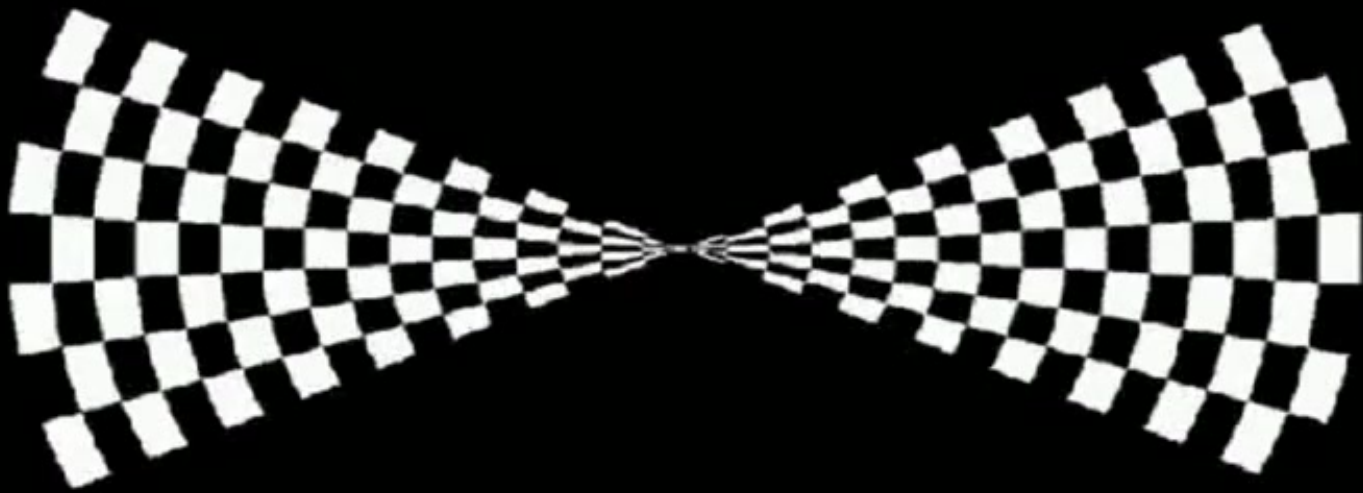
Unsupervised component analysis for neuroimaging data

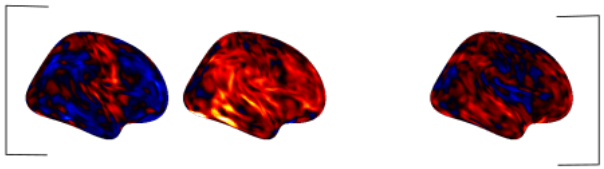
Hugo Richard

PHD supervisor	Prof. Bertrand Thirion	*Inria, Université Paris-Saclay*
Reviewers	Prof. Tülay Adali	*University of Maryland Baltimore County*
	Prof. Moritz Grosse-Wentrup	*University of Vienna*
Examinators	Prof. Christian Jutten	*Université Grenoble Alpes*
	Prof. Sylvain Chevalier	*Université de Versailles Saint-Quentin*
	Prof. Matthieu Kowalski	*Université Paris Sud*
	Prof. Aapo Hyvärinen	*University of Helsinki*
	Prof. Alexandre Gramfort	*Inria, Université Paris-Saclay*


