Shared Independent Component Analysis for Multi-Subject Neuroimaging

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

NeurIPS, 2021

Github: @hugorichard

Twitter: @hugorichard

https://hugorichard.github.io

Sources and sensors

Mixing

Independent component analysis (noise-free)

ICA model (Jutten, 1991)

- Independent *sources*: $\mathbf{s} \in \mathbb{R}^k$

$p(\mathbf{s}) = p(s_1) \cdots p(s_k)$

- *Sensors*: $\mathbf{x} \in \mathbb{R}^k$

$$\mathbf{x} = A\mathbf{s}$$

where A is the *Mixing matrix*.

Independent component analysis (noise-free)

ICA model (Jutten, 1991)

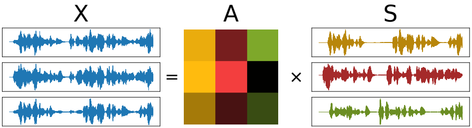
- Independent *sources*: $\mathbf{s} \in \mathbb{R}^k$

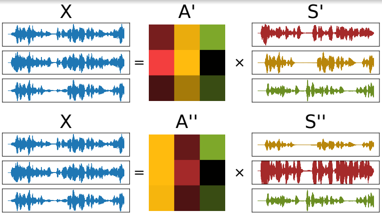
$$p(\mathbf{s}) = p(s_1) \cdots p(s_k)$$

- *Sensors*: $\mathbf{x} \in \mathbb{R}^k$

$$\mathbf{x} = A\mathbf{s}$$

where A is the *Mixing matrix*.





Independent component analysis (noise-free)

ICA model (Jutten, 1991)

- Independent *sources*: $\mathbf{s} \in \mathbb{R}^k$

$p(\mathbf{s}) = p(s_1) \cdots p(s_k)$

- *Sensors*: $\mathbf{x} \in \mathbb{R}^k$

$$\mathbf{x} = A\mathbf{s}$$

where A is the *Mixing matrix*.

Theorem (Identifiability of ICA (Common, 1994))

If $\mathbf{x} = A\mathbf{s}$ and $\mathbf{x} = A'\mathbf{s}'$ and if $\mathbf{s}$ has at most one Gaussian component, Then

- $A = PA'$
- *P is a scale and permutation matrix.*

Group ICA

Generalization to multiple subjects exposed to the same stimuli

Consider 2 subjects $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^k$ such that

- $\mathbf{x}_1 = A_1\mathbf{s} + \mathbf{n}_1$
- $\mathbf{x}_2 = A_2\mathbf{s} + \mathbf{n}_2$

Interpretation

- Shared sources $\mathbf{s}$: shared cognitive processes
- Different mixing matrices A_i: different spatial topography of each subject
- Different noises $\mathbf{n}_i$: inter-subject variability.

State of the art

ConcatICA [Calhoun, 2001]

$\mathbf{x}_1 \in \mathbb{R}^k, \mathbf{x}_2 \in \mathbb{R}^k$

- PCA of $\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}$

$\mathbf{x} = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} \mathbf{x}_{red}$, where $\mathbf{x}_{red} \in \mathbb{R}^k$ and $\begin{bmatrix} U_1 \\ U_2 \end{bmatrix}$ is orthogonal.

- ICA of reduced data $\mathbf{x}_{red} = A\mathbf{s}$

CanICA [Varoquaux, 2010]

Replace PCA with (multi-set) CCA in ConcatICA

CCA solves: $\begin{bmatrix} \mathbb{E}[\mathbf{x}_1 \mathbf{x}_1^\top] & \mathbb{E}[\mathbf{x}_1 \mathbf{x}_2^\top] \\ \mathbb{E}[\mathbf{x}_2 \mathbf{x}_1^\top] & \mathbb{E}[\mathbf{x}_2 \mathbf{x}_2^\top] \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} = \lambda \begin{bmatrix} \mathbb{E}[\mathbf{x}_1 \mathbf{x}_1^\top] & 0 \\ 0 & \mathbb{E}[\mathbf{x}_2 \mathbf{x}_2^\top] \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}$

State of the art

About CanICA and ConcatICA

- Very fast to fit
- Simple to implement
- Do not optimize a proper likelihood
- Not even clear what is the underlying model

Some other related work

- IVA [Lee, 2008]
- Unified approach [Guo, 2008]
- SRM [Chen, 2015]
- MultiViewICA [Richard, 2020]

Noisy ICA

Example (Noisy ICA likelihood with Gaussian mixtures (Bermond, Cardoso, 1999))

Our model:

