Nonlinear ICA

identifiability, applications and directions of research

Carles Balsells Rodas

cb221.ac.uk
Imperial College London

June 23, 2022

Table of Contents

Outline

Outline

- What is ICA?

Outline

- What is ICA?
  - Independent Component Analysis: Identify **latent independent sources** which generate the data via some "mixing" of the sources.

Outline

- What is ICA?
  - Independent Component Analysis: Identify **latent independent sources** which generate the data via some "mixing" of the sources.
  - Not to be confused with PCA $\rightarrow$ data compression.

Outline

- What is ICA?
  - Independent Component Analysis: Identify **latent independent sources** which generate the data via some "mixing" of the sources.
  - Not to be confused with PCA $\rightarrow$ data compression.
- **Goal**: Learn *useful* representations in data $\rightarrow$ statistically independent.

Outline

- What is ICA?
  - Independent Component Analysis: Identify **latent independent sources** which generate the data via some "mixing" of the sources.
  - Not to be confused with PCA $\rightarrow$ data compression.
- **Goal**: Learn *useful* representations in data $\rightarrow$ statistically independent.
  - independent latent components $\rightarrow$ "principled disentanglement"

Outline

- What is ICA?
  - Independent Component Analysis: Identify **latent independent sources** which generate the data via some "mixing" of the sources.
  - Not to be confused with PCA $\rightarrow$ data compression.
- **Goal**: Learn *useful* representations in data $\rightarrow$ statistically independent.
  - independent latent components $\rightarrow$ "principled disentanglement"
- Linear ICA is **successful**.

Outline

- What is ICA?
  - Independent Component Analysis: Identify **latent independent sources** which generate the data via some "mixing" of the sources.
  - Not to be confused with PCA $\rightarrow$ data compression.
- **Goal**: Learn *useful* representations in data $\rightarrow$ statistically independent.
  - independent latent components $\rightarrow$ "principled disentanglement"
- Linear ICA is **successful**.
- **Problem**: Nonlinear ICA is ill-defined $\rightarrow$ not identifiable

Outline

- What is ICA?
  - Independent Component Analysis: Identify **latent independent sources** which generate the data via some "mixing" of the sources.
  - Not to be confused with PCA → data compression.
- **Goal**: Learn *useful* representations in data → statistically independent.
  - independent latent components → "principled disentanglement"
- Linear ICA is **successful**.
- **Problem**: Nonlinear ICA is ill-defined → not identifiable
  1. Use **temporal structure**.

Outline

- What is ICA?
  - Independent Component Analysis: Identify **latent independent sources** which generate the data via some "mixing" of the sources.
  - Not to be confused with PCA → data compression.
- **Goal**: Learn *useful* representations in data → statistically independent.
  - independent latent components → "principled disentanglement"
- Linear ICA is **successful**.
- **Problem**: Nonlinear ICA is ill-defined → not identifiable
  1. Use **temporal structure**.
  2. Use **auxiliary observed variables**.

Outline

- What is ICA?
  - Independent Component Analysis: Identify **latent independent sources** which generate the data via some "mixing" of the sources.
  - Not to be confused with PCA $\rightarrow$ data compression.
- **Goal**: Learn *useful* representations in data $\rightarrow$ statistically independent.
  - independent latent components $\rightarrow$ "principled disentanglement"
- Linear ICA is **successful**.
- **Problem**: Nonlinear ICA is ill-defined $\rightarrow$ not identifiable
  1. Use **temporal structure**.
  2. Use **auxiliary observed variables**.
  3. Consider extra assumptions on the **mixing** function.

Linear ICA

Linear independent component analysis

$$x_i = \sum_{j=1}^{n} a_{ij} s_j \quad i = 1, \dots, n \tag{1}$$

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \quad p(\mathbf{s}) = \prod_{i=1}^{n} p_i(s_i) \tag{2}$$

- $\mathbf{x}$ denotes the observation.
- $\mathbf{A} = \{a_{ij}\}_{ij}^{n}$ is the linear mixing.
- $\mathbf{s}$ denotes the independent latent sources.