MAGNETOM Prisma
A Tim+Dot System

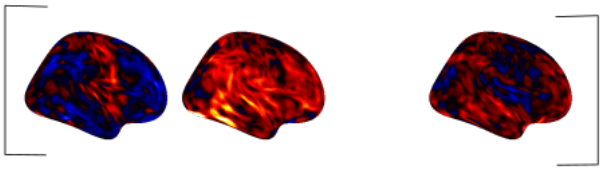


General linear model

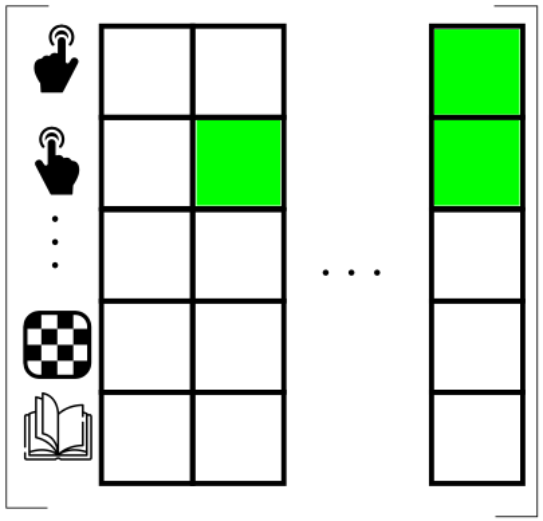


$$X \in \mathbb{R}^{v,n}$$

General linear model



$$X \in \mathbb{R}^{v,n}$$



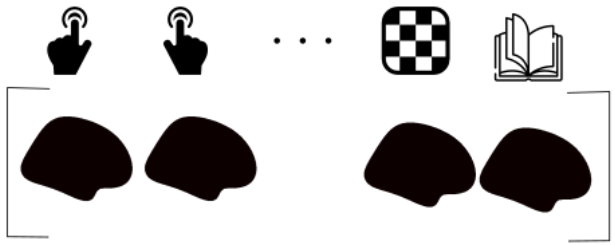
$$S \in \mathbb{R}^{p,n}$$

General linear model

$X \in \mathbb{R}^{v,n}$ $A \in \mathbb{R}^{v,p}$ $S \in \mathbb{R}^{p,n}$

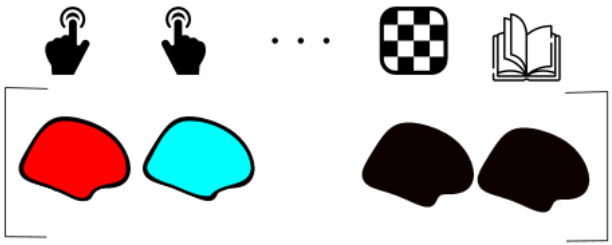
$+$ noise

General linear model



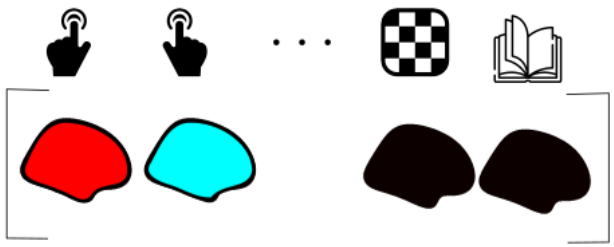
$$A \in \mathbb{R}^{v,p}$$

General linear model

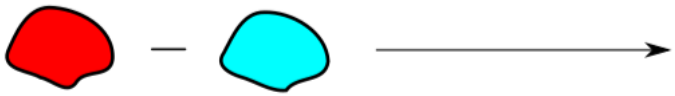


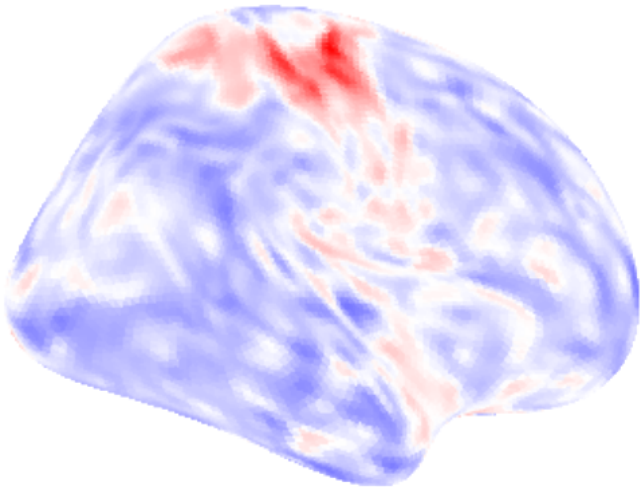
$$A \in \mathbb{R}^{v,p}$$

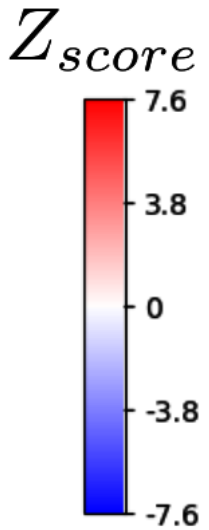
General linear model



$$A \in \mathbb{R}^{v,p}$$







General linear model

One subject

$$X = AS + \text{noise}$$

General linear model

One subject

$$X = AS + \text{noise}$$

m subjects

$$X_1 = A_1 S + \text{noise} \quad \cdots \quad X_m = A_m S + \text{noise}$$

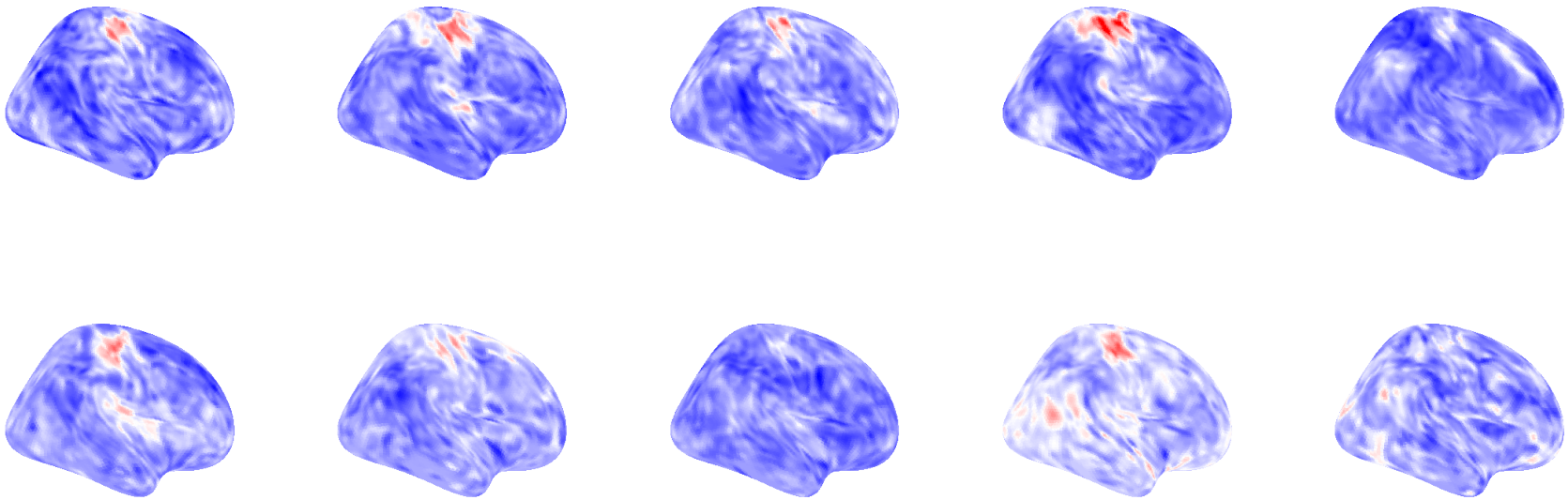
General linear model

One subject

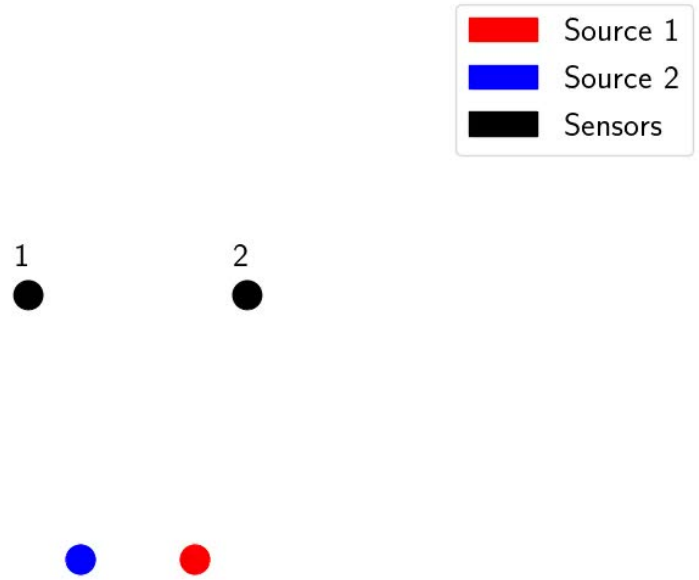
$$X = AS + \text{noise}$$

m subjects

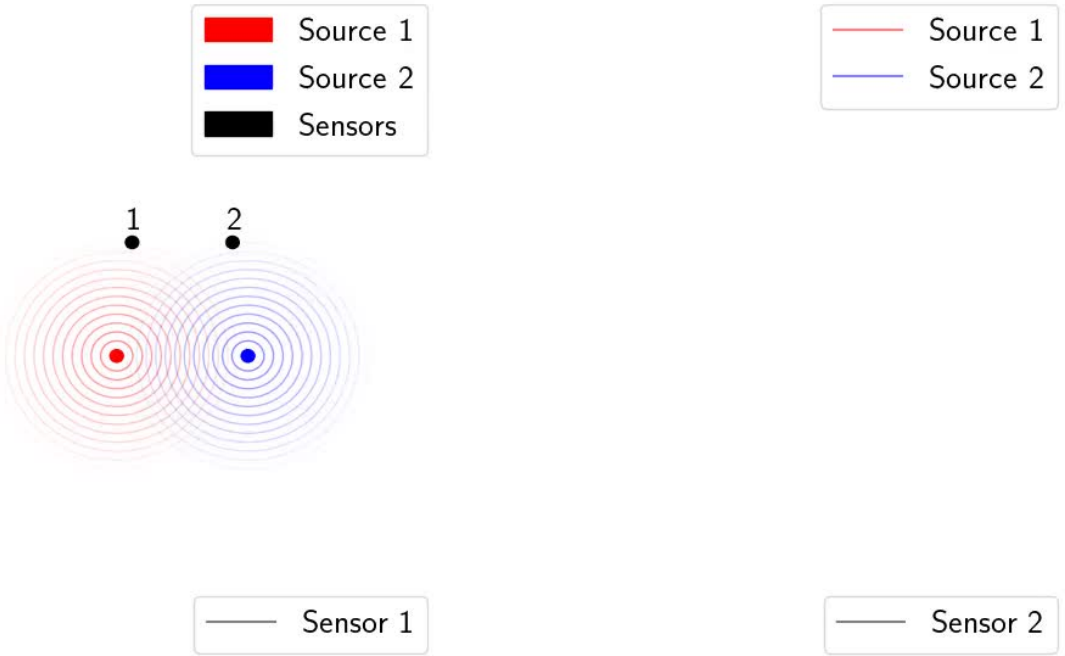
$$X_1 = A_1 S + \text{noise} \quad \cdots \quad X_m = A_m S + \text{noise}$$



Sources and sensors



Mixing


Source 1
Source 2
Sensors
1
2
Source 1
Source 2
Sensor 1
Sensor 2

Independent component analysis (noise-free)

ICA model (Jutten, 1991)

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

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

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

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

where A is the *Mixing matrix*.