- $\mathbf{x}_i = A_i\mathbf{s} + \mathbf{n}_i, \mathbf{n}_i \sim N(0, \sigma^2 I_k)$ $I_k \in \mathbb{R}^{k,k}$ is the identity matrix.
- $p(s_j) = \frac{1}{q}\sum_{\alpha_j \in \mathcal{A}} \mathcal{N}(s_j; 0, \alpha_j)$, $\mathcal{A} \in (\mathbb{R}_+^*)^q$

Solving via EM

E-step:

- $p(\mathbf{s}|\mathbf{x}) = \sum_{\boldsymbol{\alpha}, \alpha_j \in \mathcal{A}} p(\mathbf{s}|\mathbf{x}, \boldsymbol{\alpha})p(\boldsymbol{\alpha}|\mathbf{x})$, with $\mathbf{x} = \mathbf{x}_1, \cdots, \mathbf{x}_n$ and $\boldsymbol{\alpha} = \alpha_1, \cdots, \alpha_k$

But $\underline{\{\boldsymbol{\alpha}, \alpha_j \in \mathcal{A}\} \text{ has size } q^k}$ making the E-step intractable.

Our contribution: Shared ICA (ShICA)

ShICA model

$$\mathbf{x}_i = A_i(\mathbf{s} + \mathbf{n}_i), i = 1, \ldots, m$$

- $\mathbf{n}_i \sim \mathcal{N}(0, \Sigma_i)$ where Σ_i is diagonal positive.
- $\mathbf{s}$ are independent components some of which may be Gaussian
- $\mathbb{E}[\mathbf{x}_i] = 0$, A_i invertible, $\mathbb{E}[\mathbf{s}\mathbf{s}^\top] = I_k$ and $m \geq 3$

ShICA-J:

- In theory Multiset CCA solves ShICA (under some conditions).
- In practice, sampling noise causes some issues.
- Joint diagonalization solves it: ShICA-J = MCCA + Joint diag

ShICA-ML

- A maximum likelihood approach to ShICA

ShICA is identifiable

Definition (Noise diversity in Gaussian components)

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

Theorem (Identifiability)

Assuming noise diversity, let $\Theta = (A_1, \dots, A_m, \Sigma_1, \dots, \Sigma_m)$ be the set of parameter that generates $\mathbf{x}_1, \dots, \mathbf{x}_m$ from the ShICA model. We let $\Theta' = (A'_1, \dots, A'_m, \Sigma'_1, \dots, \Sigma'_m)$ another set of parameters, and assume that they also generate the data. Then, there exists a sign and permutation matrix P such that for all i, $A'_i = A_i P$, and $\Sigma'_i = P^\top \Sigma_i P$.

Note that noise diversity in Gaussian component is also a necessary condition.

Multiset CCA solves GroupICA

Theorem (Solving GroupICA with Multiset CCA)

We assume $\mathbf{x}_i$ *follows* $\mathbf{x}_i = A_i(\mathbf{s} + \mathbf{n}_i)$ *where* $\mathbf{n}_i \sim \mathcal{N}(0, \Sigma_i)$ *where* Σ_i *is diagonal and consider the multiset CCA problem*

$C\mathbf{u} = \lambda D\mathbf{u}$

where block i, j *of* C *is* $\mathbb{E}[\mathbf{x}_i \mathbf{x}_j^\top]$ *and* D *is block diagonal with block* i, i *given by* $\mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\top]$*. Let* $U = [\mathbf{u}_1 \dots \mathbf{u}_k] = \begin{bmatrix} W_1^\top \\ \vdots \\ W_m^\top \end{bmatrix}$ *where* $W_i \in \mathbb{R}^{k,k}$*.*

Then $\underline{\text{if } \lambda_1 \dots \lambda_k \text{ are distincts}}$*,* $W_i = P\Gamma_i A_i^{-1}$ *where* P *is a permutation matrix and* Γ_i *a scaling matrix.*

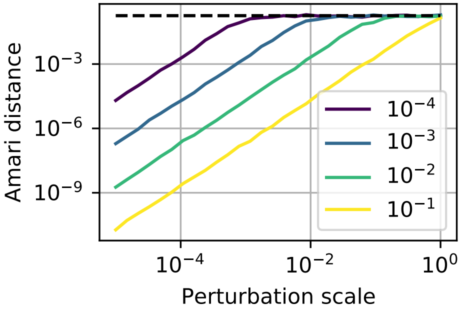
Note that the distinct eigenvalue condition is also necessary.
Note that the condition is stronger than noise diversity (we can exhibit an identifiable example on which MCCA fails).

Practical issues with Multiset CCA

The mapping from matrices to eigenvectors is highly non smooth...

Practical example

$m = 3$, $k = 2$ and Σ_i such that $\lambda_1 = 2 + \epsilon$ and $\lambda_2 = 2$.
W_i: Solution of multiset CCA on true covariance matrices C_{ij}
$\tilde{W}_i$: Solution of multiset CCA on perturbed covariance matrices $\tilde{C}_{ij} = C_{ij} + \delta S$ where S positive symmetric matrix of norm 1.