Linear independent component analysis

$$x_i = \sum_{j=1}^{n} a_{ij} s_j \quad i = 1, \ldots, n \tag{1}$$

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \quad p(\mathbf{s}) = \prod_{i=1}^{n} p_i(s_i) \tag{2}$$

- $\mathbf{x}$ denotes the observation.
- $\mathbf{A} = \{a_{ij}\}_{ij}^{n}$ is the linear mixing.
- $\mathbf{s}$ denotes the independent latent sources.

Linear ICA is **identifiable** assuming non-Gaussian sources $\mathbf{s}$ [Comon, 1994].

- Using only observations $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$, we can recover both $\mathbf{A}$ and $\mathbf{s}$.

Linear independent component analysis

$$x_i = \sum_{j=1}^{n} a_{ij} s_j \quad i = 1, \dots, n \tag{1}$$

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \quad p(\mathbf{s}) = \prod_{i=1}^{n} p_i(s_i) \tag{2}$$

- $\mathbf{x}$ denotes the observation.
- $\mathbf{A} = \{a_{ij}\}_{ij}^{n}$ is the linear mixing.
- $\mathbf{s}$ denotes the independent latent sources.

Linear ICA is **identifiable** assuming non-Gaussian sources $\mathbf{s}$ [Comon, 1994].

- Using only observations $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$, we can recover both $\mathbf{A}$ and $\mathbf{s}$.
- Gaussian sources are not identifiable
  $\rightarrow$ any orthogonal transformation $\mathbf{s}' = \mathbf{R}\mathbf{s}$ leaves the distribution unchanged.

Linear ICA identifiability proof idea

Recall the linear ICA model

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \quad p(\mathbf{s}) = \prod_{i=1}^{n} p_i(s_i) \tag{3}$$

where $\mathbf{A}$ is the true linear mixing.

Linear ICA identifiability proof idea

Recall the linear ICA model

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \quad p(\mathbf{s}) = \prod_{i=1}^{n} p_i(s_i) \tag{3}$$

where $\mathbf{A}$ is the true linear mixing. Compute estimate of the sources $\mathbf{s}'$

$$\mathbf{x} = \mathbf{F}\mathbf{s}' \tag{4}$$

Linear ICA identifiability proof idea

Recall the linear ICA model

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \quad p(\mathbf{s}) = \prod_{i=1}^{n} p_i(s_i) \tag{3}$$

where $\mathbf{A}$ is the true linear mixing. Compute estimate of the sources $\mathbf{s}'$

$$\mathbf{x} = \mathbf{F}\mathbf{s}' \tag{4}$$

Identifiability is achieved by showing the following relation

$$\mathbf{F} = \mathbf{A}\mathbf{D}\mathbf{P}, \tag{5}$$

where $\mathbf{D}$ is a diagonal matrix and $\mathbf{P}$ is a permutation [Comon, 1994].

Linear ICA identifiability proof idea

Recall the linear ICA model

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \quad p(\mathbf{s}) = \prod_{i=1}^{n} p_i(s_i) \tag{3}$$

where $\mathbf{A}$ is the true linear mixing. Compute estimate of the sources $\mathbf{s}'$

$$\mathbf{x} = \mathbf{F}\mathbf{s}' \tag{4}$$

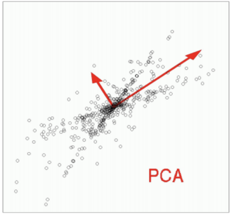
Identifiability is achieved by showing the following relation

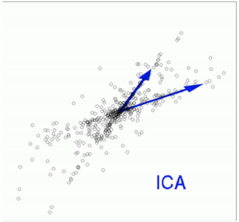
$$\mathbf{F} = \mathbf{A}\mathbf{D}\mathbf{P}, \tag{5}$$

where $\mathbf{D}$ is a diagonal matrix and $\mathbf{P}$ is a permutation [Comon, 1994]. Linear ICA is easily estimated by maximizing non-Gaussianity.

ICA vs PCA

ICA should not be confused with PCA!

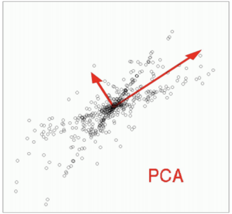


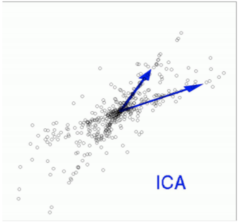


- PCA estimates directions with greatest variance in data (principal components)
- ICA estimates the statistically independent components.