Independent component analysis (noise-free)

ICA model (Jutten, 1991)

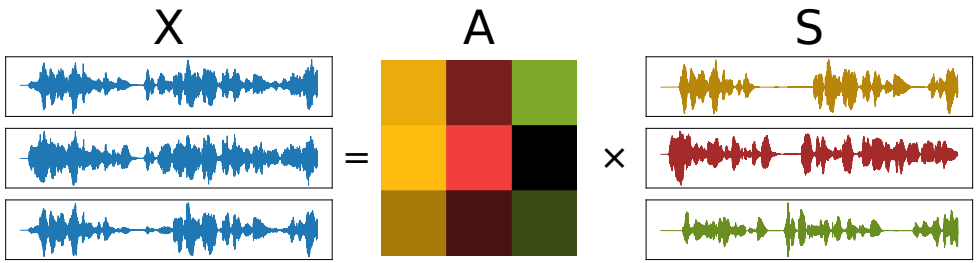
- Independent *sources*: $\mathbf{s} \in \mathbb{R}^p$

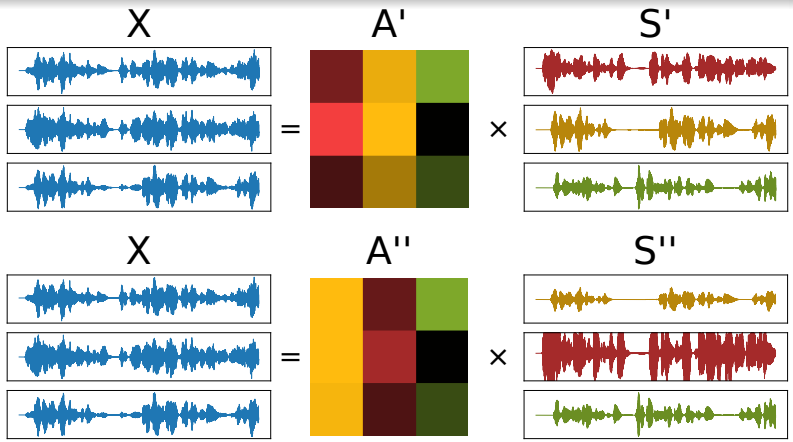
$$p(\mathbf{s}) = p(s_1) \cdots p(s_p)$$

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

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

where A is the *Mixing matrix*.





Independent component analysis (noise-free)

ICA model (Jutten, 1991)

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

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

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

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

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.*

Modeling complex stimuli

ICA

$$X = AS$$

S "as independent as possible"

Modeling complex stimuli

ICA

$$X = AS$$

S "as independent as possible"

Interpretation

- $S \in \mathbb{R}^{p,n}$: independent brain processes
- $A \in \mathbb{R}^{v,p}$: corresponding spatial maps

Modeling complex stimuli

ICA

$$X = AS$$

S "as independent as possible"

Interpretation

- $S \in \mathbb{R}^{p,n}$: independent brain processes
- $A \in \mathbb{R}^{v,p}$: corresponding spatial maps

Dimension reduction

- In ICA, we assume $p = v$.
- So we need to reduce the dimensionality of the data.

Group ICA

m subjects exposed to the same stimuli

Consider m subjects $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^k$ such that $\forall i \in \{1, \ldots, m\}$:

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

Interpretation

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

State of the art

ConcatICA [Calhoun, Adali, 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 $\mathbf{x}_{red} = A\mathbf{s}$

State of the art

ConcatICA [Calhoun, Adali, 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

State of the art

Typical noisy ICA likelihood [Moulines, Cardoso 1997]

- $\mathbf{x}_i = A_i\mathbf{s} + \mathbf{n}_i, \mathbf{n}_i \sim N(0, \sigma^2 I_p)$ $I_p \in \mathbb{R}^{p,p}$ is the identity matrix.
- $p(\mathbf{s}) = d(s_1) \cdots d(s_p)$

  Denoting $A = \begin{bmatrix} A_1 \\ \vdots \\ A_n \end{bmatrix}$, $\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{bmatrix}$, we have

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{s})p(\mathbf{s})d\mathbf{s} \tag{1}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \int e^{-\frac{1}{2\sigma^2}||\mathbf{x}-A\mathbf{s}||_2^2} p(\mathbf{s})d\mathbf{s} \tag{2}$$

- Intractable in general (it does not factorize)

State of the art

A Gaussian mixture model [Moulines, Cardoso 1997]

- $\mathbf{x}_i = A_i \mathbf{s} + \mathbf{n}_i, \mathbf{n}_i \sim N(0, \sigma^2 I_k)$
- $p(s_j) = \frac{1}{q} \sum_{\alpha_j \in \mathcal{A}} p(s_j | \alpha_j)$
- $p(s_j | \alpha_j) = \mathcal{N}(s_j; 0, \alpha_j)$

where $\mathcal{A}$ is a known finite set.

State of the art

A Gaussian mixture model [Moulines, Cardoso 1997]

- $\mathbf{x}_i = A_i \mathbf{s} + \mathbf{n}_i, \mathbf{n}_i \sim N(0, \sigma^2 I_k)$
- $p(s_j) = \frac{1}{q} \sum_{\alpha_j \in \mathcal{A}} p(s_j | \alpha_j)$
- $p(s_j | \alpha_j) = \mathcal{N}(s_j; 0, \alpha_j)$

where $\mathcal{A}$ is a known finite set.

Can we do an EM ?

$p(\mathbf{s}|\mathbf{x}, \boldsymbol{\alpha}) = \mathcal{N}(\mathbf{s}, \mu_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}})$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$ and $V_{\boldsymbol{\alpha}} = (\frac{1}{\sigma^2} A^\top A + \mathrm{diag}(\boldsymbol{\alpha})^{-1})^{-1}$, $\mu_{\boldsymbol{\alpha}} = \frac{1}{\sigma^2} V_{\boldsymbol{\alpha}} A^\top \mathbf{x}$.

State of the art

A Gaussian mixture model [Moulines, Cardoso 1997]

- $\mathbf{x}_i = A_i\mathbf{s} + \mathbf{n}_i, \mathbf{n}_i \sim N(0, \sigma^2 I_k)$
- $p(s_j) = \frac{1}{q}\sum_{\alpha_j \in \mathcal{A}} p(s_j|\alpha_j)$
- $p(s_j|\alpha_j) = \mathcal{N}(s_j; 0, \alpha_j)$

where $\mathcal{A}$ is a known finite set.

Can we do an EM ?

$p(\mathbf{s}|\mathbf{x}, \boldsymbol{\alpha}) = \mathcal{N}(\mathbf{s}, \mu_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}})$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$
and $V_{\boldsymbol{\alpha}} = (\frac{1}{\sigma^2}A^\top A + \mathrm{diag}(\boldsymbol{\alpha})^{-1})^{-1}$, $\mu_{\boldsymbol{\alpha}} = \frac{1}{\sigma^2}V_{\boldsymbol{\alpha}}A^\top\mathbf{x}$.

No we can't

$p(\mathbf{s}|\mathbf{x}) = \sum_{\boldsymbol{\alpha}\in\mathcal{A}^p} \mathcal{N}(\mathbf{s}, \mu_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}})$
It is a sum with q^p terms ($q = 2, 3$, $p = 20, 50$).