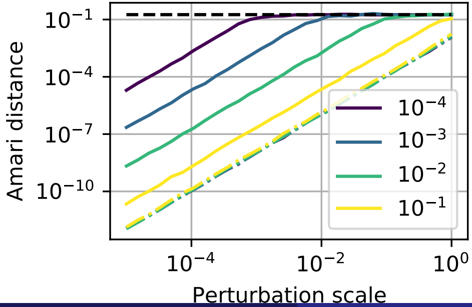
Solving practical issues with joint diagonalization

Large gap between the first k eigenvalues and others

$$\lambda_k - \lambda_{k+1} > \frac{m-1}{1+\max_{ij}\Sigma_{ij}} + \frac{1}{1+\min_{ij}\Sigma_{ij}}$$

Practical implications

The span of the p leading eigenvectors is preserved: $W_i \approx Q\tilde{W}_i$. We recover Q by joint diagonalization of $Q\tilde{W}_i\frac{1}{n}X_iX_i^\top \tilde{W}_i^\top Q^\top$



ShICA-J

Use Multiset CCA and joint diagonalization to obtain W_i up to a scaling Ψ_i.

Find the scalings

We solve $\min_\Psi \sum_{i \neq j} \|\Psi_i \mathrm{diag}(Q\tilde{W}_i \tilde{C}_{ij} \tilde{W}_j Q^\top)\Psi_j - I_k\|^2$. The estimates of W_i are then given by $\hat{W}_i = \Psi_i Q \tilde{W}_i$.

Find the noise variances

We use the maximum likelihood estimate of $\mathbf{x}_i = \hat{W}_i^{-1}(\mathbf{s} + \mathbf{n}_i)$ via an EM algorithm. The E-step and M-step are in closed form yielding a fast algorithm.

ShICA-J is very fast. But it is not a maximum likelihood estimator.

ShICA-ML: the maximum likelihood estimator

ShICA-ML

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

where $\mathbf{n}_i \sim \mathcal{N}(0, \Sigma_i)$, Σ_i diagonal and $s_j \sim \frac{1}{2}\sum_{\alpha \in \{\frac{1}{2}, \frac{3}{2}\}} \mathcal{N}(0, \alpha)$.

Optimization

Optimized via an EM algorithm.

$$\mathbb{E}[s_j|\mathbf{x}] = \frac{\sum_{\alpha \in \{\frac{1}{2}, \frac{3}{2}\}} \theta_\alpha \frac{\alpha \bar{y}_j}{\alpha + \bar{\Sigma}_j}}{\sum_{\alpha \in \{\frac{1}{2}, \frac{3}{2}\}} \theta_\alpha}, \qquad \mathbb{V}[s_j|\mathbf{x}] = \frac{\sum_{\alpha \in \{\frac{1}{2}, \frac{3}{2}\}} \theta_\alpha \frac{\bar{\Sigma}_j \alpha}{\alpha + \bar{\Sigma}_j}}{\sum_{\alpha \in \{\frac{1}{2}, \frac{3}{2}\}} \theta_\alpha}$$

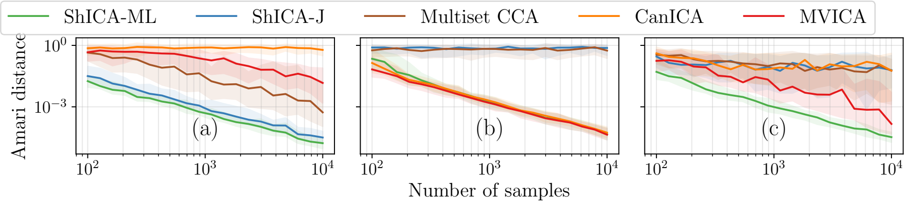
where $\theta_\alpha = \mathcal{N}(\bar{y}_j; 0, \bar{\Sigma}_j + \alpha)$, $\bar{y}_j = \frac{\sum_i \Sigma_{ij}^{-1} y_{ij}}{\sum_i \Sigma_{ij}^{-1}}$ and $\bar{\Sigma}_j = (\sum_i \Sigma_{ij}^{-1})^{-1}$ with $\mathbf{y}_i = W_i \mathbf{x}_i$. M-step: Closed form updates for noise variances and quasi-newton updates for unmixing matrices.