ICA vs PCA

ICA should not be confused with PCA!





- PCA estimates directions with greatest variance in data (principal components)
- ICA estimates the statistically independent components.

PCA is **not** identifiable $\rightarrow$ cannot find the original sources.

Nonlinear ICA

Nonlinear ICA

Can we perform a similar analysis for a nonlinear mixing?
→ generalise disentanglement

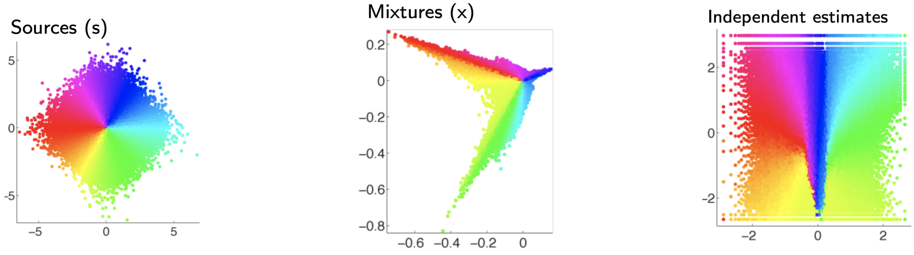
$$x_i = f_i(s_1, \dots, s_n) \quad i = 1, \dots, n \quad p(\mathbf{s}) = \prod_{i=1}^{n} p_i(s_i) \tag{6}$$

Nonlinear ICA

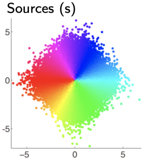
Can we perform a similar analysis for a nonlinear mixing?
$\rightarrow$ generalise disentanglement

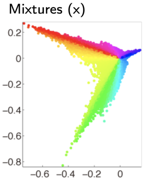
$$x_i = f_i(s_1, \ldots, s_n) \quad i = 1, \ldots, n \quad p(\mathbf{s}) = \prod_{i=1}^{n} p_i(s_i) \tag{6}$$

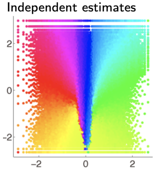
we cannot recover original sources with the same assumptions



Nonlinear ICA



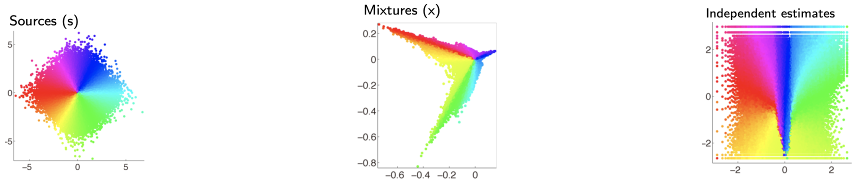




- **Identifiability**

$$p_\theta(\mathbf{x}) = p_{\theta'}(\mathbf{x}) \implies \theta = \theta', \quad \forall(\theta, \theta') \tag{7}$$

Nonlinear ICA

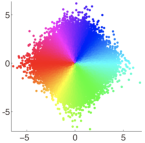


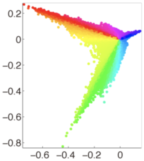
- **Identifiability**

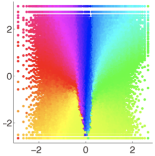
$$p_\theta(\mathbf{x}) = p_{\theta'}(\mathbf{x}) \implies \theta = \theta', \quad \forall(\theta, \theta') \tag{7}$$

- Nonlinear ICA is not identifiable!
  [Darmois, 1951, Hyvärinen and Pajunen, 1999]

Nonlinear ICA







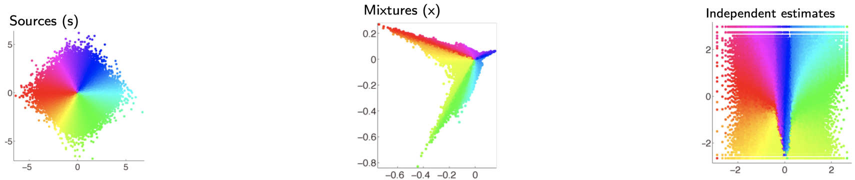
- **Identifiability**

$$p_\theta(\mathbf{x}) = p_{\theta'}(\mathbf{x}) \implies \theta = \theta', \quad \forall(\theta, \theta') \tag{7}$$

