

We have two different Assumptions, 1 and 2, the first of which guarantees theoretical identifiability as per Theorem 1 and the second guarantees consistent estimation by Multiset CCA as per Theorem 2. Next we will discuss their connections, and show some limitations of the Multiset CCA approach. To begin with, we have the following result about the eigenvalues of the problem (3) and the Σ_{ij}.

Proposition 2. *For $j \leq p$, let λ_j the largest solution of $\sum_{i=1}^{m} \frac{1}{\lambda_j(1+\Sigma_{ij})-\Sigma_{ij}} = 1$. Then, $\lambda_1, \dots, \lambda_p$ are the p largest eigenvalues of problem* (3)*.*

It is easy to see that we then have $\lambda_1, \dots, \lambda_p$ greater than 1, while the remaining eigenvalues are lower than 1. From this proposition, two things appear clearly. First, Assumption 2 implies Assumption 1. Indeed, if the λ_j's are distinct, then the sequences $(\Sigma_{ij})_i$ must also be different from the previous proposition. This is expected as from Theorem 2, Assumption 2 implies identifiability, which in turn implies Assumption 1.

Prop. 2 also allows us to derive cases where Assumption 1 holds but not Assumption 2. The following Proposition gives a simple case where the model is identifiable but it cannot be solved using Multiset CCA:

Proposition 3. *Assume that for two integers j, j', the sequence $(\Sigma_{ij})_i$ is a permutation of $(\Sigma_{ij'})_i$, i.e. that there exists a permutation of $\{1, \dots, p\}$, π, such that for all i, $\Sigma_{ij} = \Sigma_{\pi(i)j'}$. Then, $\lambda_j = \lambda_{j'}$.*

In this setting, Assumption 1 holds so ShICA is identifiable, while Assumption 2 does not hold, so Multiset CCA cannot recover the unmixing matrices.

3.2 Sampling noise and improved estimation with joint diagonalization

The consistency theory for Multiset CCA developed above is conducted under the assumption that the covariances C_{ij} are the true covariances of the model, and not approximations obtained from observed samples. In practice, however, a serious limitation of Multiset CCA is that even a slight error of estimation on the covariances, due to "sampling noise", can yield a large error in the estimation of the unmixing matrices, as will be shown next.

We begin with an empirical illustration. We take $m = 3$, $p = 2$, and Σ_i such that $\lambda_1 = 2+\varepsilon$ and $\lambda_2 = 2$ for $\varepsilon > 0$. In this way, we can control the *eigen-gap* of the problem, ε. We take W_i the outputs of Multiset CCA applied to the true covariances C_{ij}. Then, we generate a perturbation $\Delta = \delta \cdot S$, where S is a random positive symmetric $pm \times pm$ matrix of norm 1, and $\delta > 0$ controls the scale of the perturbation. We take Δ_{ij} the $p \times p$ block of Δ in position (i, j), and $\tilde{W}_i$ the output of Multiset CCA applied to the covariances $C_{ij} + \Delta_{ij}$. We finally compute the sum of the Amari distance between the W_i and $\tilde{W}_i$: the Amari distance measures how close the two matrices are, up to scale and permutation [4].

Fig 1 displays the median Amari distance over 100 random repetitions, as the perturbation scale δ increases. The different curves correspond to different values of the eigen-gap ε. We see clearly that the robustness of Multiset CCA critically depends on the eigen-gap, and when it is small, even a small perturbation of the input (due, for instance, to sampling noise) leads to large estimation errors.

This problem is very general and well studied [53]: the mapping from matrices to (generalized) eigenvectors is highly non-smooth. However, the gist of our method is that the *span* of the leading p eigenvectors is smooth, as long as there is a large enough gap between λ_p and λ_{p+1}. For our specific problem we have the following bounds, derived from Prop. 2.

Proposition 4. *We let* $\sigma_{\max} = \max_{ij} \Sigma_{ij}$ *and* $\sigma_{\min} = \min_{ij} \Sigma_{ij}$. *Then,* $\lambda_p \geq 1 + \frac{m-1}{1+\sigma_{\max}}$, *while* $\lambda_{p+1} \leq 1 - \frac{1}{1+\sigma_{min}}$.

As a consequence, we have $\lambda_p - \lambda_{p+1} \geq \frac{m-1}{1+\sigma_{\max}} + \frac{1}{1+\sigma_{\min}} \geq \frac{m}{1+\sigma_{\max}}$: the gap between these eigenvalues increases with m, and decreases with the noise power.

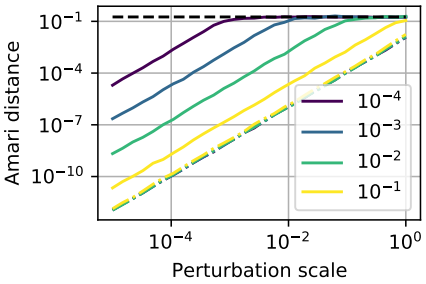


Figure 1: Amari distance between true mixing matrices and estimates of Multiset CCA when covariances are perturbed. Different solid curves correspond to different eigen-gaps. The black dotted line shows the chance level. When the gap is small, a small perturbation can lead to complete mixing. Joint-diagonalization (colored dotted lines) fixes the problem.

In this setting, when the magnitude of the perturbation Δ is smaller than $\lambda_p - \lambda_{p+1}$, [53] indicates that $\mathrm{Span}([W_1, \ldots, W_m]^\top) \simeq \mathrm{Span}([\tilde{W}_1, \ldots, \tilde{W}_m]^\top)$, where $[W_1, \ldots, W_m]^\top \in \mathbb{R}^{pm \times p}$ is the vertical concatenation of the W_i's. In turn, this shows that there exists a matrix $Q \in \mathbb{R}^{p \times p}$ such that

$$W_i \simeq Q\tilde{W}_i \quad \text{for all} \quad i. \tag{4}$$

We propose to use joint-diagonalization to recover the matrix Q. Given the $\tilde{W}_i$'s, we consider the set of symmetric matrices $\tilde{K}_i = \tilde{W}_i \tilde{C}_{ii} \tilde{W}_i^\top$, where $\tilde{C}_{ii}$ is the contaminated covariance of $\mathbf{x}_i$. Following Eq. (4), we have $Q\tilde{K}_i Q^\top = W_i \tilde{C}_{ii} W_i^\top$, and using Theorem 2, we have $Q\tilde{K}_i Q^\top = P\Gamma_i A_i^{-1} \tilde{C}_{ii} A_i^{-\top} \Gamma_i P^\top$. Since $\tilde{C}_{ii}$ is close to $C_{ii} = A_i(I_p + \Sigma_i)A_i^\top$, the matrix $P\Gamma_i A_i^{-1} \tilde{C}_{ii} A_i^{-\top} \Gamma_i P^\top$ is almost diagonal. In other words, the matrix Q is an approximate diagonalizer of the $\tilde{K}_i$'s, and we approximate Q by joint-diagonalization of the $\tilde{K}_i$'s. In Fig 1, we see that this procedure mitigates the problems of multiset-CCA, and gets uniformly better performance regardless of the eigen-gap. In practice, we use a fast joint-diagonalization algorithm [1] to minimize a joint-diagonalization criterion for positive symmetric matrices [48]. The estimated unmixing matrices $U_i = Q\tilde{W}_i$ correspond to the true unmixing matrices only up to some scaling which may be different from subject to subject: the information that the components are of unit variance is lost. As a consequence, naive averaging of the recovered components may lead to inconsistant estimation. We now describe a procedure to recover the correct scale of the individual components across subjects.

Algorithm 1 ShICA-J

Input : Covariances $\tilde{C}_{ij} = \mathbb{E}[\mathbf{x}_i \mathbf{x}_j^\top]$
$(\tilde{W}_i)_i \leftarrow \mathrm{MultisetCCA}((\tilde{C}_{ij})_{ij})$
$Q \leftarrow \mathrm{JointDiag}((\tilde{W}_i \tilde{C}_{ii} \tilde{W}_i^\top)_i)$
$\Gamma_{ij} \leftarrow Q\tilde{W}_i \tilde{C}_{ij} W_j^\top Q^\top$
$(\Phi_i)_i \leftarrow \mathrm{Scaling}((\Gamma_{ij})_{ij})$
Return : Unmixing matrices $(\Phi_i Q \tilde{W}_i)_i$.

Scale estimation We form the matrices $\Gamma_{ij} = U_i \tilde{C}_{ij} U_j^\top$. In order to estimate the scalings, we solve $\min_{(\Phi_i)} \sum_{i \neq j} \|\Phi_i \mathrm{diag}(\Gamma_{ij}) \Phi_j - I_p\|_F^2$ where the Φ_i are diagonal matrices. This function is readily minimized with respect to one of the Φ_i by the formula $\Phi_i = \frac{\sum_{j \neq i} \Phi_j \mathrm{diag}(Y_{ij})}{\sum_{j \neq i} \Phi_j^2 \mathrm{diag}(Y_{ij})^2}$ (derivations in Appendix 20). We then iterate the previous formula over i until convergence. The final estimates of the unmixing matrices are given by $(\Phi_i U_i)_{i=1}^m$. The full procedure, called ShICA-J, is summarized in Algorithm 1.

3.3 Estimation of noise covariances

In practice, it is important to estimate noise covariances Σ_i in order to take advantage of the fact that

some views are noisier than others. As it is well known in classical factor analysis, modelling noise variances allows the model to virtually discard variables, or subjects, that are particularly noisy.

Using the ShICA model with Gaussian components, we derive an estimate for the noise covariances directly from maximum likelihood. We use an expectation-maximization (EM) algorithm, which is especially fast because noise updates are in closed-form. Following derivations given in appendix D.1, the sufficient statistics in the E-step are given by

$$\mathbb{E}[\mathbf{s}|\mathbf{x}] = \left(\sum_{i=1}^{m} \Sigma_i^{-1} + I\right)^{-1} \sum_{i=1}^{m} \left(\Sigma_i^{-1}\mathbf{y}_i\right) \qquad \mathbb{V}[\mathbf{s}|\mathbf{x}] = (\sum_{i=1}^{m} \Sigma_i^{-1} + I)^{-1} \tag{5}$$

Incorporating the M-step we get the following updates that only depend on the covariance matrices: $\Sigma_i \leftarrow \operatorname{diag}(\hat{C}_{ii} - 2\mathbb{V}[\mathbf{s}|\mathbf{x}] \sum_{j=1}^{m} \Sigma_j^{-1}\hat{C}_{ji} + \mathbb{V}[\mathbf{s}|\mathbf{x}] \sum_{j=1}^{m}\sum_{l=1}^{m} \left(\Sigma_j^{-1}\hat{C}_{jl}\Sigma_l^{-1}\right) \mathbb{V}[\mathbf{s}|\mathbf{x}] + \mathbb{V}[\mathbf{s}|\mathbf{x}])$

4 ShICA-ML: Maximum likelihood for non-Gaussian components

ShICA-J only uses second order statistics. However, the ShICA model (1) allows for non-Gaussian components. We now propose an algorithm for fitting the ShICA model that combines covariance information with non-Gaussianity in the estimation to optimally separate both Gaussian and non-Gaussian components. We estimate the parameters by maximum likelihood. Since most non-Gaussian components in real data are super-Gaussian [21, 13], we assume that the non-Gaussian components $\mathbf{s}$ have the super-Gaussian density
$p(s_j) = \frac{1}{2}\left(\mathcal{N}(s_j; 0, \frac{1}{2}) + \mathcal{N}(s_j; 0, \frac{3}{2})\right)$.

We propose to maximize the log-likelihood using a generalized EM [41, 22]. Derivations are available in Appendix E. Like in the previous section, the E-step is in closed-form yielding the following sufficient statistics:

$$\mathbb{E}[s_j|\mathbf{x}] = \frac{\sum_{\alpha\in\{\frac{1}{2},\frac{3}{2}\}} \theta_\alpha \frac{\alpha\bar{y}_j}{\alpha+\bar{\Sigma}_j}}{\sum_{\alpha\in\{0.5,1.5\}} \theta_\alpha} \quad \text{and} \quad \mathbb{V}[s_j|\mathbf{x}] = \frac{\sum_{\alpha\in\{\frac{1}{2},\frac{3}{2}\}} \theta_\alpha \frac{\bar{\Sigma}_j\alpha}{\alpha+\bar{\Sigma}_j}}{\sum_{\alpha\in\{0.5,1.5\}} \theta_\alpha} \tag{6}$$

where $\theta_\alpha = \mathcal{N}(\bar{y}_j; 0, \bar{\Sigma}_j + \alpha)$, $\bar{y}_j = \frac{\sum_i \Sigma_{ij}^{-1} y_{ij}}{\sum_i \Sigma_{ij}^{-1}}$ and $\bar{\Sigma}_j = (\sum_i \Sigma_{ij}^{-1})^{-1}$ with $\mathbf{y}_i = W_i\mathbf{x}_i$. Noise updates are in closed-form and given by: $\Sigma_i \leftarrow \operatorname{diag}((\mathbf{y}_i - \mathbb{E}[\mathbf{s}|\mathbf{x}])(\mathbf{y}_i - \mathbb{E}[\mathbf{s}|\mathbf{x}])^\top + \mathbb{V}[\mathbf{s}|\mathbf{x}])$. However, no closed-form is available for the updates of unmixing matrices. We therefore perform quasi-Newton updates given by $W_i \leftarrow (I - \rho(\widehat{\mathcal{H}^{W_i}})^{-1}\mathcal{G}^{W_i})W_i$ where $\rho \in \mathbb{R}$ is chosen by backtracking line-search, $\widehat{\mathcal{H}^{W_i}_{a,b,c,d}} = \delta_{ad}\delta_{bc} + \delta_{ac}\delta_{bd}\frac{(y_{ib})^2}{\bar{\Sigma}_{ia}}$ is an approximation of the Hessian of the negative complete likelihood and $\mathcal{G}^{W_i} = -I + (\Sigma_i)^{-1}(\mathbf{y}_i - \mathbb{E}[\mathbf{s}|\mathbf{x}])(\mathbf{y}_i)^\top$ is the gradient.