ShICA-J provides a great initialization to ShICA-ML

Synthetic experiments

Separation performance depending on the density of sources

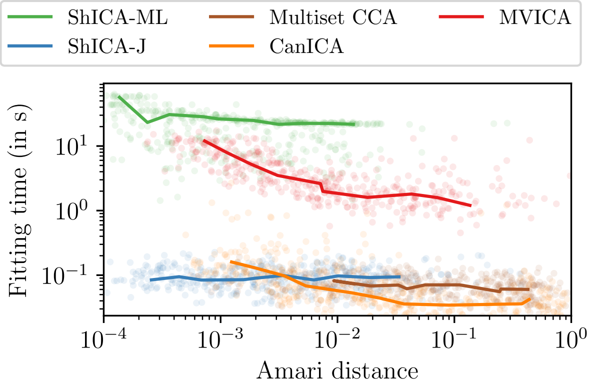
$m = 4$ views, $k = 5$ components, non-Gaussian sources are from a Laplace density, we use the ShICA model using:
(a) Gaussian components with noise diversity
(b) non-Gaussian components without noise diveristy
(c) Half of components are Gaussian with noise diversity, the other half is non-Gaussian without diversity



Synthetic experiments

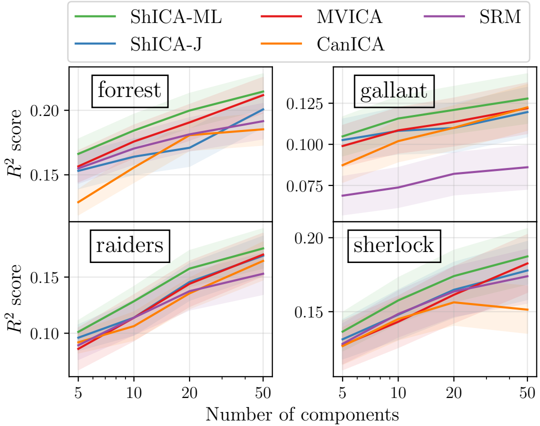
Computation time

We generate components from a slightly super-Gaussian density $s_j = d(x)$ with $d(x) = x|x|^{0.2}$ and $x \sim \mathcal{N}(0, 1)$ vary the number of samples $n = 10^2 \dots 10^4$.



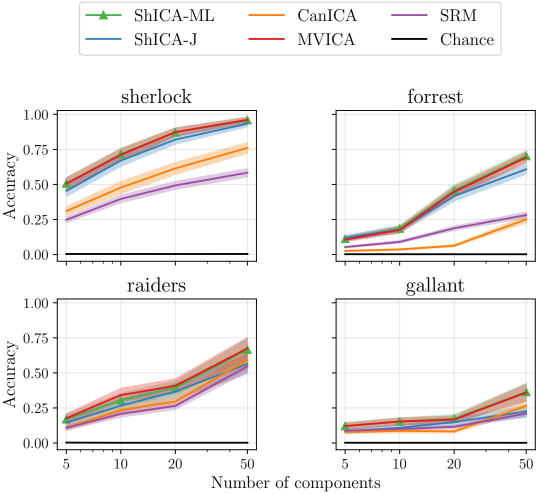
Reconstruction experiment fMRI

- Train data: 100% subjects 80% runs -> Learn unmixing matrices
- Test data: 80% subjects 20% runs -> Compute sources
- Validation data: 20% subjects 20% runs -> Measure R2 score



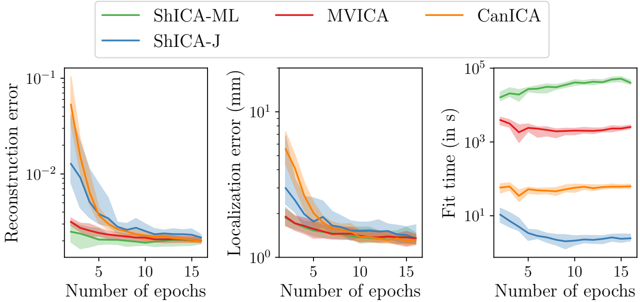
Timesegment matching fMRI

- Timesegment matching accuracy: Locate a 9 timeframes timesegment in a left out subject by correlation with the average response of other subjects.
- Train data: 80% runs $\rightarrow$ Learn unmixing matrices
- Test data: 20% runs $\rightarrow$ Measure accuracy



MEG Phantom experiment

- 8 dipoles in a plastic head at different locations
- Dipoles separately emit the same known signal S_{true} during n epochs
- 20 sources estimated: the best one is compared with S_{true}



Conclusion

Take home message

- ShICA is a powerful framework to extract shared sources
- ShICA-J yields a fast approach but only uses second order information, ShICA-ML is a bit slower but uses in addition non-Gaussianity.
- Yields better results in practice: extensive comparison on multiple neuroscience modalities.

Future Work

- These methods work on reduced data. How to provide the best dimension reduction method ?
- The non-Gaussian density of the shared sources in ShICA-ML could be learned.