- Nonlinear ICA is not identifiable! [Darmois, 1951, Hyvärinen and Pajunen, 1999]
- **Darmois construction**
  - For any x_1, x_2, construct $y = h(x_1, x_2)$ independent of x_1 as

$$h(z_1, z_2) = p(x_2 < z_2 | x_1 = z_1) \tag{8}$$

Nonlinear ICA



- **Identifiability**

$$p_\theta(\mathbf{x}) = p_{\theta'}(\mathbf{x}) \implies \theta = \theta', \quad \forall(\theta, \theta') \tag{7}$$

- Nonlinear ICA is not identifiable! [Darmois, 1951, Hyvärinen and Pajunen, 1999]
- **Darmois construction**
  - For any x_1, x_2, construct $y = h(x_1, x_2)$ independent of x_1 as

$$h(z_1, z_2) = p(x_2 < z_2 | x_1 = z_1) \tag{8}$$

- Independence alone is too weak for identifiability.

ICA in time series

Time-Contrastive Learning (TCL) and ICA

First proof for identifiable nonlinear ICA [Hyvarinen and Morioka, 2016].

Time-Contrastive Learning (TCL) and ICA

First proof for identifiable nonlinear ICA [Hyvarinen and Morioka, 2016].

Time-contrastive Learning

1. Observe time series $\mathbf{x}(t) \in \mathbb{R}^n$.
2. Divide $\mathbf{x}(t)$ into T segments.
3. Train MLP to discriminate segments.
4. Last hidden layer $\mathbf{h}(\mathbf{x}; \theta)$ should account for **nonstationarity**.

Time-Contrastive Learning (TCL) and ICA

First proof for identifiable nonlinear ICA [Hyvarinen and Morioka, 2016].

Time-contrastive Learning

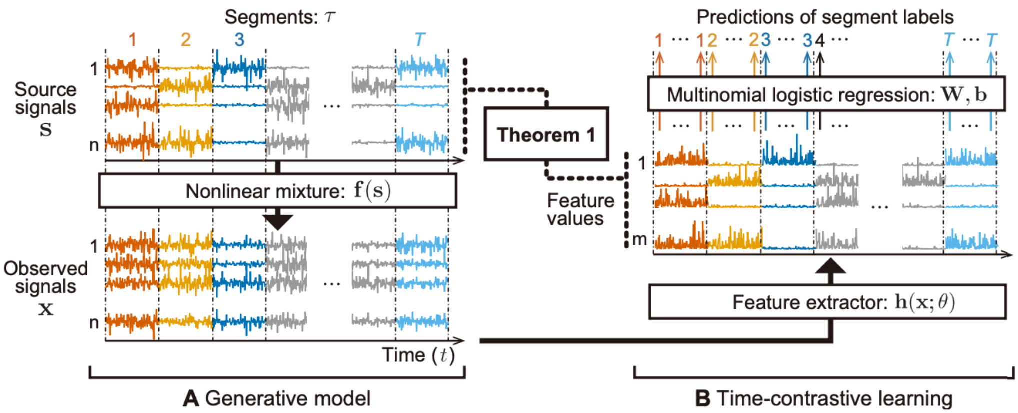
1. Observe time series $\mathbf{x}(t) \in \mathbb{R}^n$.
2. Divide $\mathbf{x}(t)$ into T segments.
3. Train MLP to discriminate segments.
4. Last hidden layer $\mathbf{h}(\mathbf{x}; \theta)$ should account for **nonstationarity**.

Nonstationary ICA

- $\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t))$
- $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$, smooth invertible and nonlinear.
- sources $s_i(t)$ are **nonstationary**

$$p_\tau(s_i) \sim q_{i,0}(s_i) + \sum_{v=1}^{V} \lambda_{i,v}(\tau) q_{i,v}(s_i) \quad (9)$$

Time-Contrastive Learning (TCL) and ICA



Time-Contrastive Learning (TCL) and ICA

- Assume we apply TCL on $\mathbf{x}(t)$.

Time-Contrastive Learning (TCL) and ICA

- Assume we apply TCL on $\mathbf{x}(t)$.
- TCL finds $\mathbf{s}(t)^2 = \mathbf{A}\mathbf{h}(\mathbf{x}(t))$ for some linear mixing $\mathbf{A}$
- TCL demixes nonlinear ICA up to linear mixing and squaring!
- Under further assumptions $\rightarrow$ $\mathbf{A}$ can be estimated by linear ICA.