We alternate between computing the statistics $\mathbb{E}[\mathbf{s}|\mathbf{x}]$, $\mathbb{V}[\mathbf{s}|\mathbf{x}]$ (E-step) and updates of parameters Σ_i and W_i for $i = 1 \dots m$ (M-step). Let us highlight that our EM algorithm and in particular the E-step resembles the one used in [40]. However because they assume noise on the sensors and not on the components, their formula for $\mathbb{E}[\mathbf{s}|\mathbf{x}]$ involves a sum with 2^p terms whereas we have only 2 terms. The resulting method is called ShICA-ML.

Minimum mean square error estimates in ShICA In ShICA-J as well as in ShICA-ML, we have a closed-form for the expected components given the data $\mathbb{E}[\mathbf{s}|\mathbf{x}]$, shown in equation (5) and (6) respectively. This provides minimum mean square error estimates of the shared components, and is an important benefit of explicitly modelling shared components in a probabilistic framework.

5 Related Work

ShICA combines theory and methods coming from different branches of "component analysis". It can be viewed as a GroupICA method, as an extension of Multiset CCA, as an Independent Vector Analysis method or, crucially, as an extension of the shared response model. In the setting studied here, ShICA improves upon all existing methods.

GroupICA GroupICA methods extract independent components from multiple datasets. In its original form[14], views are concatenated and then a PCA is applied yielding reduced data on which ICA is applied. One can also reduce the data using Multiset CCA instead of PCA, giving a method called *CanICA* [58]. Other works [24, 32] apply ICA separately on the datasets and attempt to match the decompositions afterwards. Although these works provide very fast methods, they do not rely on a well defined model like ShICA. Other GroupICA methods impose some structure on the mixing matrices such as the tensorial method of [10] or the group tensor model in [27] (which assumes identical mixing matrices up to a scaling) or [54] (which assumes identical mixing matrices but different components). In ShICA the mixing matrices are only constrained to be invertible. Lastly, maximum-likelihood based methods exist such as *MultiViewICA* [51] (MVICA) or the full model of [27]. These methods are weaker than ShICA as they use the same noise covariance across views and lack a principled method for shared response inference.

Multiset CCA In its basic formulation, CCA identifies a shared space between two datasets. The extension to more than two datasets is ambiguous, and many different generalized CCA methods have been proposed. [33] introduces 6 objective functions that reduce to CCA when $m = 2$ and [42] considered 4 different possible constrains leading to 24 different formulations of Multiset CCA. The formulation used in ShICA-J is refered to in [42] as SUMCORR with constraint 4 which is one of the fastest as it reduces to solving a generalized eigenvalue problem. The fact that CCA solves a well defined probabilistic model has first been studied in [9] where it is shown that CCA is identical to multiple battery factor analysis [12] (restricted to 2 views). This latter formulation differs from our model in that the noise is added on the sensors and not on the components which makes the model unidentifiable. Identifiable variants and generalizations can be obtained by imposing sparsity on the mixing matrices such as in [8, 34, 62] or non-negativity [20]. The work in [38] exhibits a set of sufficient (but not necessary) conditions under which a well defined model can be learnt by the formulation of Multiset CCA used in ShICA-J. The set of conditions we exhibit in this work are necessary and sufficient. We further emphasize that basic Multiset CCA provides a poor estimator as explained in Section 3.2.

Independent vector analysis Independent vector analysis [36] (IVA) models the data as a linear mixture of independent components $\mathbf{x}_i = A_i\mathbf{s}_i$ where each component s_{ij} of a given view i can depend on the corresponding component in other views ($(s_{ij})_{i=1}^m$ are not independent). Practical implementations of this very general idea assume a distribution for $p((s_{ij})_{i=1}^m)$. In IVA-L [36], $p((s_{ij})_{i=1}^m) \propto \exp(-\sqrt{\sum_i (s_{ij})^2})$ (so the variance of each component in each view is assumed to be the same), in IVA-G [5] or in [60], $p((s_{ij})_{i=1}^m) \sim \mathcal{N}(0, R_{ss})$ and [23] proposed a normal inverse-Gamma density. Let us also mention IVA-L-SOS [11], IVA-GGD [7] and IVA with Kotz distribution [6] that assume a non-Gaussian density general enough so that they can use both second and higher order statistics to extract view-specific components. The model of ShICA can be seen as an instance of IVA which specifically enables extraction of shared components from the subject specific components, unlike previous versions of IVA. In fact, ShICA comes with minimum mean square error estimates for the shared components that is often the quantity of interest. The IVA theory provides global identifiability conditions in the Gaussian case (IVA-G) [59] and local identifiability conditions in the general case [7] from which local identifiability conditions of ShICA could be derived. However, in this work, we provide global identifiability conditions for ShICA. Lastly, IVA can be performed using joint diagonalization of cross covariances [37, 19] although multiple matrices have to be learnt and cross-covariances are not necessarily symmetric positive definite which makes the algorithm slower and less principled.

Shared response model ShICA extracts shared components from multiple datasets, which is also the goal of the shared response model (SRM) [16]. The robust SRM [57] also allows to capture subject specific noise. However these models impose orthogonality constraints on the mixing matrices while ShICA does not. Deep variants of SRM exist such as [17] but while they release the orthogonality constrain, they are not very easy to train or interpret and have many hyper-parameters to tune. ShICA leverages ICA theory to provide a much more powerful model of shared responses.

Limitations The main limitation of this work is that the model cannot reduce the dimension inside each view : there are as many estimated sources as sensors. This might be problematic when the number of sensors is very high. In line with other methods, view-specific dimension reduction has to be done by some external method, typically view-specific PCA. Using specialized methods

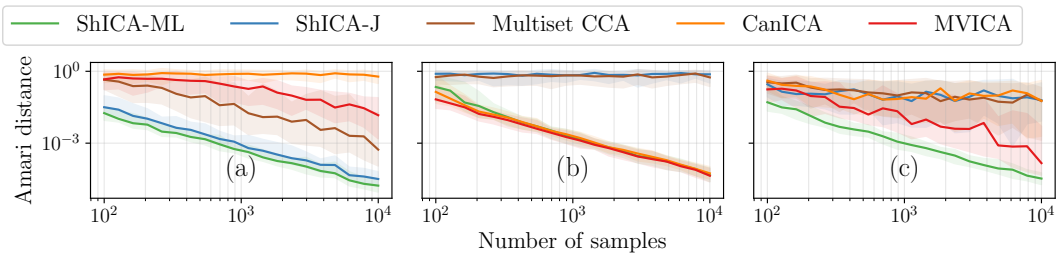


Figure 2: **Separation performance**: Algorithms are fit on data following model 1 **(a)** Gaussian components with noise diversity **(b)** Non-Gaussian components without noise diversity **(c)** Half of the components are Gaussian with noise diversity, the other half is non-Gaussian without noise diversity.

for the estimation of covariances should also be of interest for ShICA-J, where it only relies on sample covariances. Finally, ShICA-ML uses a simple model of a super-Gaussian distribution, while modelling the non-gaussianities in more detail in ShICA-ML should improve the performance.

6 Experiments

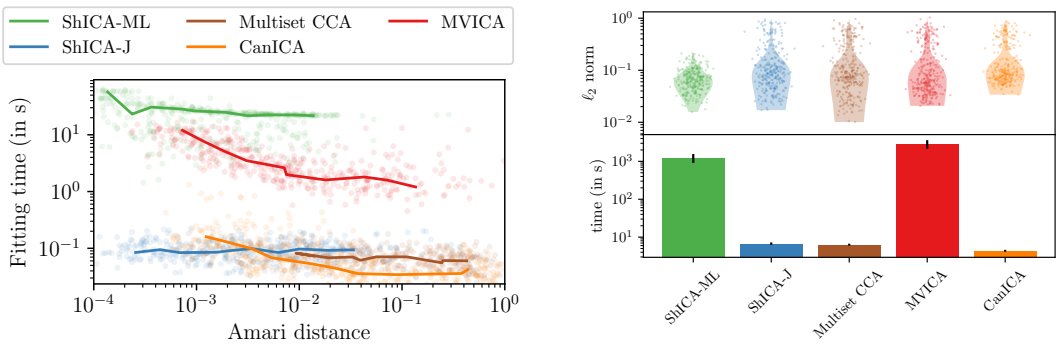
Experiments used Nilearn [3] and MNE [26] for fMRI and MEG data processing respectively, as well as the scientific Python ecosystem: Matplotlib [31], Scikit-learn [45], Numpy [29] and Scipy [61]. We use the Picard algorithm for non-Gaussian ICA [2], and mvlearn for multi-view ICA [47]. The above libraries use open-source licenses. fMRI experiments used the following datasets: sherlock [15], forrest [28] , raiders [49] and gallant [49]. The data we use do not contain offensive content or identifiable information and consent was obtained before data collection. Computations were run on a large server using up to 100 GB of RAM and 20 CPUs in parallel.

Separation performance In the following synthetic experiments, data are generated according to model (1) with $p = 4$ components and $m = 5$ views and mixing matrices are generated by sampling coefficients from a standardized Gaussian. Gaussian components are generated from a standardized Gaussian and their noise has standard deviation $\Sigma_i^{\frac{1}{2}}$ (obtained by sampling from a uniform density between 0 and 1) while non-Gaussian components are generated from a Laplace distribution and their noise standard deviations are equal. We study 3 cases where either all components are Gaussian, all components are non-Gaussian or half of the components are Gaussian and half are non-Gaussian. We vary the number of samples n between 10^2 and 10^5 and display in Fig 2 the mean Amari distance across subjects between the true unmixing matrices and estimates of algorithms as a function of n. The experiment is repeated 100 times using different seeds. We report the median result and error bars represent the first and last deciles.

When all components are Gaussian (Fig. 2 (a)), CanICA cannot separate the components at all. In contrast ShICA-J, ShICA-ML, Multiset CCA and MVICA are able to separate them, but Multiset CCA needs many more samples than ShICA-J or ShICA-ML to reach a low amari distance, which shows that correcting for the rotation due to sampling noise improves the results. Looking at error bars, we also see that the performance of Multiset CCA varies quite a lot with the random seeds: this shows that depending on the sampling noise, the rotation can be very different from identity. MVICA needs even more sample than Multiset CCA to reach a low amari distance but still outperforms CanICA.

When none of the components are Gaussian (Fig. 2 (b)), only CanICA, ShICA-ML and MVICA are able to separate the components, as other methods do not make use of non-Gaussianity. Finally, in the hybrid case (Fig. 2 (c)), ShICA-ML is able to separate the components as it can make use of both non-Gaussianity and noise diversity. MVICA is a lot less reliable than ShICA-ML, it is uniformly worse and error bars are very large showing that for some seeds it gives poor results. CanICA, ShICA-J and MultisetCCA cannot separate the components at all. Additional experiments illustrating the separation powers of algorithms are available in Appendix H.1.

As we can see, MVICA can separate Gaussian components to some extent and therefore does not completely fail when Gaussian and non-Gaussian components are present. However MVICA is a lot



(a) (b)

Figure 3: **Left: Computation time.** Algorithms are fit on data generated from model (1) with a super-Gaussian density. For different values of the number of samples, we plot the Amari distance and the fitting time. Thick lines link median values across seeds. **Right: Robustness w.r.t intra-subject variability in MEG.** (**top**) ℓ_2 distance between shared components corresponding to the same stimuli in different trials. (**bottom**) Fitting time.

less reliable than ShICA-ML: MVICA is uniformly worse than ShICA-ML and the error bars are very large showing that for some seeds it gives poor results.