State of the art

The model [Guo, 2008]

- $\mathbf{x}_i = A_i \mathbf{s} + \mathbf{n}_i$
- $p(s_j) = \sum_{z=1}^{q} \mathcal{N}(\mu_{zj}, \sigma_{zj})$
- μ_{zj}, σ_{zj} must be learned.

It is even more difficult !

State of the art

Independent vector analysis [Lee, 2008]

- $\mathbf{x}_i = A_i \mathbf{s}_i$
- $\mathbf{s}_{[j]} = (s_{ij})_{i=1}^m$ are not independent

State of the art

Independent vector analysis [Lee, 2008]

- $\mathbf{x}_i = A_i \mathbf{s}_i$
- $\mathbf{s}_{[j]} = (s_{ij})_{i=1}^{m}$ are not independent

Advantages

- A closed form likelihood
- Takes naturally into account inter-subject variability

State of the art

Independent vector analysis [Lee, 2008]

- $\mathbf{x}_i = A_i \mathbf{s}_i$
- $\mathbf{s}_{[j]} = (s_{ij})_{i=1}^m$ are not independent

Advantages

- A closed form likelihood
- Takes naturally into account inter-subject variability

Some instances of IVA

- IVA-L [Lee, 2008]: $p_{\mathbf{s}_{[j]}}(\mathbf{y}_{[j]}) \propto \exp(-\sqrt{\sum_i (y_{ij})^2})$
- IVA-G [Anderson, 2011] [Via, 2011]: $p_{\mathbf{s}_{[j]}}(\mathbf{y}_{[j]}) = \mathcal{N}(\mathbf{y}_{[j]}; 0, \Sigma_j)$
- Many others: IVA-L-SOS [Bhinge, 2019]

State of the art

Independent vector analysis [Lee, 2008]

- $\mathbf{x}_i = A_i \mathbf{s}_i$
- $\mathbf{s}_{[j]} = (s_{ij})_{i=1}^m$ are not independent

Advantages

- A closed form likelihood
- Takes naturally into account inter-subject variability

Some instances of IVA

- IVA-L [Lee, 2008]: $p_{\mathbf{s}_{[j]}}(\mathbf{y}_{[j]}) \propto \exp(-\sqrt{\sum_i (y_{ij})^2})$
- IVA-G [Anderson, 2011] [Via, 2011]: $p_{\mathbf{s}_{[j]}}(\mathbf{y}_{[j]}) = \mathcal{N}(\mathbf{y}_{[j]}; 0, \Sigma_j)$
- Many others: IVA-L-SOS [Bhinge, 2019]

No *shared response*.

State of the art

Some other related work

- Tensor ICA [Beckmann, 2005]
- SRM [Chen, 2015], FastSRM [Richard, Submitted to Aperture]
- Hyperalignment [Haxby, 2011]

Our contribution

Introducing MultiViewICA

- A model with a closed form likelihood
- A fast optimization algorithm
- Theoretical guarantees
- Improved source identification and localization

Our contribution

Our model

$\mathbf{x}_i \in \mathbb{R}^k, A_i \in \mathbb{R}^{k,k}$

$$\boxed{\mathbf{x}_i = A_i(\mathbf{s} + \mathbf{n}_i), \quad i = 1, \dots, m} \tag{3}$$

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

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

- Gaussian noise $\mathbf{n}_i \sim \mathcal{N}(0, \sigma^2 I_k)$
- Independent noise and independent from sources
- Noise on the source side

A closed-form likelihood

Example (Typical noisy ICA likelihood (Bermond, Cardoso, 1999))

- $\mathbf{x}_i = A_i\mathbf{s} + \mathbf{n}_i, \mathbf{n}_i \sim N(0, \sigma^2 I_k)$ $I_k \in \mathbb{R}^{k,k}$ is the identity matrix.
- $p(\mathbf{s}) = d(s_1) \cdots d(s_k)$

  Denoting $A = \begin{bmatrix} A_1 \\ \vdots \\ A_n \end{bmatrix}$, $\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{bmatrix}$, we have

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{s})p(\mathbf{s})d\mathbf{s} \tag{4}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \int e^{-\frac{1}{2\sigma^2}||\mathbf{x}-A\mathbf{s}||_2^2} p(\mathbf{s})d\mathbf{s} \tag{5}$$

- Intractable in general (it does not factorize)

A closed-form likelihood

Our model

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

$\mathbf{n}_{ik} \sim \mathcal{N}(0, \sigma^2)$, $p(\mathbf{s}) = d(s_1) \cdots d(s_k)$

Closed form likelihood of our model.

- Bla
- $p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{s})p(\mathbf{s})d\mathbf{s}$

A closed-form likelihood

Our model

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

$\mathbf{n}_{ik} \sim \mathcal{N}(0, \sigma^2)$, $p(\mathbf{s}) = d(s_1) \cdots d(s_k)$

Closed form likelihood of our model.

- Bla
- $p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{s})p(\mathbf{s})d\mathbf{s}$

A closed-form likelihood

Our model

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

$\mathbf{n}_{ik} \sim \mathcal{N}(0, \sigma^2)$, $p(\mathbf{s}) = d(s_1) \cdots d(s_k)$

Closed form likelihood of our model.

- Bla
- $p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{s}) p(\mathbf{s}) d\mathbf{s}$
- $p(\mathbf{x}) = \int \prod_i p(\mathbf{x}_i|\mathbf{s}) \prod_j d(s_j) d\mathbf{s}$

A closed-form likelihood

Our model

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

$\mathbf{n}_{ik} \sim \mathcal{N}(0, \sigma^2)$, $p(\mathbf{s}) = d(s_1) \cdots d(s_k)$

Closed form likelihood of our model.

- Bla
- $p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{s}) p(\mathbf{s}) d\mathbf{s}$
- $p(\mathbf{x}) = \int \prod_i (|W_i| p_n^i (W_i \mathbf{x}_i - \mathbf{s})) \prod_j d(s_j) ds_j$, $W_i = {A_i}^{-1}$

□

A closed-form likelihood

Our model

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

$\mathbf{n}_{ik} \sim \mathcal{N}(0, \sigma^2)$, $p(\mathbf{s}) = d(s_1) \cdots d(s_k)$

Closed form likelihood of our model.

- Bla
- $p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{s})p(\mathbf{s})d\mathbf{s}$
- $p(\mathbf{x}) = \int \prod_i (|W_i| p_n^i(W_i\mathbf{x}_i - \mathbf{s})) \prod_j d(s_j)ds_j$, $W_i = {A_i}^{-1}$

A closed-form likelihood

Our model

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

$\mathbf{n}_{ik} \sim \mathcal{N}(0, \sigma^2)$, $p(\mathbf{s}) = d(s_1) \cdots d(s_k)$

Closed form likelihood of our model.

- Bla
- $p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{s})p(\mathbf{s})d\mathbf{s}$
- $p(\mathbf{x}) = \int \prod_i (|W_i| p_n^i(W_i\mathbf{x}_i - \mathbf{s})) \prod_j d(s_j)ds_j$, $W_i = {A_i}^{-1}$
- $p(\mathbf{x}) \propto \int \prod_i (|W_i| \exp(-\frac{1}{2\sigma^2} \| W_i\mathbf{x}_i - \mathbf{s}\|^2) \prod_j d(s_j)ds_j$

A closed-form likelihood

Our model

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

$\mathbf{n}_{ik} \sim \mathcal{N}(0, \sigma^2)$, $p(\mathbf{s}) = d(s_1) \cdots d(s_k)$

Closed form likelihood of our model.