Time-Contrastive Learning (TCL) and ICA

- Assume we apply TCL on $\mathbf{x}(t)$.
- TCL finds $\mathbf{s}(t)^2 = \mathbf{Ah}(\mathbf{x}(t))$ for some linear mixing $\mathbf{A}$
- TCL demixes nonlinear ICA up to linear mixing and squaring!
- Under further assumptions $\rightarrow$ $\mathbf{A}$ can be estimated by linear ICA.
- **Important result**: This opens the direction of nonlinear ICA in time-series

Time-Contrastive Learning (TCL) and ICA

- Assume we apply TCL on $\mathbf{x}(t)$.
- TCL finds $\mathbf{s}(t)^2 = \mathbf{Ah}(\mathbf{x}(t))$ for some linear mixing $\mathbf{A}$
- TCL demixes nonlinear ICA up to linear mixing and squaring!
- Under further assumptions $\rightarrow$ $\mathbf{A}$ can be estimated by linear ICA.
- **Important result**: This opens the direction of nonlinear ICA in time-series
  - Independence at every **time step** and **point** $\rightarrow$ more constraints $\rightarrow$ identifiability

Permutation-Contrastive Learning and ICA

- Similar idea as TCL for autorregressive time series → Sources are **stationary**

Permutation-Contrastive Learning and ICA

- Similar idea as TCL for autorregressive time series → Sources are **stationary**
  1. Observe time series $\mathbf{x}(t) \in \mathbb{R}^n$.

Permutation-Contrastive Learning and ICA

- Similar idea as TCL for autorregressive time series → Sources are **stationary**
  1. Observe time series $\mathbf{x}(t) \in \mathbb{R}^n$.
  2. Take short time windows:

$$\mathbf{y}(t) = \Big(\mathbf{x}(t), \mathbf{x}(t-1)\Big) \tag{10}$$

Permutation-Contrastive Learning and ICA

- Similar idea as TCL for autorregressive time series $\rightarrow$ Sources are **stationary**
  1. Observe time series $\mathbf{x}(t) \in \mathbb{R}^n$.
  2. Take short time windows:

$$\mathbf{y}(t) = \Big(\mathbf{x}(t), \mathbf{x}(t-1)\Big) \tag{10}$$

  3. Create randomly time-permuted data:

$$\mathbf{y} * (t) = \Big(\mathbf{x}(t), \mathbf{x}(t*)\Big), \tag{11}$$

where t* is a random time step

Permutation-Contrastive Learning and ICA

- Similar idea as TCL for autorregressive time series → Sources are **stationary**
  1. Observe time series $\mathbf{x}(t) \in \mathbb{R}^n$.
  2. Take short time windows:

$$\mathbf{y}(t) = \Big(\mathbf{x}(t), \mathbf{x}(t-1)\Big) \tag{10}$$

  3. Create randomly time-permuted data:

$$\mathbf{y}*(t) = \Big(\mathbf{x}(t), \mathbf{x}(t*)\Big), \tag{11}$$

     where t* is a random time step
  4. Train an MLP to classify $\mathbf{y}$ and $\mathbf{y}*$

Permutation-Contrastive Learning and ICA

- Similar idea as TCL for autorregressive time series $\rightarrow$ Sources are **stationary**
  1. Observe time series $\mathbf{x}(t) \in \mathbb{R}^n$.
  2. Take short time windows:

$$\mathbf{y}(t) = \Big(\mathbf{x}(t), \mathbf{x}(t-1)\Big) \tag{10}$$

  3. Create randomly time-permuted data:

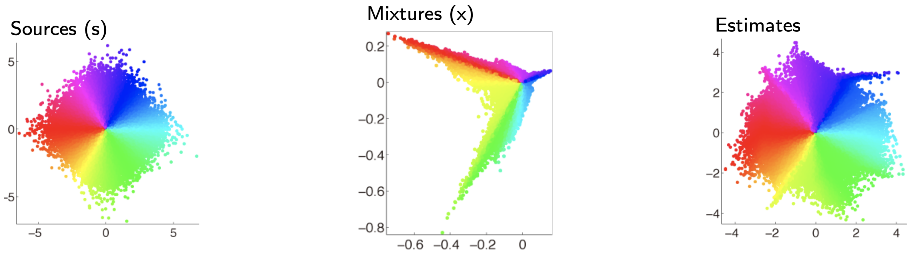
$$\mathbf{y} * (t) = \Big(\mathbf{x}(t), \mathbf{x}(t*)\Big), \tag{11}$$