Computation time We generate components using a slightly super Gaussian density: $s_j = d(x)$ with $d(x) = x|x|^{0.2}$ and $x \sim \mathcal{N}(0, 1)$. We vary the number of samples n between 10^2 and 10^4. We compute the mean Amari distance across subjects and record the computation time. The experiment is repeated 40 times. We plot the Amari distance as a function of the computation time in Fig 3a. Each point corresponds to the Amari distance/computation time for a given number of samples and a given seed. We then consider for a given number of samples, the median Amari distance and computation time across seeds and plot them in the form of a thick line. From Fig 3a, we see that ShICA-J is the method of choice when speed is a concern while ShICA-ML yields the best performance in terms of Amari distance at the cost of an increased computation time. The thick lines for ShICA-J and Multiset CCA are quasi-flat, indicating that the number of samples does not have a strong impact on the fitting time as these methods only work with covariances. On the other hand CanICA or MVICA computation time is more sensitive to the number of samples.

Robustness w.r.t intra-subject variability in MEG In the following experiments we consider the Cam-CAN dataset [56]. We use the magnetometer data from the MEG of $m = 100$ subjects chosen randomly among 496. In appendix F we give more information about Cam-CAN dataset. Each subject is repeatedly presented three audio-visual stimuli. For each stimulus, we divide the trials into two sets and within each set, the MEG signal is averaged across trials to isolate the evoked response. This procedure yields 6 chunks of individual data (2 per stimulus). We study the similarity between shared components corresponding to repetitions of the same stimulus. This gives a measure of robustness of each ICA algorithm with respect to intra-subject variability. Data are first reduced using a subject-specific PCA with $p = 10$ components. The initial dimensionality of the data before PCA is 102 as we only use the 102 magnetometers. Algorithms are run 10 times with different seeds on the 6 chunks of data, and shared components are extracted. When two chunks of data correspond to repetitions of the same stimulus they should yield similar components. For each component and for each stimulus, we therefore measure the ℓ_2 distance between the two repetitions of the stimulus. This yields 300 distances per algorithm that are plotted on Fig 3b.

The components recovered by ShICA-ML have a much lower variability than other approaches. The performance of ShICA-J is competitive with MVICA while being much faster to fit. Multiset CCA yields satisfying results compared with ShICA-J. However we see that the number of components that do not match at all across trials is greater in Multiset CCA. Additional experiments on MEG data are available in Appendix H.3.

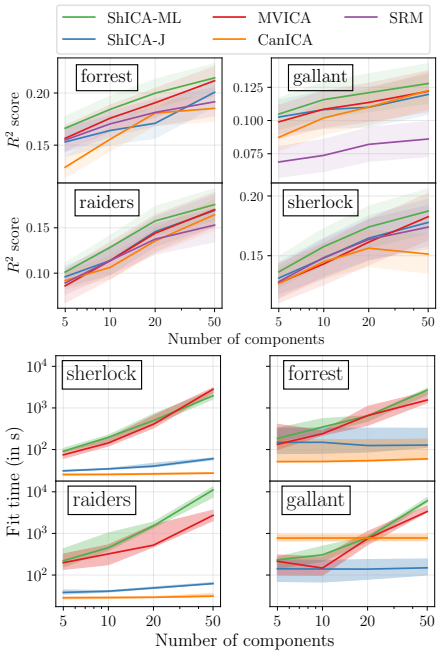


Figure 4: **Reconstructing the BOLD signal of missing subjects**. (**top**) Mean R^2 score between reconstructed data and true data. (**bottom**) Fitting time.

Reconstructing the BOLD signal of missing subjects We reproduce the experimental pipeline of [51] to benchmark GroupICA methods using their ability to reconstruct fMRI data of a left-out subject. The preprocessing involves a dimension reduction step performed using the shared response model [16]. Detailed preprocessing pipeline is described in Appendix F. We call an *unmixing operator* the product of the dimension reduction operator and an unmixing matrix and a *mixing operator* its pseudoinverse. There is one unmixing operator and one mixing operator per view. The unmixing operators are learned using all subjects and 80% of the runs. Then they are applied on the remaining 20% of the runs using 80% of the subjects yielding unmixed data from which shared components are extracted. The unmixed data are combined by averaging (for SRM and other baselines) or using the MMSE estimate for ShICA-J and ShICA-ML. We then apply the mixing operator of the remaining 20% subjects on the shared components to reconstruct their data. Reconstruction accuracy is measured via the coefficient of determination, *a.k.a.* R^2 score, that yields for each voxel the relative discrepancy between the true time course and the predicted one. For each compared algorithm, the experiment is run 25 times with different seeds to obtain error bars. We report the mean R^2 score across voxels in a region of interest (see Appendix F for details) and display the results in Fig 4. The error bars represent a 95% confidence interval. The chance level is given by the R^2 score of an algorithm that samples the coefficients of its unmixing matrices and dimension reduction operators from a standardized Gaussian. The median chance level is below 10^{-3} on all datasets. ShICA-ML yields the best R^2 score in all datasets and for any number of components. ShICA-J yields competitive results with respect to MVICA while being much faster to fit. A popular benchmark especially in the SRM community is the time-segment matching experiment [16]: we include such experiments in Appendix H.2. In appendix G, we give the performance of ShICA-ML, ShICA-J and MVICA in form of a table.

7 Conclusion, Future work and Societal impact

We introduced the ShICA model as a principled unifying solution to the problems of shared response modelling and GroupICA. ShICA is able to use both the diversity of Gaussian variances and non-Gaussianity for optimal estimation. We presented two algorithms to fit the model: ShICA-J, a fast algorithm that uses noise diversity, and ShICA-ML, a maximum likelihood approach that can use non-Gaussianity on top of noise diversity. ShICA algorithms come with principled procedures for shared components estimation, as well as adaptation and estimation of noise levels in each view (subject) and component. On simulated data, ShICA clearly outperforms all competing methods in terms of the trade-off between statistical accuracy and computation time. On brain imaging data, ShICA gives more stable decompositions for comparable computation times, and more accurately predicts the data of one subject from the data of other subjects, making it a good candidate to perform transfer learning. Our code is available at `https://github.com/hugorichard/ShICA`. *

*Regarding the ethical aspects of this work, we think this work presents exactly the same issues as any brain imaging analysis method related to ICA.

Acknowledgement and funding disclosure This work has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3), the KARAIB AI chair (ANR-20-CHIA-0025-01), the Grant SLAB ERC-StG-676943 and the BrAIN AI chair (ANR-20-CHIA-0016). PA acknowledges funding by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR19-P3IA-0001 (PRAIRIE 3IA Institute). AH received funding from a CIFAR Fellowship.

References

[1] Pierre Ablin, Jean-François Cardoso, and Alexandre Gramfort. Beyond pham's algorithm for joint diagonalization. *arXiv preprint arXiv:1811.11433*, 2018.

[2] Pierre Ablin, Jean-François Cardoso, and Alexandre Gramfort. Faster independent component analysis by preconditioning with hessian approximations. *IEEE Transactions on Signal Processing*, 66(15):4040–4049, 2018.

[3] Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics*, 8:14, 2014.

[4] Shun-ichi Amari, Andrzej Cichocki, Howard Hua Yang, et al. A new learning algorithm for blind signal separation. In *Advances in neural information processing systems*, pages 757–763. Morgan Kaufmann Publishers, 1996.

[5] Matthew Anderson, Tuelay Adali, and Xi-Lin Li. Joint blind source separation with multivariate gaussian model: Algorithms and performance analysis. *IEEE Transactions on Signal Processing*, 60(4):1672–1683, 2011.

[6] Matthew Anderson, Geng-Shen Fu, Ronald Phlypo, and Tülay Adali. Independent vector analysis, the kotz distribution, and performance bounds. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3243–3247. IEEE, 2013.

[7] Matthew Anderson, Geng-Shen Fu, Ronald Phlypo, and Tülay Adalı. Independent vector analysis: Identification conditions and performance bounds. *IEEE Transactions on Signal Processing*, 62(17):4399–4410, 2014.

[8] Cédric Archambeau and Francis R Bach. Sparse probabilistic projections. In *NIPS*, pages 73–80, 2008.

[9] Francis R Bach and Michael I Jordan. A probabilistic interpretation of canonical correlation analysis. Technical Report 688, Department of Statistics, University of California, Berkeley, 2005.

[10] Christian F Beckmann and Stephen M Smith. Tensorial extensions of independent component analysis for multisubject fMRI analysis. *Neuroimage*, 25(1):294–311, 2005.

[11] Suchita Bhinge, Rami Mowakeaa, Vince D Calhoun, and Tülay Adalı. Extraction of time-varying spatiotemporal networks using parameter-tuned constrained iva. *IEEE transactions on medical imaging*, 38(7):1715–1725, 2019.

[12] Michael W Browne. Factor analysis of multiple batteries by maximum likelihood. *British Journal of Mathematical and Statistical Psychology*, 33(2):184–199, 1980.

[13] Vince D Calhoun and Tülay Adali. Unmixing fmri with independent component analysis. *IEEE Engineering in Medicine and Biology Magazine*, 25(2):79–90, 2006.

[14] Vince D Calhoun, Tülay Adali, Godfrey D Pearlson, and James J Pekar. A method for making group inferences from functional MRI data using independent component analysis. *Human brain mapping*, 14(3):140–151, 2001.

[15] Janice Chen, Yuan C Leong, Christopher J Honey, Chung H Yong, Kenneth A Norman, and Uri Hasson. Shared memories reveal shared structure in neural activity across individuals. *Nature neuroscience*, 20(1):115–125, 2017.

[16] Po-Hsuan Chen, Janice Chen, Yaara Yeshurun, Uri Hasson, James V Haxby, and Peter J Ramadge. A reduced-dimension fMRI shared response model. In *NIPS*, volume 28, pages 460–468, 2015.

[17] Po-Hsuan Chen, Xia Zhu, Hejia Zhang, Javier S Turek, Janice Chen, Theodore L Willke, Uri Hasson, and Peter J Ramadge. A convolutional autoencoder for multi-subject fMRI data aggregation. *arXiv preprint arXiv:1608.04846*, 2016.

[18] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[19] Marco Congedo, Ronald Phlypo, and Jonas Chatel-Goldman. Orthogonal and non-orthogonal joint blind source separation in the least-squares sense. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pages 1885–1889. IEEE, 2012.

[20] Filip Deleus and Marc M. Van Hulle. Functional connectivity analysis of fMRI data based on regularized multiset canonical correlation analysis. *Journal of Neuroscience Methods*, 197(1):143–157, 2011.

[21] Arnaud Delorme, Jason Palmer, Julie Onton, Robert Oostenveld, and Scott Makeig. Independent eeg sources are dipolar. *PloS one*, 7(2):e30135, 2012.

[22] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[23] Astrid ME Engberg, Kasper W Andersen, Morten Mørup, and Kristoffer H Madsen. Independent vector analysis for capturing common components in fMRI group analysis. In *2016 international workshop on pattern recognition in neuroimaging (prni)*, pages 1–4. IEEE, 2016.

[24] F. Esposito, T. Scarabino, A. Hyvärinen, J. Himberg, E. Formisano, S. Comani, G. Tedeschi, R. Goebel, E. Seifritz, and F. Di Salle. Independent component analysis of fMRI group studies by self-organizing clustering. *NeuroImage*, 25(1):193–205, 2005.

[25] Gene H Golub. Some modified matrix eigenvalue problems. *Siam Review*, 15(2):318–334, 1973.

[26] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. MEG and EEG data analysis with MNE-Python. *Frontiers in neuroscience*, 7:267, 2013.

[27] Ying Guo and Giuseppe Pagnoni. A unified framework for group independent component analysis for multi-subject fMRI data. *NeuroImage*, 42(3):1078–1093, 2008.

[28] Michael Hanke, Florian J Baumgartner, Pierre Ibe, Falko R Kaule, Stefan Pollmann, Oliver Speck, Wolf Zinke, and Jörg Stadler. A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie. *Scientific data*, 1:140003, 2014.

[29] Charles R. Harris, K. Jarrod Millman, St'efan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fern'andez del R'ıo, Mark Wiebe, Pearu Peterson, Pierre G'erard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[30] James V Haxby, J Swaroop Guntupalli, Andrew C Connolly, Yaroslav O Halchenko, Bryan R Conroy, M Ida Gobbini, Michael Hanke, and Peter J Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011.

[31] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(3):90–95, 2007.

[32] A. Hyvärinen. Testing the ICA mixing matrix based on inter-subject or inter-session consistency. *NeuroImage*, 58(1):122–136, 2011.

[33] Jon R Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.

[34] Arto Klami, Seppo Virtanen, Eemeli Leppäaho, and Samuel Kaski. Group factor analysis. *IEEE transactions on neural networks and learning systems*, 26(9):2136–2147, 2014.

[35] Manoj Kumar, Michael J Anderson, James W Antony, Christopher Baldassano, Paula P Brooks, Ming Bo Cai, Po-Hsuan Cameron Chen, Cameron T Ellis, Gregory Henselman-Petrusek, David Huberdeau, et al. Brainiak: The brain imaging analysis kit. 2020.