- Bla
- $p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{s})p(\mathbf{s})d\mathbf{s}$
- $p(\mathbf{x}) = \int \prod_i (|W_i| p_n^i(W_i\mathbf{x}_i - \mathbf{s})) \prod_j d(s_j)ds_j$, $W_i = {A_i}^{-1}$
- $p(\mathbf{x}) \propto \int \prod_i (|W_i| \exp(-\frac{1}{2\sigma^2} \| W_i\mathbf{x}_i - \mathbf{s}\|^2) \prod_j d(s_j)ds_j$

A closed-form likelihood

Our model

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

$\mathbf{n}_{ik} \sim \mathcal{N}(0, \sigma^2)$, $p(\mathbf{s}) = d(s_1) \cdots d(s_k)$

Closed form likelihood of our model.

- Bla
- $p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{s}) p(\mathbf{s}) d\mathbf{s}$
- $p(\mathbf{x}) = \int \prod_i (|W_i| p_n^i(W_i \mathbf{x}_i - \mathbf{s})) \prod_j d(s_j) ds_j$, $W_i = {A_i}^{-1}$
- $p(\mathbf{x}) \propto \int \prod_i (|W_i| \prod_j \exp(-\frac{1}{2\sigma^2}(\langle \mathbf{w}_{ij} | \mathbf{x}_i \rangle - s_j)^2) \prod_j d(s_j) ds_j$

A closed-form likelihood

Our model

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

$\mathbf{n}_{ik} \sim \mathcal{N}(0, \sigma^2)$, $p(\mathbf{s}) = d(s_1) \cdots d(s_k)$

Closed form likelihood of our model.

- Bla
- $p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{s}) p(\mathbf{s}) d\mathbf{s}$
- $p(\mathbf{x}) = \int \prod_i (|W_i| p_n^i(W_i \mathbf{x}_i - \mathbf{s})) \prod_j d(s_j) ds_j$, $W_i = {A_i}^{-1}$
- $p(\mathbf{x}) \propto \int \prod_i (|W_i| \prod_j \exp(-\frac{1}{2\sigma^2}(\langle \mathbf{w}_{ij} | \mathbf{x}_i \rangle - s_j)^2) \prod_j d(s_j) ds_j$

A closed-form likelihood

Our model

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

$\mathbf{n}_{ik} \sim \mathcal{N}(0, \sigma^2)$, $p(\mathbf{s}) = d(s_1) \cdots d(s_k)$

Closed form likelihood of our model.

- Bla

$\mathcal{L} = -\sum_{i=1}^{m} \log |W_i| + \frac{1}{2\sigma^2} \sum_{i=1}^{m} \|W_i \mathbf{x}_i - \tilde{\mathbf{s}}\|^2 + f(\tilde{\mathbf{s}})$

- $\tilde{\mathbf{s}} = \sum_{i=1}^{n} \frac{W_i \mathbf{x}_i}{n}$
- $f(s) = f(s_1) + \cdots f(s_k)$
- f is d smoothed by a Gaussian kernel:
  $f(s_j) = \int \exp(-\frac{1}{2\sigma^2} m z^2) d(s_j - z) dz$

Optimization

Fast optimization

- Alternate optimization (alternate between subjects)
- Quasi newton with an approximation of the Hessian

Quasi-Newton

- Updates $W_i = (I + \rho D)W_i$, $D = {H_i}^{-1} G_i$ such that

$$\mathcal{L}((I+\varepsilon)W_i)) = \mathcal{L}(W_i) + < \varepsilon | G_i > + < \varepsilon | H_i \varepsilon >, \quad H_i \in \mathbb{R}^{k \times k \times k \times k}$$

$$(\mathcal{H}_i)_{abcd} = \delta_{ad}\delta_{bc} + \delta_{ac}\left(\frac{1}{m^2} f''(\tilde{s}_a) + \frac{1 - 1/m}{\sigma^2}\right) y_{ib} y_{id} \tag{6}$$

$$\text{where } \mathbf{y}_i = W_i \mathbf{x}_i \tag{7}$$

- $\mathcal{H}_i$ has $O(k^3)$ non zero coefficients.

Optimization

Fast optimization

- Alternate optimization (alternate between subjects)
- Quasi newton with an approximation of the Hessian

Quasi-Newton

- Updates $W_i = (I + \rho D)W_i$, $D = {H_i}^{-1} G_i$ such that

$$\mathcal{L}((I+\varepsilon)W_i)) = \mathcal{L}(W_i) + <\varepsilon|G_i> + <\varepsilon|H_i\varepsilon>, \quad H_i \in \mathbb{R}^{k\times k\times k\times k}$$

$$(H_i)_{abcd} = \delta_{ad}\delta_{bc} + \delta_{ac}\delta_{bd}\left(\frac{1}{m^2}f''(\tilde{s}_a) + \frac{1-1/m}{\sigma^2}\right)(y_{ib})^2 \quad (6)$$

$$\text{where } \mathbf{y}_i = W_i\mathbf{x}_i \quad (7)$$

- H_i has $O(k^2)$ non zero coefficients. It is 2x2 block diagonal:

Theoretical contributions

Theorem (Identifiability of MultiViewICA)

- *Consider* $\mathbf{x}_i = A_i(\mathbf{s} + \mathbf{n}_i)$, $i = 1, \dots, m$ $\mathbf{n}_i \sim \mathcal{N}(0, \sigma^2 I_k)$, $p(\mathbf{s}) = d(s_1) \cdots d(s_k)$, $\mathbf{s}$ *has at most one Gaussian component*
- *Assume* $\mathbf{x}_i = A'_i(\mathbf{s}' + \mathbf{n}'_i)$, $i = 1, \dots, m$, $\mathbf{n}'_i \sim \mathcal{N}(0, \sigma'^2 I_k)$, $p(\mathbf{s}') = d'(s'_1) \cdots d'(s'_k)$

Then, $\exists P \in \mathbb{R}^{k \times k}$ *a scale and permutation matrix s. t.* $\forall i \;\; A'_i = A_i P$.

Theoretical contributions

Theorem (Identifiability of MultiViewICA)

- *Consider* $\mathbf{x}_i = A_i(\mathbf{s} + \mathbf{n}_i)$, $i = 1, \dots, m$ $\mathbf{n}_i \sim \mathcal{N}(0, \sigma^2 I_k)$, $p(\mathbf{s}) = d(s_1) \cdots d(s_k)$, $\mathbf{s}$ *has at most one Gaussian component*
- *Assume* $\mathbf{x}_i = A'_i(\mathbf{s}' + \mathbf{n}'_i)$, $i = 1, \dots, m$, $\mathbf{n}'_i \sim \mathcal{N}(0, \sigma'^2 I_k)$, $p(\mathbf{s}') = d'(s'_1) \cdots d'(s'_k)$

Then, $\exists P \in \mathbb{R}^{k \times k}$ *a scale and permutation matrix s. t.* $\forall i \;\; A'_i = A_i P$.

Theorem (Robustness to choice of hyperparameter σ)

- *Assume sub-linear growth on* f': $|f'(x)| \leq c|x|^\alpha + d$ *for some* $c, d > 0$ *and* $0 < \alpha < 1$ *(Verified in practice)*
- *Assume that* $\mathbf{x}_i$ *follows the model with* σ' *and* d'.

Then, $\exists \Lambda$ *a diagonal matrix s.t.* $(\Lambda(A^1)^{-1}, \dots, \Lambda(A^m)^{-1})$ *is a stationary point of* $\mathcal{L}$.