     where t* is a random time step
  4. Train an MLP to classify $\mathbf{y}$ and $\mathbf{y}*$
- Under certain assumptions, we have $s_i(t) = k_i(h_j(\mathbf{x}(t)))$ for some ordering of j and scalar nonlinearities k_i.

Permutation-Contrastive Learning and ICA

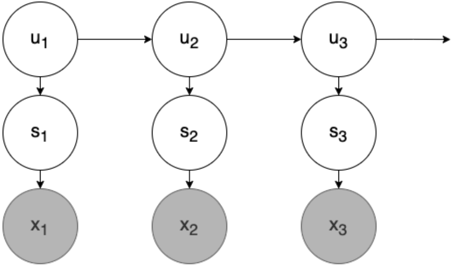
Example: Autoregressive model with Laplacian innovations

$$\log p(s(t)|s(t-1)) = -|s(t) - \rho s(t-1)| \quad (12)$$



Hidden Markov Nonlinear ICA

- Learn states and dynamics using the Hidden Markov model (HMM) framework.
- Similar to TCL with **latent** conditioning variables.

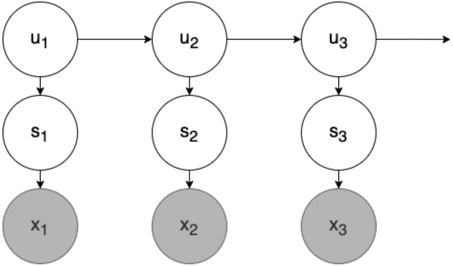


$$p(\mathbf{s}^{(t)}|u^{(t)};\boldsymbol{\lambda}_{u^{(t)}}) \tag{13}$$

$$= \prod_{i=1}^{n} \frac{h(s_i^{(t)})}{Z(\boldsymbol{\lambda}_{i,u^{(t)}})} \exp\{\langle \boldsymbol{\lambda}_{i,u^{(t)}}, \mathbf{T}_i(s_i)\rangle\} \tag{14}$$

Hidden Markov Nonlinear ICA

- Learn states and dynamics using the Hidden Markov model (HMM) framework.
- Similar to TCL with **latent** conditioning variables.



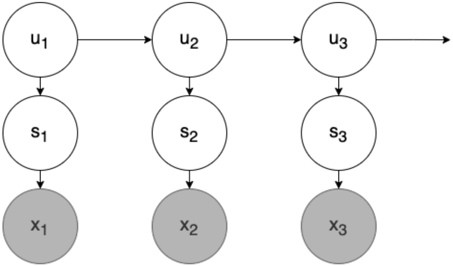
$$p(\mathbf{s}^{(t)}|u^{(t)};\boldsymbol{\lambda}_{u^{(t)}}) \tag{13}$$

$$= \prod_{i=1}^{n} \frac{h(s_i^{(t)})}{Z(\boldsymbol{\lambda}_{i,u^{(t)}})} \exp\{\langle \boldsymbol{\lambda}_{i,u^{(t)}}, \mathbf{T}_i(s_i)\rangle\} \tag{14}$$

- **Idea:** Identify independent components using HMM identifiability [Gassiat et al., 2016].

Hidden Markov Nonlinear ICA

- Learn states and dynamics using the Hidden Markov model (HMM) framework.
- Similar to TCL with **latent** conditioning variables.