[36] Jong-Hwan Lee, Te-Won Lee, Ferenc A Jolesz, and Seung-Schik Yoo. Independent vector analysis (IVA): multivariate approach for fMRI group study. *Neuroimage*, 40(1):86–109, 2008.

[37] Xi-Lin Li, Tülay Adalı, and Matthew Anderson. Joint blind source separation by generalized joint diagonalization of cumulant matrices. *Signal Processing*, 91(10):2314–2322, 2011.

[38] Yi-Ou Li, Tülay Adali, Wei Wang, and Vince D Calhoun. Joint blind source separation by multiset canonical correlation analysis. *IEEE Transactions on Signal Processing*, 57(10):3918–3929, 2009.

[39] Thomas T Liu. Noise contributions to the fmri signal: An overview. *NeuroImage*, 143:141–151, 2016.

[40] Eric Moulines, J-F Cardoso, and Elisabeth Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *1997 ieee international conference on acoustics, speech, and signal processing*, volume 5, pages 3617–3620. IEEE, 1997.

[41] Radford M Neal and Geoffrey E Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

[42] Allan Aasbjerg Nielsen. Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *IEEE transactions on image processing*, 11(3):293–305, 2002.

[43] Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19):1641–1646, 2011.

[44] Roberto Domingo Pascual-Marqui et al. Standardized low-resolution brain electromagnetic tomography (sloreta): technical details. *Methods Find Exp Clin Pharmacol*, 24(Suppl D):5–12, 2002.

[45] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[46] W Penny and A Holmes. Random effects analysis. *Statistical parametric mapping: The analysis of functional brain images*, 156:165, 2007.

[47] Ronan Perry, Gavin Mischler, Richard Guo, Theo Lee, Alexander Chang, Arman Koul, Cameron Franz, and Joshua T Vogelstein. mvlearn: Multiview machine learning in python. *arXiv preprint arXiv:2005.11890*, 2020.

[48] Dinh Tuan Pham. Joint approximate diagonalization of positive definite hermitian matrices. *SIAM Journal on Matrix Analysis and Applications*, 22(4):1136–1152, 2001.

[49] Ana Luísa Pinho, Alexis Amadon, Torsten Ruest, Murielle Fabre, Elvis Dohmatob, Isabelle Denghien, Chantal Ginisty, Séverine Becuwe-Desmidt, Séverine Roger, Laurence Laurier, et al. Individual brain charting, a high-resolution fMRI dataset for cognitive mapping. *Scientific data*, 5, 2018.

[50] Russell A Poldrack, Deanna M Barch, Jason Mitchell, Tor Wager, Anthony D Wagner, Joseph T Devlin, Chad Cumba, Oluwasanmi Koyejo, and Michael Milham. Toward open sharing of task-based fmri data: the openfmri project. *Frontiers in neuroinformatics*, 7:12, 2013.

[51] H. Richard, L. Gresele, A. Hyvarinen, B. Thirion, A. Gramfort, and P. Ablin. Modeling shared responses in neuroimaging studies through multiview ICA. In *Advances in Neural Information Processing Systems 33*, December 2020.

[52] Thomas J Rothenberg. Identification in parametric models. *Econometrica: Journal of the Econometric Society*, pages 577–591, 1971.

[53] Gilbert W Stewart. Error and perturbation bounds for subspaces associated with certain eigenvalue problems. *SIAM review*, 15(4):727–764, 1973.

[54] Markus Svensén, Frithjof Kruggel, and Habib Benali. ICA of fMRI group study data. *NeuroImage*, 16(3):551–563, 2002.

[55] François Tadel, Sylvain Baillet, John C Mosher, Dimitrios Pantazis, and Richard M Leahy. Brainstorm: a user-friendly application for meg/eeg analysis. *Computational intelligence and neuroscience*, 2011, 2011.

[56] Jason R Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A Shafto, Marie Dixon, Lorraine K Tyler, Richard N Henson, et al. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *Neuroimage*, 144:262–269, 2017.

[57] Javier S Turek, Cameron T Ellis, Lena J Skalaban, Nicholas B Turk-Browne, and Theodore L Willke. Capturing shared and individual information in fmri data. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 826–830. IEEE, 2018.

[58] Gaël Varoquaux, Sepideh Sadaghiani, Jean-Baptiste Poline, and Bertrand Thirion. CanICA: Model-based extraction of reproducible group-level ICA patterns from fMRI time series. *arXiv preprint arXiv:0911.4650*, 2009.

[59] Javier Vía, Matthew Anderson, Xi-Lin Li, and Tülay Adalı. Joint blind source separation from second-order statistics: Necessary and sufficient identifiability conditions. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2520–2523. IEEE, 2011.

[60] Javier Vía, Matthew Anderson, Xi-Lin Li, and Tülay Adalı. A maximum likelihood approach for independent vector analysis of gaussian data sets. In *2011 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2011.

[61] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[62] Daniela M Witten and Robert J Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology*, 8(1), 2009.

A Proofs and Lemmas

A.1 Proof of Theorem 1

Proof. By hypothesis, the covariances verify $C_{ij} = \mathbb{E}[\mathbf{x}_i \mathbf{x}_j^\top] = A_i(I_p + \delta_{ij}\Sigma_i)A_j^\top = A_i'(I_p + \delta_{ij}\Sigma_i')A_j'^\top$ for all i, j. We let $P_i = A_i^{-1}A_i'$. The previous relationship for $j \neq i$ gives $P_i P_j^\top = I_p$. Because there are more than 3 views, there is another integer $k \notin \{i, j\}$, and we have $P_i P_k^\top = P_j P_k^\top = I_p$. This shows that $P_i = P_j$: all these matrices are equal, and we call P their common value. The previous equation also gives $PP^\top = I_p$, so P is orthogonal. We have that $s + n_i$ and $s' + n_i'$ have independent components and $s + n_i = P(s' + n_i')$. Lemma 1 (a direct consequence of classical ICA results [18], Theorem 10) gives $P = \Pi^{-1}\Omega\Pi'$ where Π and Π' are sign and permutation matrices such that the first g components of $\Pi(s + n_i)$ and $\Pi'(s' + n_i')$ are Gaussian, and Ω is a block diagonal matrix given by

$$\Omega = \begin{bmatrix} \Omega_g & 0 \\ 0 & I_{p-g} \end{bmatrix}$$

where Ω_g is orthogonal. We call $A^{(g)}$ the first $g \times g$ block of a matrix A so that $\Omega^{(g)} = \Omega_g$.

Then, considering only the Gaussian components, we can write for $i = j$: $(\Pi\Sigma_i)^{(g)} = \Omega_g(\Pi'\Sigma_i')^{(g)}\Omega_g^\top$ for all i. This, combined with Assumption 1, implies that Ω_g is a sign and permutation matrix (see Lemma 2) and therefore P is a sign and permutation matrix. Then it follows that $I + \Sigma_i = P(I + \Sigma_i')P^\top$ and therefore $\Sigma_i = P\Sigma_i' P^\top$ so $\Sigma_i' = P^\top \Sigma_i P$. □

A.2 Proof of Theorem 2

Proof. Let us denote $W \in \mathbb{R}^{mp \times mp}$ the block diagonal matrix with block i given by $(A^i)^{-1}$. We have $\mathcal{C}\mathbf{u} = \lambda\mathcal{D}\mathbf{u} \iff W\mathcal{C}W^\top\mathbf{z} = \lambda W\mathcal{D}W^\top\mathbf{z}$ where $\mathbf{u} = W^\top\mathbf{z}$. We call $\mathbf{z}$ a reduced eigenvector. Each block in $W\mathcal{C}W^\top$ and in $W\mathcal{D}W^\top$ is diagonal so any reduced eigenvector $\mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_m \end{bmatrix}$ is such that the matrix $Z = [\mathbf{z}_1 \dots \mathbf{z}_m]$ has exactly one non-zero line. Following Lemma 3, the first p leading reduced eigenvectors $\mathbf{z}^1, \dots, \mathbf{z}^p$ all have different first non-zero coordinates. Therefore the concatenation of the first p leading reduced eigenvectors is given by $[\mathbf{z}^1, \dots \mathbf{z}^p] = \begin{bmatrix} \Gamma_1 \\ \vdots \\ \Gamma_m \end{bmatrix} P^\top$ where $P^\top \in \mathbb{R}^{p \times p}$ is a permutation matrix and $\Gamma_i \in \mathbb{R}^{p \times p}$ is a diagonal matrix. Therefore, the first p eigenvectors are given by $[\mathbf{u}^1 \dots \mathbf{u}^p] = \begin{bmatrix} W_1^\top \\ \vdots \\ W_m^\top \end{bmatrix} = \begin{bmatrix} (A_1^{-1})^\top\Gamma_1 P^\top \\ \vdots \\ (A_m^{-1})^\top\Gamma_m P^\top \end{bmatrix}$ and so $W_i = P\Gamma_i A_i^{-1}$ □

Lemma 1. *Let $\mathbf{s} \in \mathbb{R}^k$ and $\mathbf{s}' \in \mathbb{R}^k$ have independent components among which g are Gaussian, and P a rotation matrix such that $\mathbf{s} = P\mathbf{s}'$. Then, $P = \Pi^{-1}O\Pi'$ where Π and Π' are sign and permutation matrices such that the first g components of $\Pi\mathbf{s}$ and $\Pi'\mathbf{s}'$ are Gaussian and O is a block diagonal matrix such that $O^{(g)}$, the first $g \times g$ block of O, is orthogonal and the other block is identity.*

Proof. From [18], Theorem 10: Assume $\mathbf{s} = P\mathbf{s}'$, if the column j of P has more than one non-zero element then s_j' is Gaussian.

Let us define permutations Π_1, Π_1' such that the first g components of $\Pi_1\mathbf{s}$ and $\Pi_1'\mathbf{s}'$ are Gaussian and $P_1 = \Pi_1 P(\Pi_1')^{-1}$. We can see that P_1 is orthogonal.

We have $\Pi_1\mathbf{s} = P_1\Pi_1'\mathbf{s}'$. So the last $p - g$ columns of P_1 contain at most one non-zero element. Using orthogonality of P_1 this non-zero element has value 1 or -1 and is also the only one in its line. Let us focus on column $l > g$. Assume column l has its non-zero element at index $k \leq g$. Then line k in P_1 is only non-zero at index l and therefore $(\Pi_1\mathbf{s})_k$ (which is Gaussian) is equal to $(\Pi_1'\mathbf{s}')_l$ (which is not). Therefore column l can only have its non-zero element at an index greater than g.

This shows that P_1 is block diagonal $P_1 = \begin{bmatrix} O_g & 0 \\ 0 & P_2 \end{bmatrix}$ where O_g is orthogonal and P_2 is a sign and permutation matrix.

$$\begin{bmatrix} O_g & 0 \\ 0 & P_2 \end{bmatrix} = \Pi_1 P (\Pi_1')^{-1} \tag{7}$$

$$\iff \begin{bmatrix} O_g & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & P_2 \end{bmatrix} = \Pi_1 P (\Pi_1')^{-1} \tag{8}$$

$$\iff \Pi_1^{-1} \begin{bmatrix} O_g & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & P_2 \end{bmatrix} \Pi_1' = P \tag{9}$$

Therefore setting $\Pi' = \begin{bmatrix} I & 0 \\ 0 & P_2 \end{bmatrix} \Pi_1'$ and $\Pi = \Pi_1$ and $O = \begin{bmatrix} O_g & 0 \\ 0 & I \end{bmatrix}$ concludes the proof.

□

Lemma 2. *Assume that Assumption 2 holds for Σ_i, and that there is an orthogonal matrix P and diagonal matrices Σ_i' such that for all i, $\Sigma_i' = P\Sigma_i P^\top$. Then, P is a permutation matrix.*

Proof. The proof is in two parts. First, we show that there exist some coefficients $\alpha_1, \dots, \alpha_m$ such that the matrix $\sum_i \alpha_i \Sigma_i$ has distinct coefficients on the diagonal. Then, since we have $\sum_i \alpha_i \Sigma_i' = P\left(\sum_i \alpha_i \Sigma_i\right) P^\top$, and the diagonal $\sum_i \alpha_i \Sigma_i$ has distinct entries, we can invoke the unicity of the eigenvalue decomposition for symmetric matrices, which shows that P is necessarily a permutation matrix. Now, the only thing left is to prove is that Assumption 2 implies the existence of this linear combination.

We assume by contradiction that any linear combination of the Σ_i has two equal entries.