Small recap

What do we have so far

- A principled approach to perform GroupICA
- Theoretical guarantees

Some questions

- Can we have a more general model while still keeping the likelihood in closed form ?
- How to deal with Gaussian sources ?

Our contribution: Shared ICA (ShICA)

ShICA model

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

- $\mathbf{n}_i \sim \mathcal{N}(0, \Sigma_i)$ where Σ_i is diagonal positive.
- $\mathbf{s}$ are independent components some of which may be Gaussian

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-ML can separate Gaussian and non-Gaussian data

ShICA is identifiable

Theorem (Identifiability)

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

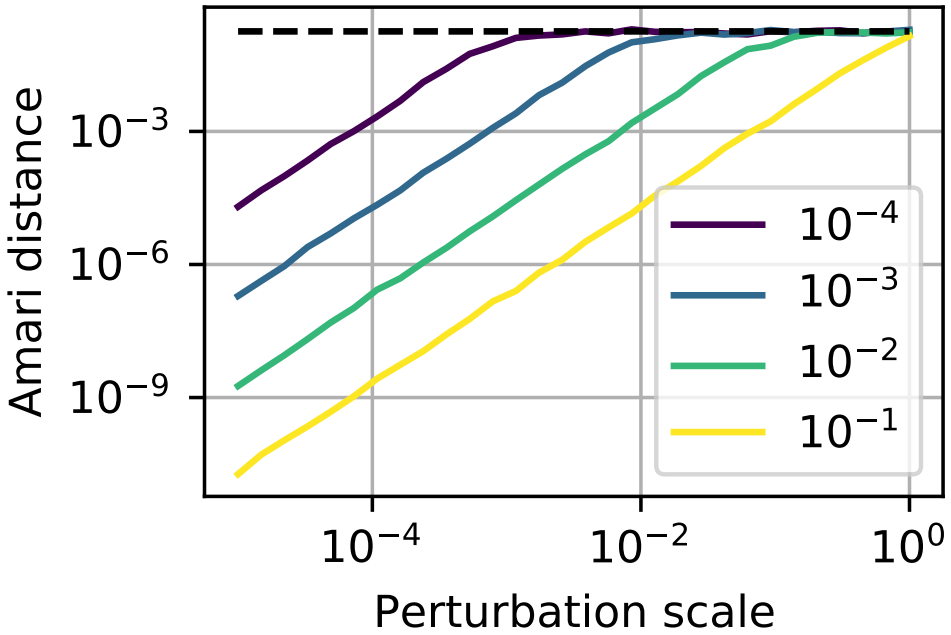
Note that the distinct eigenvalue condition is also necessary.
Note that the condition is stronger than noise diversity (we can exhibit an identifiable example on which MCCA fails).

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.



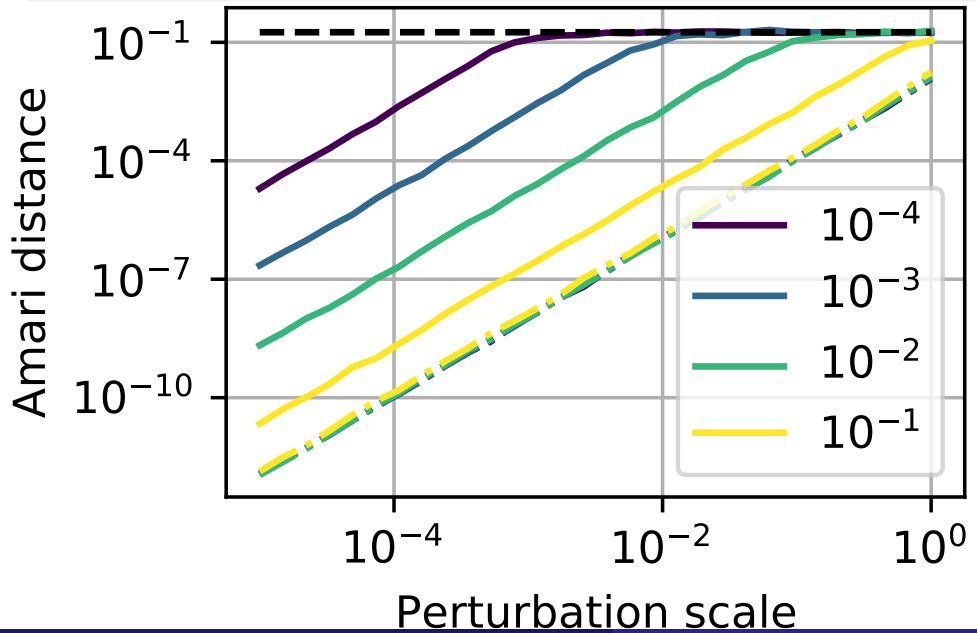
Solving practical issues with joint diagonalization

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 $Q\tilde{W}_i \frac{1}{n} X_i X_i^\top \tilde{W}_i^\top Q^\top$



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.

E-step: $\mathbb{E}[\mathbf{s}|\mathbf{x}] = \frac{\sum_{\alpha \in \{\frac{1}{2}, \frac{3}{2}\}} \theta_\alpha \mathbf{e}_\alpha}{\sum_{\alpha \in \{\frac{1}{2}, \frac{3}{2}\}} \theta_\alpha}$ $\mathbb{V}[\mathbf{s}|\mathbf{x}] = \frac{\sum_{\alpha \in \{\frac{1}{2}, \frac{3}{2}\}} \theta_\alpha V_\alpha}{\sum_{\alpha \in \{\frac{1}{2}, \frac{3}{2}\}} \theta_\alpha}$

$\mathbf{e}_\alpha = \left(\sum_{i=1}^m \Sigma_i^{-1} + \frac{1}{\alpha} I\right)^{-1} \sum_{i=1}^m \left(\Sigma_i^{-1} W_i \mathbf{x}_i\right)$, $V_\alpha = (\sum_{i=1}^m \Sigma_i^{-1} + \frac{1}{\alpha} I)^{-1}$

and $\theta_\alpha = \mathcal{N}((\sum_{i=1}^m \Sigma_i^{-1})^{-1} \Sigma_i^{-1} \mathbf{y}_i; 0, (\alpha I + (\sum_{i=1}^m \Sigma_i^{-1})^{-1})^{\frac{1}{2}})$

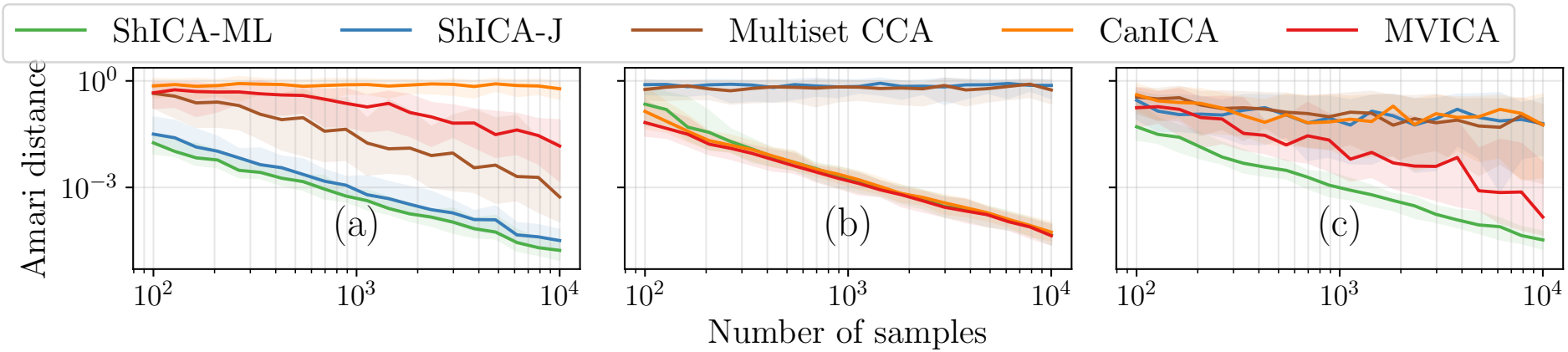
M-step: Closed form updates for noise variances and quasi-newton updates for unmixing matrices.