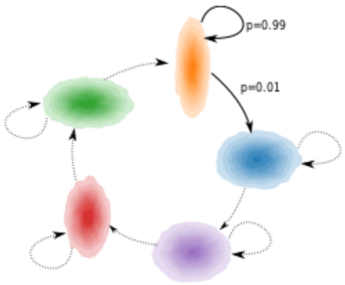
$$p(\mathbf{s}^{(t)}|u^{(t)}; \boldsymbol{\lambda}_{u^{(t)}}) \tag{13}$$

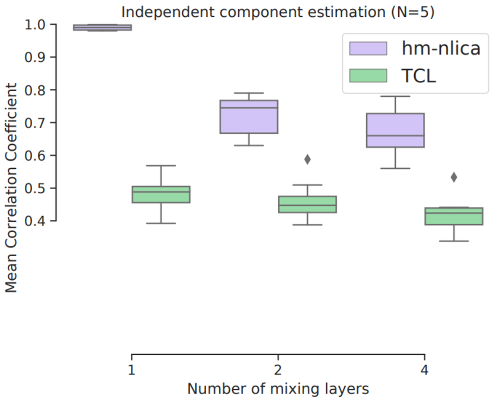
$$= \prod_{i=1}^{n} \frac{h(s_i^{(t)})}{Z(\boldsymbol{\lambda}_{i,u^{(t)}})} \exp\{\langle \boldsymbol{\lambda}_{i,u^{(t)}}, \mathbf{T}_i(s_i)\rangle\} \tag{14}$$

- **Idea:** Identify independent components using HMM identifiability [Gassiat et al., 2016].
- Strong identifiability results $s_i = w_{ij}\hat{g}_j(\mathbf{x}) + b_{ij}$ [Hälvä and Hyvarinen, 2020].

Hidden Markov Nonlinear ICA

- Learning can be done by EM (Baum-Welch).





Structured Nonlinear ICA (SNICA)

- Generalise identifiable nonlinear ICA for structured noisy data (e.g. time series) [Hälvä et al., 2021].

Structured Nonlinear ICA (SNICA)

- Generalise identifiable nonlinear ICA for structured noisy data (e.g. time series) [Hälvä et al., 2021].
- **Q:** What type of structures allow identifiable disentanglement?

Structured Nonlinear ICA (SNICA)

- Generalise identifiable nonlinear ICA for structured noisy data (e.g. time series) [Hälvä et al., 2021].
- **Q:** What type of structures allow identifiable disentanglement?
- Main assumptions:

Structured Nonlinear ICA (SNICA)

- Generalise identifiable nonlinear ICA for structured noisy data (e.g. time series) [Hälvä et al., 2021].
- **Q:** What type of structures allow identifiable disentanglement?
- Main assumptions:
  1. Weak stationarity: distributions for $s_i^{(t)}$ and $s_i^{(t')}$ are the same for any $t, t' \in \mathbb{T}, \forall i$.

Structured Nonlinear ICA (SNICA)

- Generalise identifiable nonlinear ICA for structured noisy data (e.g. time series) [Hälvä et al., 2021].
- **Q:** What type of structures allow identifiable disentanglement?
- Main assumptions:
  1. Weak stationarity: distributions for $s_i^{(t)}$ and $s_i^{(t')}$ are the same for any $t, t' \in \mathbb{T}, \forall i$.
  2. Unconditional independence between components

$$p(\mathbf{s}^{(t_1)}, \ldots, \mathbf{s}^{(t_m)}) = \prod_{i=1}^{n} p(s_i^{(t_1)}, \ldots, s_i^{(t_m)}) \tag{15}$$

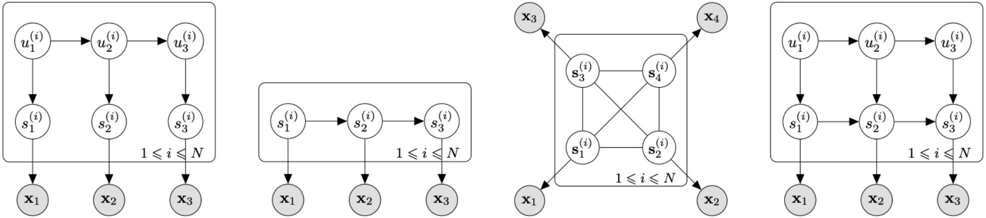
Structured Nonlinear ICA (SNICA)

- Generalise identifiable nonlinear ICA for structured noisy data (e.g. time series) [Hälvä et al., 2021].
- **Q:** What type of structures allow identifiable disentanglement?
- Main assumptions:
  1. Weak stationarity: distributions for $s_i^{(t)}$ and $s_i^{(t')}$ are the same for any $t, t' \in \mathbb{T}, \forall i$.
  2. Unconditional independence between components

$$p(\mathbf{s}^{(t_1)}, \ldots, \mathbf{s}^{(t_m)}) = \prod_{i=1}^{n} p(s_i^{(t_1)}, \ldots, s_i^{(t_m)}) \tag{15}$$

  3. Noisy model $\mathbf{x}^t = \mathbf{f}(\mathbf{s}^t) + \varepsilon^t$, with ε^t i.i.d. noise with unknown distribution.