For $\alpha = [\alpha_1, \dots, \alpha_m]$, we let $\mathcal{S}(\alpha) = \operatorname{diag}(\sum_i \alpha_i \Sigma_i) \in \mathbb{R}^p$, where $\operatorname{diag}(\cdot)$ extracts the diagonal entries. The operator $\mathcal{S}$ is linear. We now define for $j, j' \leq p$ the linear form $\ell_{jj'}(\alpha) = \mathcal{S}(\alpha)_j - \mathcal{S}(\alpha)_{j'} \in \mathbb{R}$. The assumption on the linear combinations of Σ_i simply rewrites: For all $\alpha \in \mathbb{R}^m$, there exists $j, j' \leq p$ such that $\ell_{jj'}(\alpha) = 0$.

From a set point of view, this relationship writes

$$\bigcup_{j,j'} \operatorname{Ker}(\ell_{jj'}) = \mathbb{R}^m \ .$$

Since the $\ell_{jj'}$ are all linear forms, the $\operatorname{Ker}(\ell_{jj'})$ are subspaces of dimensions m or $m-1$, and since their union is of dimension m, there exists j, j' such that $\operatorname{Ker}(\ell_{jj'}) = \mathbb{R}^m$, i.e. such that $\ell_{jj'} = 0$.

As a consequence, we have for all α, $\mathcal{S}(\alpha)_j = \mathcal{S}(\alpha)_{j'}$. This implies that the sequences $(\Sigma_{ij})_i$ and $(\Sigma_{ij'})_i$ are equal, which contradicts Assumption 2.

We have therefore shown that Assumption 2 implies the existence of a linear combination of the Σ_i that has distinct entries, which concludes the proof. □

Lemma 3. *Let us consider the following eigenvalue problem:*

$$\begin{bmatrix} I+\Sigma_1 & I & \dots & I \\ I & I+\Sigma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & I \\ I & \dots & I & I+\Sigma_m \end{bmatrix} \mathbf{z} = \lambda \begin{bmatrix} I+\Sigma_1 & 0 & \dots & 0 \\ 0 & I+\Sigma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & I+\Sigma_m \end{bmatrix} \mathbf{z} \tag{10}$$

where $\forall i,\ 1 \leq i \leq m,\ \Sigma_m \in \mathbb{R}^{p,p}$ are positive diagonal matrices and I is the identity matrix. If the first p eigenvalues are distincts, the first p eigenvectors $\mathbf{z}^1, \dots, \mathbf{z}^p, \mathbf{z}^i \in \mathbb{R}^{mp}$ have different first non-zero coordinates.

Proof. We sort the eigenvectors in p groups of m vectors so that all vectors in group l have their l-th coordinate different from 0. Let $\mathbf{z}^{(l)}$ be an eigenvector in group l and let us call $\mathbf{w}_l \in \mathbb{R}^m$ the non-zero coordinates of this eigenvector: $\forall i \in \{1 \dots m\}, w_{li} = z^{(l)}_{l+(i-1)p}$.

We have:

$$\begin{bmatrix} 1+\Sigma_{1l} & 1 & \dots & 1 \\ 1 & 1+\Sigma_{2l} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \dots & 1 & 1+\Sigma_{ml} \end{bmatrix} \mathbf{w}_l = \begin{bmatrix} 1+\Sigma_{1l} & 0 & \dots & 0 \\ 0 & 1+\Sigma_{2l} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1+\Sigma_{ml} \end{bmatrix} \mathbf{w}_l \lambda_l \quad (11)$$

We now show that the biggest eigenvalue of (11) is strictly above 1 while all others are strictly below 1. The core of the proof comes from the study of the eigenvalues of a matrix modified by a rank 1 matrix. The reasoning we use here follows [25] (end of section 5).

Let us introduce $K^l = \mathrm{diag}(\Sigma_{1l} \dots \Sigma_{ml})$ and $\mathbf{u} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$. Let us drop the index l in the notations for simplicity.

The problem can be rewritten

$$(\mathbf{u}\mathbf{u}^\top + K)\mathbf{w} = (I+K)\mathbf{w}\lambda \quad (12)$$

$$\iff (I+K)^{-1}(\mathbf{u}\mathbf{u}^\top + K)\mathbf{w} = \mathbf{w}\lambda \quad (13)$$

The characteristic polynomial is given by:

$$\mathcal{P}(\lambda) = \det((I+K)^{-1}K - \lambda I + (I+K)^{-1}\mathbf{u}\mathbf{u}^\top) \quad (14)$$

$$\propto \det(I + ((I+K)^{-1}K - \lambda I)^{-1}(I+K)^{-1}\mathbf{u}\mathbf{u}^\top) \quad (15)$$

where we implicitly focus here on eigenvalues λ such that $\det((I+K)^{-1}K - \lambda I) \neq 0 \iff \forall i, \lambda \neq \frac{k_i}{1+k_i}$.

We then use the following property: Let $A \in \mathbb{R}^{a,b}$ and $B \in \mathbb{R}^{b,a}$ we have $\det(I_a + AB) = \det(I_b + BA)$.

Let us call $\chi(\lambda) = \det(I + ((I+K)^{-1}K - \lambda I)^{-1}(I+K)^{-1}\mathbf{u}\mathbf{u}^\top)$ we have:

$$\chi(\lambda) = 1 + \mathbf{u}^\top((I+K)^{-1}K - \lambda I)^{-1}(I+K)^{-1}\mathbf{u} \quad (16)$$

$$= 1 + \sum_{i=1}^{m} \frac{1}{1+k_i} \frac{1}{\frac{k_i}{1+k_i} - \lambda} \quad (17)$$

where $k_i = \Sigma_{il} > 0$. Taking the derivative we get

$$\chi'(\lambda) = \sum_{i=1}^{m} \frac{1}{1+k_i} \frac{1}{(\frac{k_i}{1+k_i} - \lambda)^2} > 0 \quad (18)$$

Trivially, $\forall i, \frac{k_i}{1+k_i} < 1$. We also have

$$\chi(1) = 1 + \sum_{i=1}^{m} \frac{1}{1+k_i} \frac{1}{\frac{k_i}{1+k_i} - 1} = 1 - m < 0 \quad (19)$$

and $\lim_{\lambda \to +\infty} \chi(\lambda) = 1$ so as χ is continuous and strictly increasing on $[1, +\infty[$. Therefore, it reaches 0 only once on this interval (excluding 1 since we know $\chi(1) \neq 0$). Therefore the greatest eigenvalue λ^* is strictly above 1 while all other eigenvalues are strictly below 1.

Note that because $\chi' > 0$, λ^* is of multiplicity 1. In the analysis above we ignored those eigenvalues λ such that $\lambda = \frac{k_i}{1+k_i}$ for some i. However since $\frac{k_i}{1+k_i} < 1$, none of these eigenvalues can be the largest one.

Finally, the p first eigenvectors belong to different groups (the corresponding eigenvalues are all strictly above 1). This shows that these eigenvectors have different first non-zero coordinates.

□

B Identifiability results for $m < 3$

We have a slightly weaker identifiability result when $m = 2$.

Proposition 5. *Let $m = 2$, and suppose that the scalars $(1 + \Sigma_{1j})(1 + \Sigma_{2j})$ for $j = 1 \dots p$ are all different. We let $\Theta' = (A'_1, A'_2, \Sigma'_1, \Sigma'_2)$ that also generates $\mathbf{x}_1, \mathbf{x}_2$. Then, there exists a permutation and scale matrix P such that $A'_1 = A_1 P$ and $A'_2 = A_2 P^{-\top}$.*

Proof. We let $P = A_1^{-1} A'_1$. Since $C_{12} = I_p$, it holds $A_2^{-1} A'_2 = P^{-\top}$. Then, we have $I_p + \Sigma_1 = P(I_p + \Sigma'_1)P^\top$. This means that there exists $U \in \mathcal{O}_p$ such that $P = (I_p + \Sigma_1)^{\frac{1}{2}} U (I_p + \Sigma'_1)^{-\frac{1}{2}}$. Since $P^{-\top}(I_p + \Sigma'_2)P^{-1} = I_p + \Sigma_2$, we find $U(I_p + \Sigma'_1)(I_p + \Sigma'_2)U^\top = (I_p + \Sigma_1)(I_p + \Sigma_2)$. By identification, U is a permutation matrix, and P is a scale and permutation matrix. □

As a consequence, when there are only two subjects, it is possible to recover the components and noise levels up to a scaling factor. When there is only one view, $m = 1$, there is a global rotation indeterminacy: $A_1(I_p + \Sigma_1)A_1^\top = A'_1(I_p + \Sigma_1){A'_1}^\top$ for $A'_1 = A_1(I_p + \Sigma_1)^{\frac{1}{2}} U (I_p + \Sigma_1)^{-\frac{1}{2}}$ where U is any orthogonal matrix. In this case, we lose identifiability.

C Derivation of fixed point updates for scalings

We want to minimize

$$L((\Phi_i)_{i=1}^m) = \sum_i \sum_{j \neq i} \|\Phi_i \mathrm{diag}(Y_{ij}) \Phi_j - I_p\|_F^2 \tag{20}$$

for Φ_i diagonal. With respect to each Φ_i, this function is strongly-convex, which means that the minimization w.r.t Φ_i can be done by cancelling the gradient. The gradient is given by

$$\frac{\partial L}{\partial \Phi_i} = 2 \sum_{j \neq i} (\Phi_i \mathrm{diag}(Y_{ij}) \Phi_j - I_p) \Phi_j \tag{21}$$

Therefore we get

$$\frac{\partial L}{\partial \Phi_i} = 0 \tag{22}$$

$$\iff 2 \sum_{j \neq i} (\Phi_i \mathrm{diag}(Y_{ij}) \Phi_j - I_p) \Phi_j = 0 \tag{23}$$

$$\iff \Phi_i \sum_{j \neq i} \mathrm{diag}(Y_{ij}) \Phi_j^2 - \sum_{j \neq i} \Phi_j = 0 \tag{24}$$

$$\iff \Phi_i = \frac{\sum_{j \neq i} \Phi_j}{\sum_{j \neq i} \mathrm{diag}(Y_{ij}) \Phi_j^2} \tag{25}$$

This update equation ensures that $\Phi_i = \arg\min_{\Phi_i} L((\Phi_j)_{i=1}^m)$, and we then loop through the Φ_i to get an alternate minimization scheme, which is guaranteed to converge to a stationary point of (20).

D EM E-step and M-step for ShICA with Gaussian components

D.1 E-step

The derivations are the same as in section E.1 but the sum over $\alpha \in \frac{1}{2}, \frac{3}{2}$ is replaced by just $\alpha = 1$.

D.2 M-step

The function to minimize in the M-step is then given by:

$$\mathcal{J} = -\log p(\mathbf{x}, \mathbf{s}) \tag{26}$$

$$= \sum_{i=1}^{m} \log(|\Sigma_i|) + \frac{1}{2} \operatorname{tr}(\Sigma_i^{-1} \left[(\mathbf{y}_i - \mathbb{E}[\mathbf{s}|\mathbf{x}])(\mathbf{y}_i - \mathbb{E}[\mathbf{s}|\mathbf{x}])^\top + \mathbb{V}[\mathbf{s}|\mathbf{x}] \right]) + c \tag{27}$$

where c does not depend on Σ_i

Therefore we get closed-form updates for Σ_i:

$$\Sigma_i \leftarrow \operatorname{diag}((\mathbf{y}_i - \mathbb{E}[\mathbf{s}|\mathbf{x}])(\mathbf{y}_i - \mathbb{E}[\mathbf{s}|\mathbf{x}])^\top + \mathbb{V}[\mathbf{s}|\mathbf{x}]) \tag{28}$$

Plugging in the closed-form formula for $\mathbb{E}[\mathbf{s}|\mathbf{x}]$ and $\mathbb{V}[\mathbf{s}|\mathbf{x}]$ we get updates that only depends on the covariances $\hat{C_{ij}} = \mathbb{E}[\mathbf{x}_i \mathbf{x}_j^\top]$.

$$\Sigma_i \leftarrow \operatorname{diag}(\hat{C_{ii}} - 2\mathbb{V}[\mathbf{s}|\mathbf{x}] \sum_{j=1}^{m} \Sigma_j^{-1} \hat{C}_{ji} + \mathbb{V}[\mathbf{s}|\mathbf{x}] \sum_{j=1}^{m} \sum_{l=1}^{m} \left(\Sigma_j^{-1} \hat{C}_{jl} \Sigma_l^{-1} \right) \mathbb{V}[\mathbf{s}|\mathbf{x}] + \mathbb{V}[\mathbf{s}|\mathbf{x}])$$

E EM E-step and M-step for ShICA with non-Gaussian components

E.1 E-step

The complete likelihood is given by

$$p(\mathbf{x}, \mathbf{s}) = \prod_i p(\mathbf{x}_i|\mathbf{s}) p(\mathbf{s}) \tag{29}$$

$$= \prod_i p(\mathbf{x}_i|\mathbf{s}) \prod_j \sum_{\alpha \in \{0.5, 1.5\}} p(s_j|\alpha) \tag{30}$$

$$\tag{31}$$

where

$$p(s_j|\alpha) = \mathcal{N}(s_j; 0, \alpha) \tag{32}$$

We have

$$p(\mathbf{x}_i|\mathbf{s}) = |W_i| \mathcal{N}(\mathbf{y}_i; \mathbf{s}, \Sigma_i) \tag{33}$$

$$= |W_i| \prod_j \mathcal{N}(y_{ij}; s_j, \Sigma_{ij}) \tag{34}$$

where Σ_{ij} is the coefficient j, j of Σ_i and $\mathbf{y}_i = W \mathbf{x}_i$.