ShICA-J provides a great initialization to ShICA-ML

Synthetic experiments

Separation performance depending on the density of sources

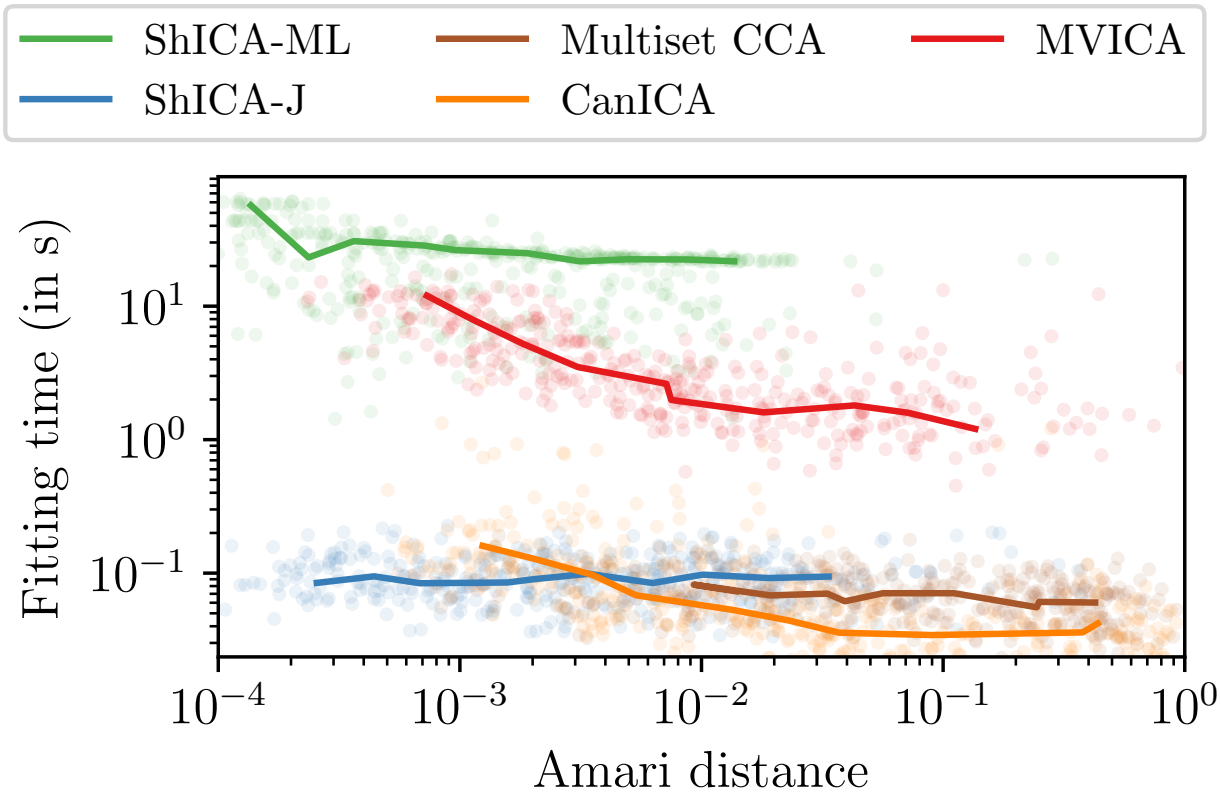
$m = 4$ views, $k = 5$ components, non-Gaussian sources are from a Laplace density, we use the ShICA model using:
(a) Gaussian components with noise diversity
(b) non-Gaussian components without noise diveristy
(c) Half of components are Gaussian with noise diversity, the other half is non-Gaussian without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



Synthetic experiments

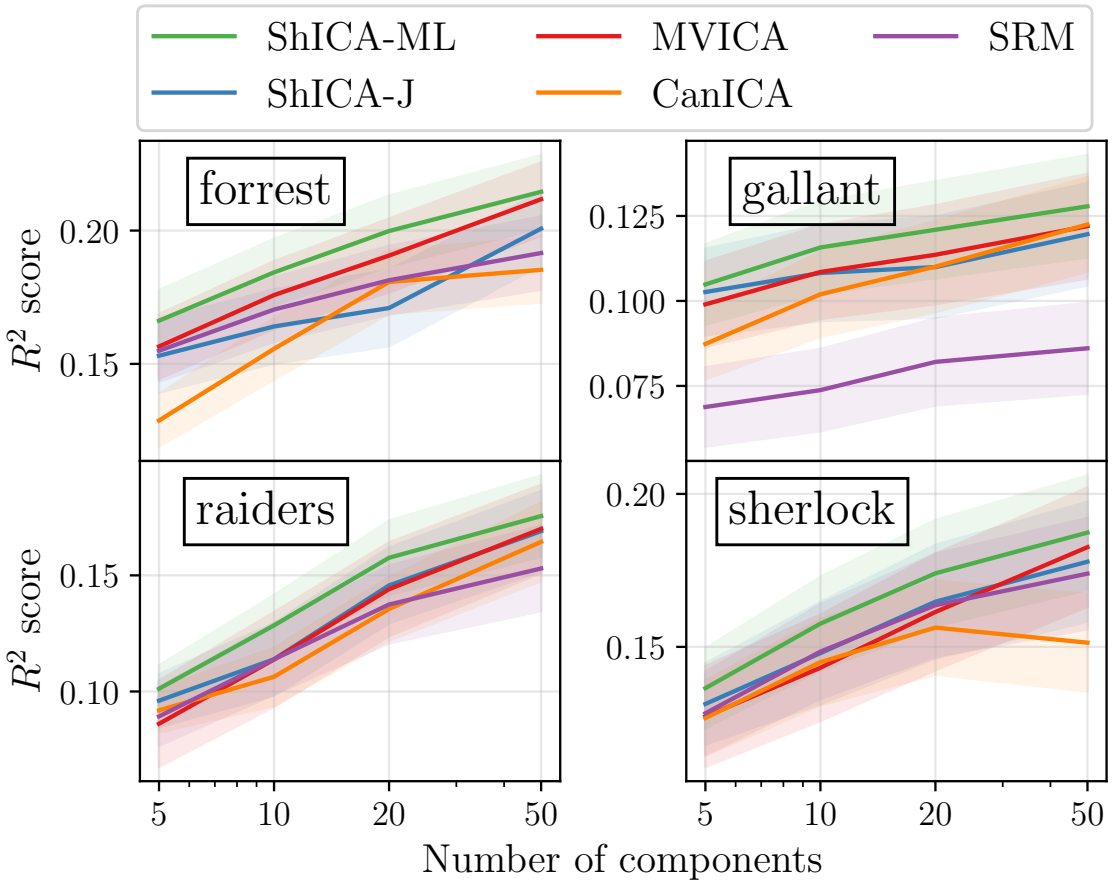
Computation time

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 $n = 10^2 \dots 10^4$.