Let us introduce a first lemma:

Lemma 4.

$$\prod_{i=1}^{m} \mathcal{N}(x_i; u, v_i) = \prod_{i=1}^{m} \mathcal{N}(x_i; \bar{x}, v_i) \sqrt{2\pi \bar{v}} \mathcal{N}(\bar{x}; u, \bar{v})$$

where $\bar{v} = (\sum_{i=1}^{m} v_i^{-1})^{-1}$ *and* $\bar{x} = \frac{\sum_i v_i^{-1} x_i}{\sum_i v_i^{-1}}$.

Proof. We have that

$$\sum_i \frac{1}{v_i} (x_i - u)^2 = \sum_i \frac{1}{v_i} (x_i - u)^2 \tag{35}$$

$$= \sum_i \frac{1}{v_i} (x_i - \bar{x} + \bar{x} - u)^2 \tag{36}$$

$$= \sum_i \frac{1}{v_i} (x_i - \bar{x})^2 + \sum_i \frac{1}{v_i} (\bar{x} - u)^2 \tag{37}$$

and therefore

$$\prod_i (\frac{1}{\sqrt{2\pi v_i}} \exp(-\frac{1}{2v_i}(x_i - \mu)^2)) \tag{38}$$

$$= \prod_i \frac{1}{\sqrt{2\pi v_i}} \exp(\sum_i -\frac{1}{2}(\frac{1}{v_i}(x_i - \bar{x})^2 + \frac{1}{v_i}(\bar{x} - u)^2)) \tag{39}$$

$$= \prod_i \mathcal{N}(x_i, \bar{x}, v_i) \exp(-\frac{1}{2}(\sum_i \frac{1}{v_i})(\bar{x} - u)^2)) \tag{40}$$

$$\tag{41}$$

so the desired result follow. □

By Lemma 4, we have

$$\prod_i p(\mathbf{x}_i|\mathbf{s}) = \prod_i |W_i| \prod_j \mathcal{N}(y_{ij}; \bar{y}_j, \Sigma_{ij}) \sqrt{2\pi\bar{\Sigma_j}} \mathcal{N}(\bar{y}_j; s_j, \bar{\Sigma_j}) \tag{42}$$

$$\tag{43}$$

where $\bar{y}_j = \frac{\sum_i \Sigma_{ij}^{-1} y_{ij}}{\sum_i \Sigma_{ij}^{-1}}$ and $\bar{\Sigma_j} = (\sum_i \Sigma_{ij}^{-1})^{-1}$. Hiding variable that do not depend on $\mathbf{s}$ we obtain

$$\prod_i p(\mathbf{x}_i|\mathbf{s}) \propto \prod_j \mathcal{N}(\bar{y}_j; s_j, \bar{\Sigma_j}) \tag{44}$$

$$\tag{45}$$

Then we get

$$p(\mathbf{x}, \mathbf{s}) \propto \prod_j \sum_{\alpha \in \{0.5, 1.5\}} \mathcal{N}(s_j; \bar{y}_j, \bar{\Sigma_j}) \mathcal{N}(s_j; 0, \alpha) \tag{46}$$

Let us now prove a second Lemma:

Lemma 5.

$$\mathcal{N}(x; y, \nu)\mathcal{N}(x, 0, \alpha) = \mathcal{N}(y; 0, \nu + \alpha)\mathcal{N}(x; \frac{\alpha y}{\alpha + \nu}, \frac{\nu\alpha}{\alpha + \nu})$$

Proof. We have

$$\mathcal{N}(x; y, \nu)\mathcal{N}(x, 0, \alpha) = \frac{\exp\left(-\frac{(x-y)^2}{2\nu}\right)}{\sqrt{2\pi\nu}} \frac{\exp\left(-\frac{x^2}{2\alpha}\right)}{\sqrt{2\pi\alpha}} \tag{47}$$

Then,

$$\exp\left(-\frac{(x-y)^2}{2\nu}\right) \tag{48}$$

$$= \exp\left(-\frac{\alpha(x-y)^2+\nu x^2}{2\alpha\nu}\right) \tag{49}$$

$$= \exp\left(-\frac{\alpha(x^2-2xy+y^2)+\nu x^2}{2\alpha\nu}\right) \tag{50}$$

$$= \exp\left(-\frac{x^2(\alpha+\nu)-2x(\alpha y)+\alpha y^2}{2\alpha\nu}\right) \tag{51}$$

$$= \exp\left(-\frac{x^2-2x\frac{\alpha y}{\alpha+\nu}+\frac{\alpha y^2}{\alpha+\nu}}{2\frac{\alpha\nu}{\alpha+\nu}}\right) \tag{52}$$

$$= \exp\left(-\frac{(x-\frac{\alpha y}{\alpha+\nu})^2-(\frac{\alpha y}{\alpha+\nu})^2+\frac{\alpha y^2}{\alpha+\nu}}{2\frac{\alpha\nu}{\alpha+\nu}}\right) \tag{53}$$

$$= \exp\left(-\frac{(x-\frac{\alpha y}{\alpha+\nu})^2}{2\frac{\alpha\nu}{\alpha+\nu}}\right)\exp\left(-\frac{-\alpha^2y^2+(\alpha+\nu)\alpha y^2}{2\alpha\nu(\alpha+\nu)}\right) \tag{54}$$

$$= \exp\left(-\frac{(x-\frac{\alpha y}{\alpha+\nu})^2}{2\frac{\alpha\nu}{\alpha+\nu}}\right)\exp\left(-\frac{\nu\alpha y^2}{2\alpha\nu(\alpha+\nu)}\right) \tag{55}$$

and

$$\frac{1}{\sqrt{2\pi\nu}\sqrt{2\pi\alpha}} = \frac{1}{\sqrt{2\pi(\nu+\alpha)}\sqrt{2\pi\frac{\nu\alpha}{\nu+\alpha}}} \tag{56}$$

so that the desired result follow. □

By Lemma 5, we have:

$$p(\mathbf{x},\mathbf{s}) \tag{57}$$

$$\propto \prod_j \sum_{\alpha\in\{0.5,1.5\}} \mathcal{N}(\bar{y}_j;0,\bar{\Sigma}_j+\alpha)\mathcal{N}(s_j;\frac{\alpha\bar{y}_j}{\alpha+\bar{\Sigma}_j},\frac{\bar{\Sigma}_j\alpha}{\alpha+\bar{\Sigma}_j}) \tag{58}$$

and therefore we get:

$$p(\mathbf{s}|\mathbf{x}) = \frac{p(\mathbf{s},\mathbf{x})}{\int_{\mathbf{s}} p(\mathbf{s},\mathbf{x})} \tag{59}$$

$$= \prod_j \frac{\sum_{\alpha\in\{0.5,1.5\}}\theta_\alpha\mathcal{N}(s_j;\frac{\alpha\bar{y}_j}{\alpha+\bar{\Sigma}_j},\frac{\bar{\Sigma}_j\alpha}{\alpha+\bar{\Sigma}_j})}{\sum_{\alpha\in\{0.5,1.5\}}\theta_\alpha} \tag{60}$$

where $\theta_\alpha = \mathcal{N}(\bar{y}_j;0,\bar{\Sigma}_j+\alpha)$.

So we obtain the desired result:

$$\mathbb{E}[s_j|\mathbf{x}] = \frac{\sum_{\alpha\in\{0.5,1.5\}}\theta_\alpha\frac{\alpha\bar{y}_j}{\alpha+\bar{\Sigma}_j}}{\sum_{\alpha\in\{0.5,1.5\}}\theta_\alpha} \tag{61}$$

$$\mathbb{V}[s_j|\mathbf{x}] = \frac{\sum_{\alpha\in\{0.5,1.5\}}\theta_\alpha\frac{\bar{\Sigma}_j\alpha}{\alpha+\bar{\Sigma}_j}}{\sum_{\alpha\in\{0.5,1.5\}}\theta_\alpha} \tag{62}$$

E.2 M-step

The function to minimize in the M-step is then given by:

$$\mathcal{J} = -\log p(\mathbf{x}, \mathbf{s}) \tag{63}$$

$$= \sum_{i=1}^{m} -\log(|W_i|) + \log(|\Sigma_i|) + \frac{1}{2}\operatorname{tr}(\Sigma_i^{-1}\left[(\mathbf{y}_i - \mathbb{E}[\mathbf{s}|\mathbf{x}])(\mathbf{y}_i - \mathbb{E}[\mathbf{s}|\mathbf{x}])^\top + \mathbb{V}[\mathbf{s}|\mathbf{x}]\right]) + c \tag{64}$$

where c does not depend on Σ_i or W_i

Therefore we get closed-form updates for Σ_i:

$$\Sigma_i \leftarrow \operatorname{diag}((\mathbf{y}_i - \mathbb{E}[\mathbf{s}|\mathbf{x}])(\mathbf{y}_i - \mathbb{E}[\mathbf{s}|\mathbf{x}])^\top + \mathbb{V}[\mathbf{s}|\mathbf{x}]) \tag{65}$$

We update W_i by performing a quasi-Newton step.

We use the relative gradient $\mathcal{G}^{W_i}$ and $\mathcal{H}^{W_i}$ defined by
$\mathcal{J}(W_i + \varepsilon W_i) = \mathcal{J}(W_i) + \langle \varepsilon | \mathcal{G}^{W_i} \rangle + \frac{1}{2}\langle \varepsilon | \mathcal{H}^{W_i} \varepsilon \rangle$.

We get:

$$\mathcal{J}(W_i + \varepsilon W_i) = \sum_{i=1}^{m} \left[-\log(|W_i|) - \log(|I_k + \varepsilon|) - \log(\mathcal{N}(\mathbf{y}_i + \varepsilon \mathbf{y}^i; \mathbf{s}; \Sigma_i))\right] + const \tag{66}$$

$$= \mathcal{J}(W_i) - \operatorname{tr}(\varepsilon) + \frac{1}{2}\operatorname{tr}(\varepsilon^2) \tag{67}$$

$$+ \frac{1}{2}\left[\langle \varepsilon \mathbf{y}^i | (\Sigma_i)^{-1}(\mathbf{y}^i - \mathbf{s})\rangle + \langle (\mathbf{y}^i - \mathbf{s}) | (\Sigma_i)^{-1}\varepsilon \mathbf{y}^i\rangle + \langle \varepsilon \mathbf{y}^i | (\Sigma_i)^{-1}\varepsilon \mathbf{y}^i\rangle\right] \tag{68}$$

$$+ o(\|\varepsilon\|^2) \tag{69}$$

$$= \mathcal{J}(W_i) - \sum_a \varepsilon_{a,a} + \frac{1}{2}\sum_{a,b} \varepsilon_{a,b}\varepsilon_{b,a} \tag{70}$$

$$+ \sum_{a,b} \varepsilon_{a,b}\left[(\Sigma_i)^{-1}(\mathbf{y}^i - \mathbf{s})(\mathbf{y}^i)^\top\right]_{a,b} + \frac{1}{2}\sum_{a,b} \varepsilon_{a,b}\left[(\Sigma_i)^{-1}\varepsilon \mathbf{y}^i(\mathbf{y}^i)^\top\right]_{a,b} \tag{71}$$

$$+ o(\|\varepsilon\|^2) \tag{72}$$

$$= \mathcal{J}(W_i) - \sum_a \varepsilon_{a,a} + \frac{1}{2}\sum_{a,b} \varepsilon_{a,b}\varepsilon_{b,a} \tag{73}$$

$$+ \sum_{a,b} \varepsilon_{a,b}\left[(\Sigma_i)^{-1}(\mathbf{y}^i - \mathbf{s})(\mathbf{y}^i)^\top\right]_{a,b} + \frac{1}{2}\sum_{a,b,d} \varepsilon_{a,b}(\Sigma_i)^{-1}_{a,a}\varepsilon_{a,d}\left[\mathbf{y}^i(\mathbf{y}^i)^\top\right]_{d,b} \tag{74}$$

$$+ o(\|\varepsilon\|^2) \tag{75}$$

$$\tag{76}$$

So:

$$\mathcal{G}^{W_i}_{a,b} = -\delta_{a,b} + \left[(\Sigma_i)^{-1}(\mathbf{y}^i - \mathbf{s})(\mathbf{y}^i)^\top\right]_{a,b} \tag{77}$$

and

$$\mathcal{H}^{W_i}_{a,b,c,d} = \delta_{a,d}\delta_{b,c} + \delta_{a,c}\frac{y_{ib}y_{id}}{\Sigma_{ia}} \tag{78}$$

We approximate the Hessian by

$$\widehat{\mathcal{H}^{W_i}_{a,b,c,d}} = \delta_{ad}\delta_{bc} + \delta_{ac}\delta_{bd}\frac{(y_{ib})^2}{\Sigma_{ia}} \tag{79}$$

where the Hessian approximation is exact when the unmixed data have truly independent components.

Updates for W_i are then given by $W_i \leftarrow (I - \rho(\widehat{\mathcal{H}^{W_i}})^{-1}\mathcal{G}^{W_i})W_i$, where ρ is chosen by backtracking line-search. We alternate between computing the statistics $\mathbb{E}[\mathbf{s}|\mathbf{x}]$ and $\operatorname{Var}[\mathbf{s}|\mathbf{x}]$ (E-step) and updates of parameters Σ_i and W_i for $i = 1 \ldots m$ (M-step).

Dataset	Duration	m	Description
Sherlock	50 min	16	Movie watching (BBC TV show "Sherlock")
Forrest	110 min	19	Auditory version "Forrest Gump"
Gallant	130 min	12	various short video clips
Raiders	110 min	11	Movie watching ("Raiders of the lost ark")

Table 1: Information about datasets (name, duration, number of subjects m and short description)

F Description of the datasets and the preprocessing pipeline

All datasets are resampled and masked using the brain mask available at `http://cogspaces.github.io/assets/data/hcp_mask.nii.gz`. The dimensionality of the data is given by the number of voxels in the mask: 212445. Data are detrended and standardized so that each voxels' timecourse has zero mean and unit variance.

When reconstructing the BOLD signal of missing subjects, data are preprocessed with a 6 mm smoothing. In the timesegment matching experiment, we use unsmoothed data except for the sherlock dataset for which the available data are already smoothed. Multiple acquisitions (called runs) are necessary to build the datasets. Each run lasts approximately 10 minutes.

Sherlock data are available at `http://arks.princeton.edu/ark:/88435/dsp01nz8062179`. We refer the reader to [15] for a precise description of the study cohort, experimental design and pre-processing pipeline. The data are split manually into 4 runs of 395 timeframes and one run of 396 timeframes so that cross validation can be performed. Subject 5 is removed because of missing data. The repetition time (TR) is 1.5s and the spatial resolution is of 3 mm.

Forrest data are downloaded from OpenfMRI [50]. Data are acquired with a 7T scanner with an isotropic spatial resolution of 1 mm and then resampled to a spatial resolution of 3 mm. A complete description of the experimental design and study cohort are given in `http://studyforrest.org` and [28]. Subject 10 is discarded as not all runs are available at the time of writing. Run 8 is discarded as it is missing in some subjects. We therefore uses 7 runs of respectively 451, 441, 438, 488, 462, 439 and 542 timeframes and 19 subjects. The repetition time (TR) is 2s and the spatial resolution is of 1 mm.

Raiders and Gallant dataset pertains to the Individual Brain Charting dataset. These data were acquired using a 3T scanner and resampled to an isotropic spatial resolution of 3 mm. More information is available in [49]. Gallant dataset is refered to as clips in [49]. Data are available at `https://openneuro.org/datasets/ds00268`. Datasets gallant and raiders are preprocessed using FSL `http://fsl.fmrib.ox.ac.uk/fsl` using slice time correction, spatial realignment, co-registration to the T1 image and affine transformation of the functional volumes to a template brain (MNI). The repetition time (TR) is 2s and the spatial resolution is of 3 mm. The Raiders dataset uses 9 runs of respectively 374, 297, 314, 379, 347, 346, 350, 353 and 211 timeframes. The Gallant dataset uses 17 runs of 325 timeframes each. The protocol used for Raiders is the same as the one used in [30] and the protocol used for Gallant is the same as the one used in [43].

A brief summary of the characteristics of the datasets is available in Table 1

All datasets used in MEG have dimensionality 102 since we only consider the magnetometers. The temporal resolution is 1 ms.

The *CamCAN* dataset [56] contains the MEG data of 496 different subjects exposed to an audio-visual stimuli. More precisely, subjects are presented simultaneously an auditory stimuli lasting 300ms at frequency 300, 600 or 1200 Hz and a checkerboard pattern lasting 34ms. 120 trials are available. The protocol used in the CamCAN MEG dataset is described in [56].

G Reconstructing the BOLD signal of missing subjects

We report in Table 2 the R2 score obtained with MVICA, ShICA-J and ShICA-ML with 20 components as well as a 95% confidence interval on the experiment "Reconstructing the fMRI data of left-out subjects". These data are already reported in Figure 4 but are given here in form of a table.

Dataset	Method	R^2 score	Confidence interval
forrest	ShICA-ML	0.200	[0.187, 0.213]
	ShICA-J	0.171	[0.157, 0.185]
	MVICA	0.191	[0.177, 0.204]
gallant	ShICA-ML	0.121	[0.107, 0.135]
	ShICA-J	0.110	[0.095, 0.125]
	MVICA	0.114	[0.099, 0.128]
raiders	ShICA-ML	0.158	[0.142, 0.174]
	ShICA-J	0.146	[0.129, 0.162]
	MVICA	0.144	[0.124, 0.164]
sherlock	ShICA-ML	0.174	[0.157, 0.191]
	ShICA-J	0.165	[0.146, 0.183]
	MVICA	0.161	[0.142, 0.180]

Table 2: **Reconstructing the BOLD signal of missing subjects**. Median R^2 score and 95% confidence interval.

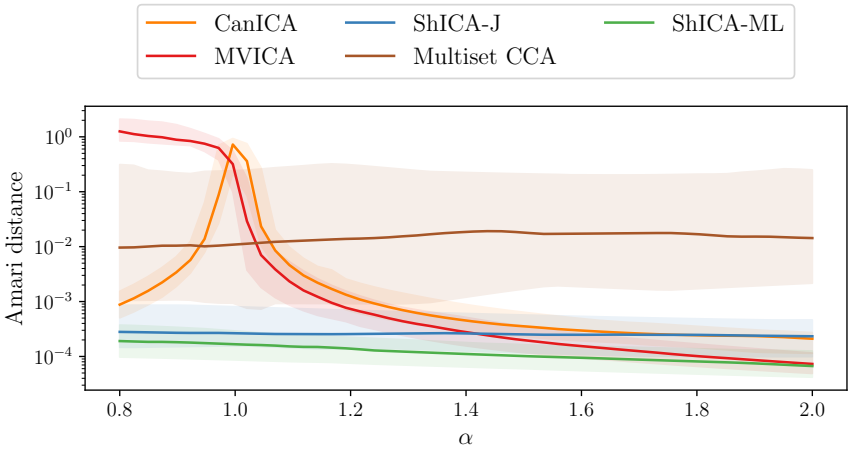


Figure 5: **Separation performance in function of non-Gaussianity** Separation performance of algorithms for sub-Gaussian $\alpha < 1$ and super-Gaussian $\alpha > 1$ components

H Additional experiments

H.1 Separation performance

H.1.1 Separation performance in function of non-Gaussianity

We generate data according to model (1). Components $\mathbf{s}$ are generated using $s_j = d(x)$ with $d(x) = x|x|^{\alpha-1}$ and $x \sim \mathcal{N}(0, 1)$. Mixing matrices A_i are generated by sampling their coefficients from a standardized Gaussian law. The number of samples is fixed to $n = 10^5$ and we vary α between 0.8 and 1.2. Each experiment is repeated 40 times using different seeds in the random number generator. We use $p = 4$ components and $m = 5$ views. We display in Fig 5 the mean Amari distance across subjects. The experiment is repeated 100 times using different seeds. We report the median result and error bars represent the first and last deciles. When α is close to 1 (components are almost Gaussian), ShICA-J, ShICA-ML and multiset CCA can separate components well (but multiset CCA reaches higher amari distance than ShICA). In this regime, MVICA yields much higher amari distance than ShICA-J, ShICA-ML or Multiset CCA but is still better than CanICA which cannot separate components at all. As non-Gaussianity (α) increases, ICA based methods yield better results but ShICA-ML yields uniformly lower amari distance.

H.2 fMRI timesegment matching experiment

We benchmark ShICA on four different real fMRI datasets via a timesegment matching experiment similar to the one in [16]. We use full brain data. The datasets and the preprocessing pipeline are

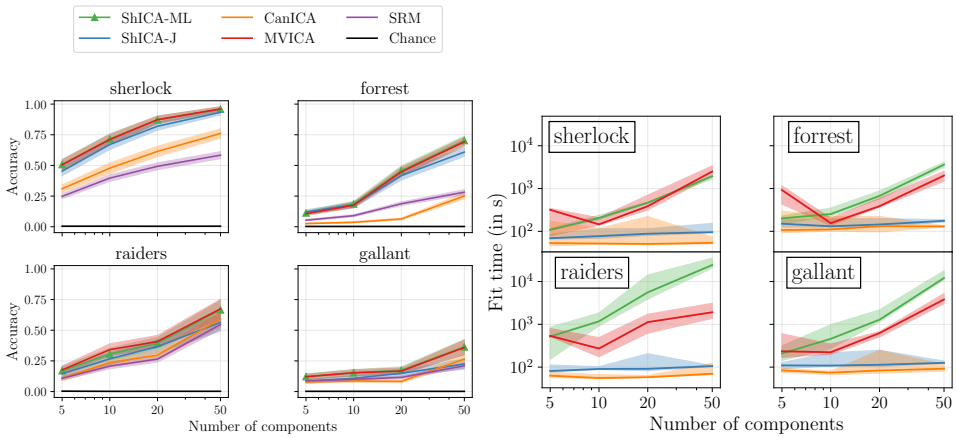


Figure 6: **Timesegment matching experiment**: (left) Accuracy (right) Fitting time (in seconds)

described in Appendix F. We split the data into a train and test set and algorithms are fitted on the train set. On the test set, we estimate the shared components from all subjects but one and select a target timesegment containing 9 consecutive samples in the shared components. We try to localize this timesegment from the components of the left-out subject using a maximum correlation classifier (all possible windows of 9 consecutive timeframes are considered in the left-out subject excluding the ones partially overlapping with the correct timesegment). The left panel in Fig 6 shows that ShICA-ML, MVICA and ShICA-J yield almost equal accuracy and outperform other methods by a large margin. The right panel in Fig 6 shows that ShICA-J is much faster to fit than MVICA or ShICA-ML.

We would like to highlight here that these experiments are not exactly the same as in [16] as we use full brain data and they use regions of interest. The code used for this experiment is very similar to the tutorial in `https://brainiak.org/tutorials/11-SRM/`. We use the SRM implementation in Brainiak [35]. Also note that the Raiders dataset is different from the one used in [16] as it involves different subjects and data were acquired in a different neuro-imaging center.

H.3 MEG Phantom experiment

H.3.1 Phantom Elektra

Dipoles in $m = 32$ various locations are emitting the same signal. Signal magnitude can be either very high, high or low, leading to 3 datasets: a very clean one, a clean one and a noisy one. These datasets are available as part of the Brainstorm application [55]. We preprocess the data using Maxwell filtering and low-pass filtering as done in the MNE tutorial `https://mne.tools/0.17/auto_tutorials/plot_brainstorm_phantom_elekta.html` and only consider data recorded by the magnetometers. We use the very clean dataset to recover the true signal by PCA with 1 component. Then we reduce the noisy dataset by applying view-specific PCA with $k = 20$ components and algorithms are applied on the reduced data. We select the component that is closer to the true one and compute the L2 norm between the predicted component and the true one after normalization. Then we attempt to recover the position of each dipole by performing dipole fitting on the mixing operator of each view (using only the column corresponding to the true component). The localization error is defined as the mean l2 distance between the true localization and the predicted localization where the mean is computed across dipoles. Each epoch corresponds to 301 samples and 20 epochs are available in total. We vary the number of epochs between 2 and 18 and display in Fig 7 the reconstruction error and the localization error in function of the number of epochs used. ShICA-ML outperforms other methods. ShICA-J gives satisfying results while being much faster.