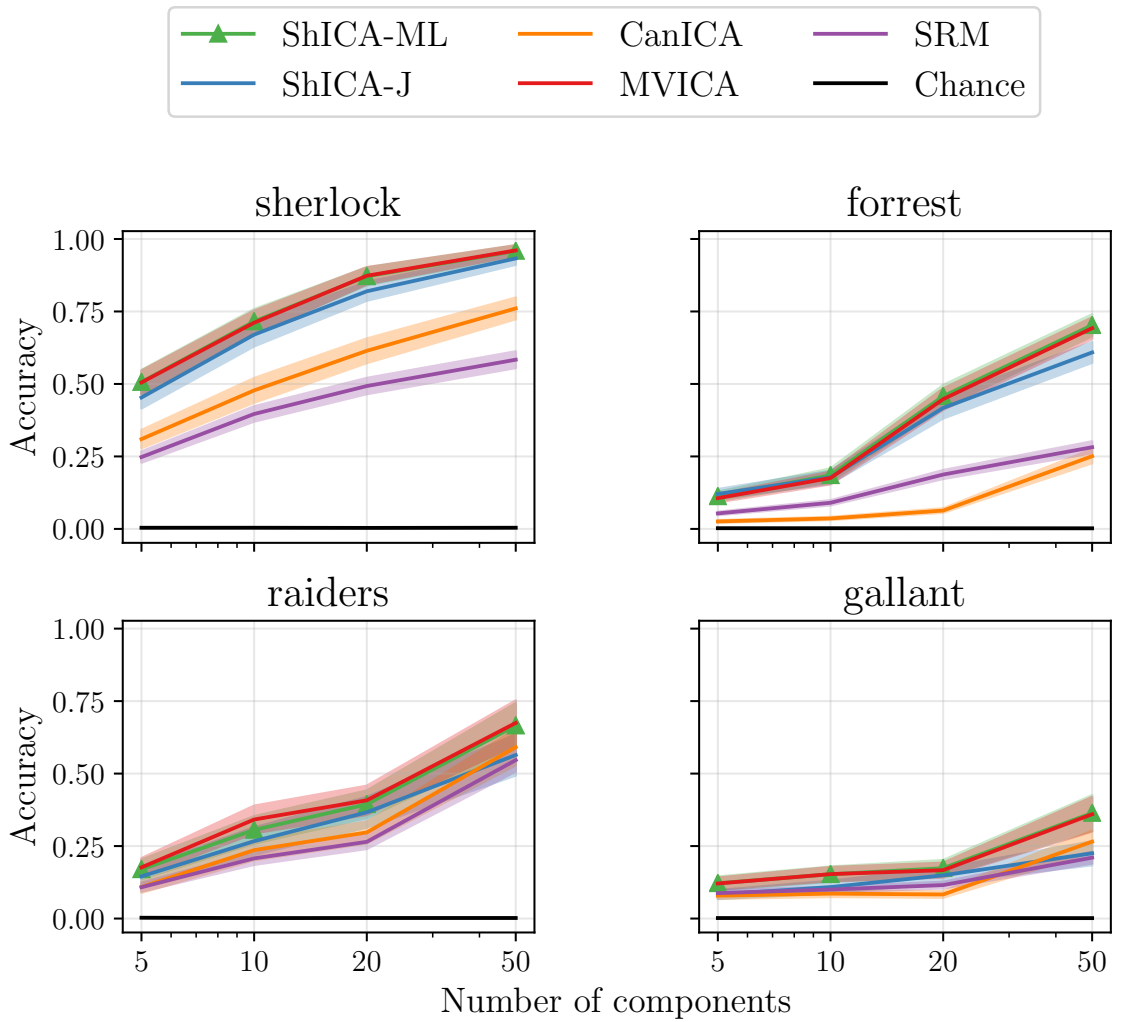
Reconstruction experiment fMRI

- 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



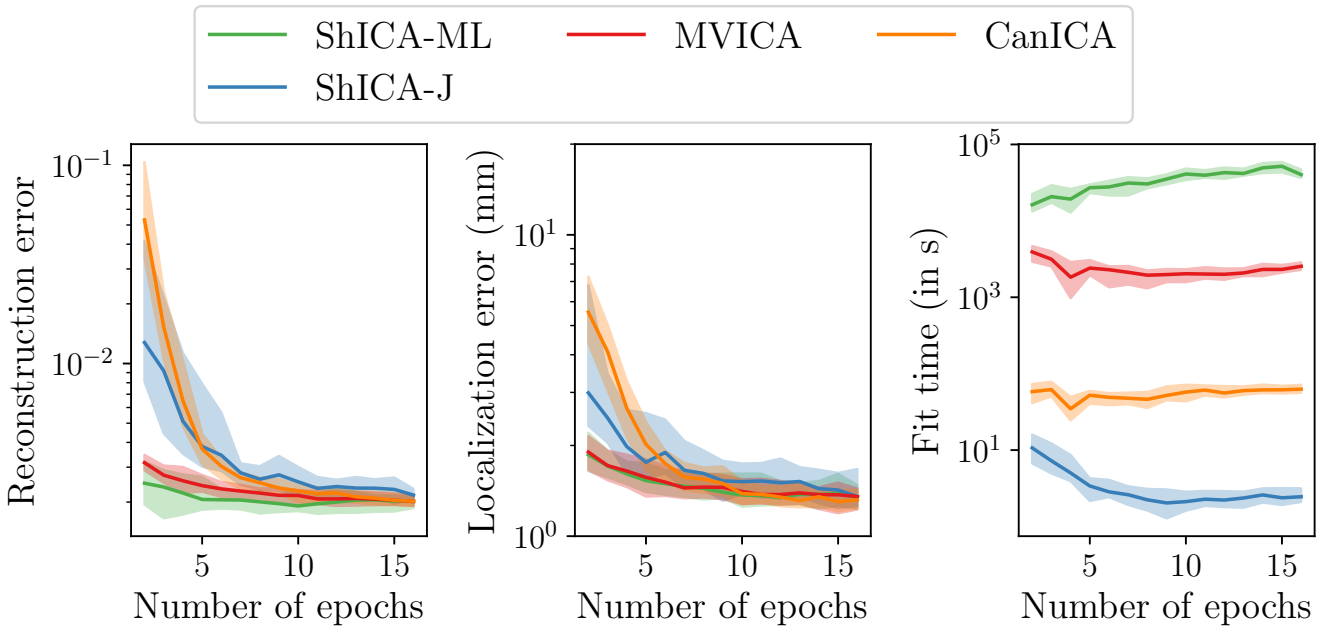
Timesegment matching fMRI

- 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.

About us

ICA may be old stuff but ideas can be reused

- Our Hessian approximation is specific to ICA but a similar technique is used for Fast joint diagonalization.
- Multiplicative algorithms provide a way to "remove the Jacobian term" when training deep generative neural networks".
- ICA is done under orthogonal or invertible constrains on mixing matrices leading to optimization on Riemannian manifold. The natural gradient in ICA coincides with the relative gradient. In neural networks, such ideas are used to obtain *equivariant* networks.

About us

Other contributions

- Mvlearn: MultiView Machine Learning in Python (JMLR, 2021 - non-first author)
- Brainiak: The brain imaging analysis kit (OSF, 2020 - non-first author)
- Local optimal transport for functional brain template estimation (IPMI, 2019 - non-first author)
- Optimizing deep video representation to match brain activity (CCN, 2018)
- A fast algorithm for performing dimension reduction under orthogonal constrains (srm) (OHBM, 2018)