H.3.2 Phantom Sinusoidal components

For completeness, we display the results obtained on another MEG dataset where the true component is a known sinusoidal and $m = 8$ different locations are considered for the dipoles. We vary the number of epochs between 2 and 16 and display in Fig 8 the reconstruction error and the localization

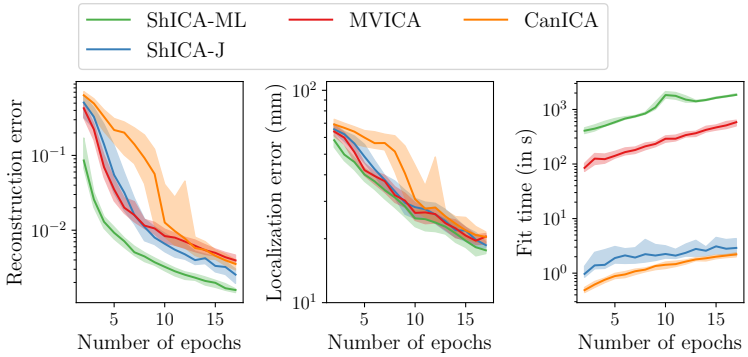


Figure 7: **MEG Phantom (Elektra)**: (left) L2 distance between the predicted and actual component (middle) Mean error (in mm) between predicted and actual dipoles localization (right) Fitting time (in seconds)

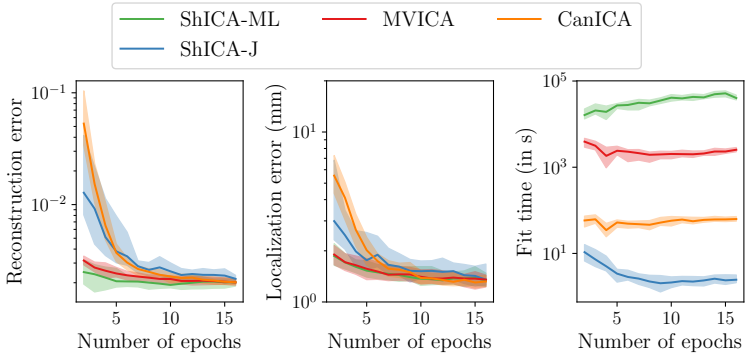
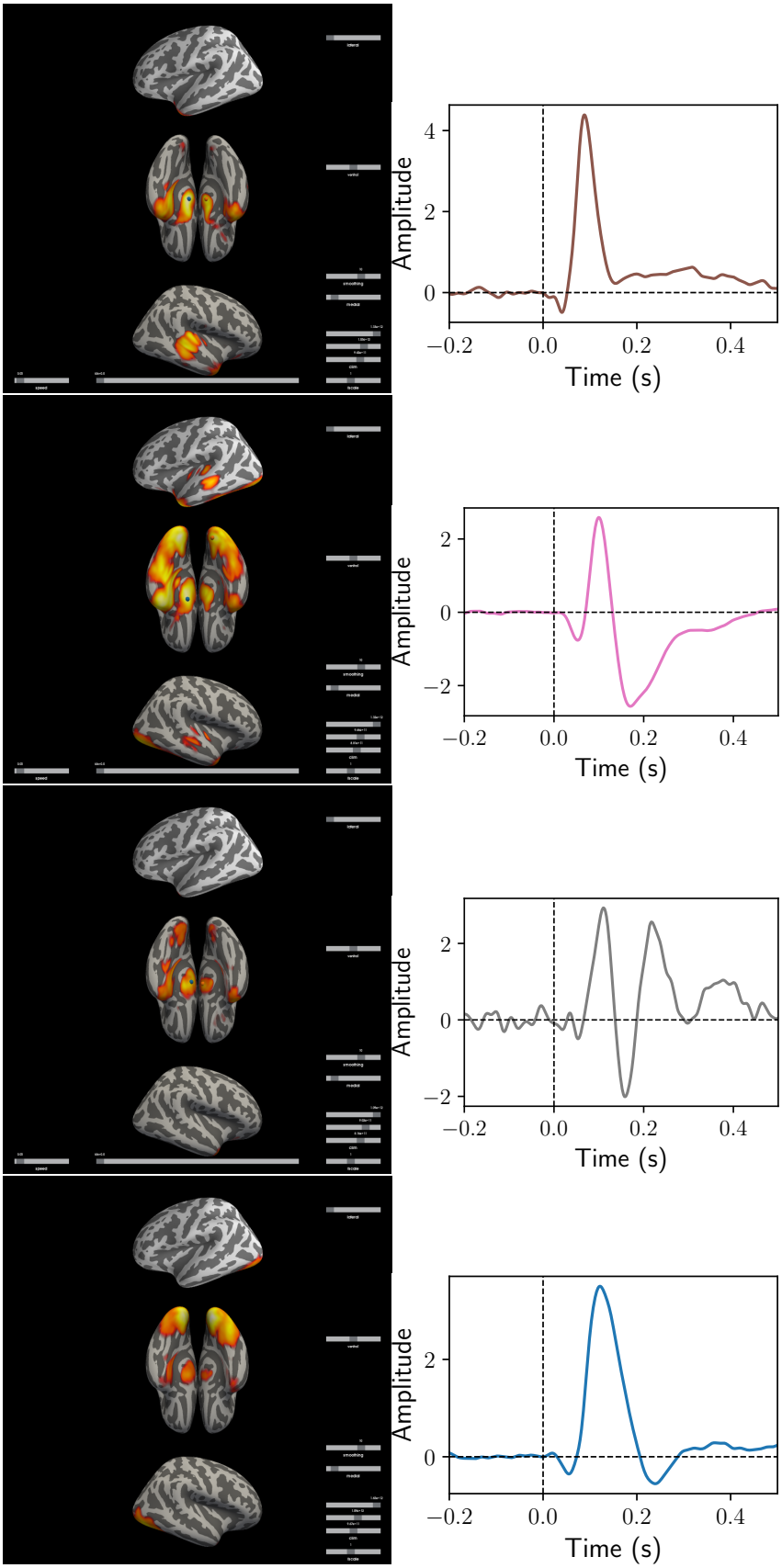


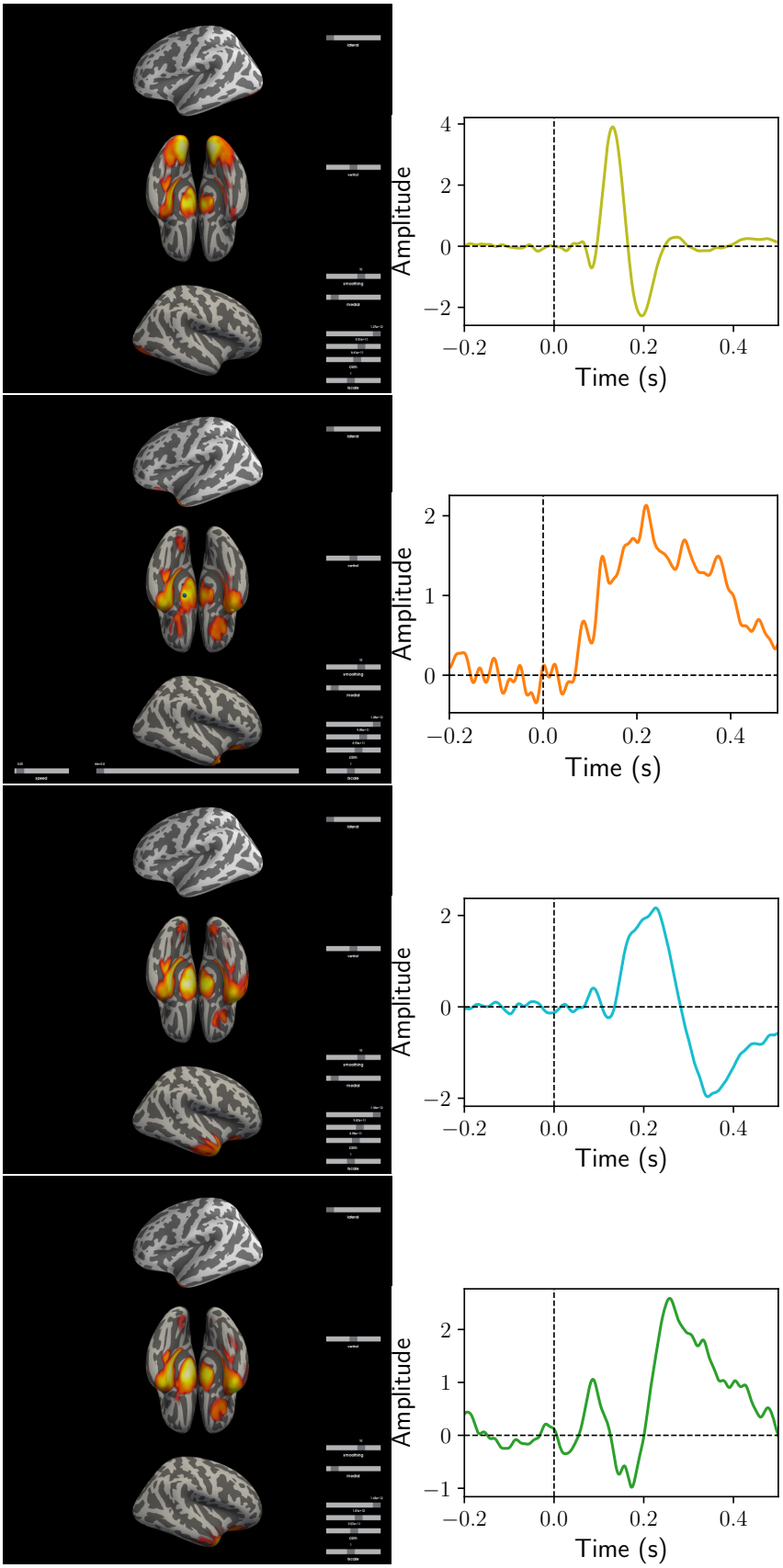
Figure 8: **MEG Phantom Sinusoidal components**: (left) L2 distance between the predicted and actual component (middle) Mean error (in mm) between predicted and actual dipoles localization (right) Fitting time (in seconds)

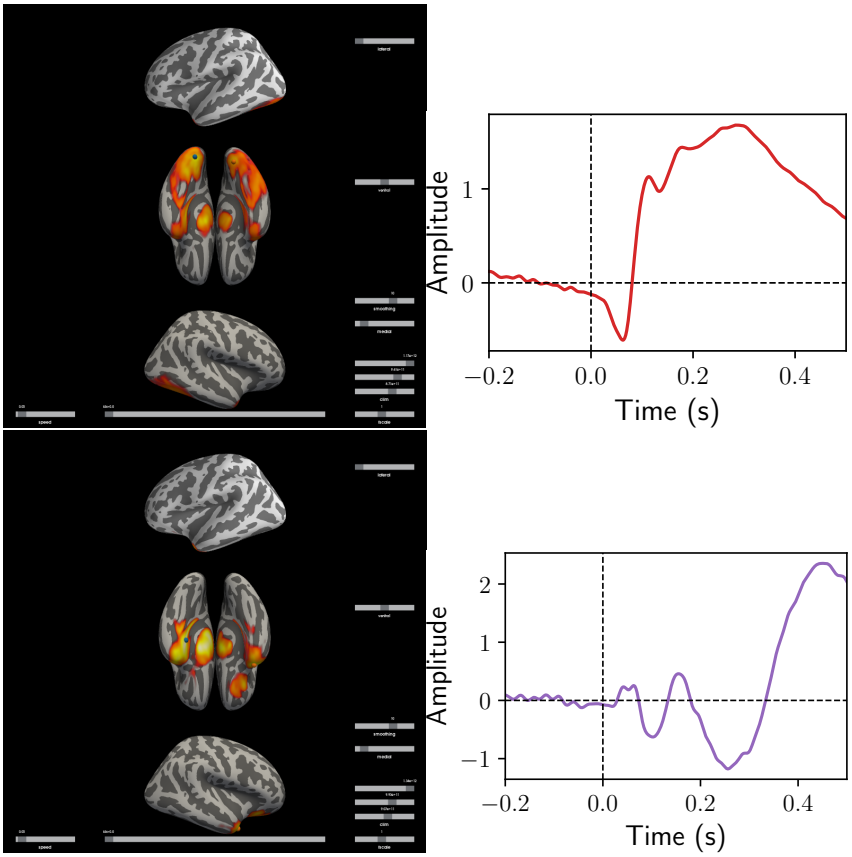
error as a function of the number of epochs used. ShICA-ML outperforms other methods. ShICA-J gives satisfying results while being much faster.

H.4 CamCAN MEG components

We consider the CamCAN dataset used to produce Fig 4. We use $m = 496$ subjects and fit ShICA-ML with $p = 10$ components. We localize the components of each subject using sLoreta [44]. Then components are registered to a common brain and averaged. Thresholded maps are displayed below along with the time courses of each component. Components obtained with ShICA-ML highlight the ventral visual cortex and auditive cortex. The results suggest that the response of the auditive cortex is faster and lasts a shorter time than the response of the ventral visual cortex.


Amplitude
Time (s)
4
2
0
−0.2
0.0
0.2
0.4
Amplitude
Time (s)
2
0
−2
−0.2
0.0
0.2
0.4
Amplitude
Time (s)
2
0
−2
−0.2
0.0
0.2
0.4
Amplitude
Time (s)
2
0
−0.2
0.0
0.2
0.4


Amplitude
Time (s)
Amplitude
Time (s)
Amplitude
Time (s)
Amplitude
Time (s)


Amplitude
1
0
−0.2
0.0
0.2
0.4
Time (s)
Amplitude
2
1
0
−1
−0.2
0.0
0.2
0.4
Time (s)

References