Structured Nonlinear ICA (SNICA) – Examples



(a) HMM modulated components c.f. (Hälvä and Hyvärinen, 2020)

(b) Temporal dependencies c.f. (Hyvärinen and Morioka, 2017)

(c) New: Spatial process on a graph (with latent states u_t integrated out)

(d) New: Δ-SNICA , a linear switching dynamics model for components

- SNICA covers and extends previous identifiable models
- It also introduces new structured models (Δ-SNICA).

ICA in DLVMs

ICA in Deep Latent Variable Models (DLVM)

- Generative framework for data $\mathbf{x}$ and latent $\mathbf{z}$, with parameters θ.

$$p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z}) \tag{16}$$

and a data generative process

$$\mathcal{D} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}\} \tag{17}$$

$$\mathbf{z}^{*(i)} \sim p_{\theta^*}(\mathbf{z}), \quad \mathbf{x}^{(i)} \sim p_{\theta^*}(\mathbf{x}|\mathbf{z}^{*(i)}) \tag{18}$$

ICA in Deep Latent Variable Models (DLVM)

- Generative framework for data $\mathbf{x}$ and latent $\mathbf{z}$, with parameters θ.

$$p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z}) \tag{16}$$

and a data generative process

$$\mathcal{D} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}\} \tag{17}$$

$$\mathbf{z}^{*(i)} \sim p_{\theta^*}(\mathbf{z}), \quad \mathbf{x}^{(i)} \sim p_{\theta^*}(\mathbf{x}|\mathbf{z}^{*(i)}) \tag{18}$$

- Data likelihood can be computed as

$$\int p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z} = p_\theta(\mathbf{x}) \approx p_{\theta*}(\mathbf{x}) \tag{19}$$

ICA in Deep Latent Variable Models (DLVM)

- Variational autoencoders (VAEs) [Kingma and Welling, 2013]:
  1. Use factorised Gaussian prior $p(\mathbf{z}) = \prod_{i=1}^{n} p(z_i)$
  2. Posterior is defined as $\mathbf{x} = \mathbf{g}(\mathbf{z}) + \varepsilon$.

ICA in Deep Latent Variable Models (DLVM)

- Variational autoencoders (VAEs) [Kingma and Welling, 2013]:
  1. Use factorised Gaussian prior $p(\mathbf{z}) = \prod_{i=1}^{n} p(z_i)$
  2. Posterior is defined as $\mathbf{x} = \mathbf{g}(\mathbf{z}) + \varepsilon$.
- Recall identifiability in ICA

$$\forall(\theta, \theta') : p_\theta(\mathbf{x}) = p_{\theta'}(\mathbf{x}) \implies \theta = \theta' \tag{20}$$

ICA in Deep Latent Variable Models (DLVM)

- Variational autoencoders (VAEs) [Kingma and Welling, 2013]:
  1. Use factorised Gaussian prior $p(\mathbf{z}) = \prod_{i=1}^{n} p(z_i)$
  2. Posterior is defined as $\mathbf{x} = \mathbf{g}(\mathbf{z}) + \varepsilon$.
- Recall identifiability in ICA

$$\forall(\theta, \theta') : p_\theta(\mathbf{x}) = p_{\theta'}(\mathbf{x}) \implies \theta = \theta' \tag{20}$$

- VAEs are **not identifiable**.

ICA in Deep Latent Variable Models (DLVM)

- Variational autoencoders (VAEs) [Kingma and Welling, 2013]:
  1. Use factorised Gaussian prior $p(\mathbf{z}) = \prod_{i=1}^{n} p(z_i)$
  2. Posterior is defined as $\mathbf{x} = \mathbf{g}(\mathbf{z}) + \boldsymbol{\varepsilon}$.
- Recall identifiability in ICA

$$\forall(\theta, \theta') : p_\theta(\mathbf{x}) = p_{\theta'}(\mathbf{x}) \implies \theta = \theta' \tag{20}$$

- VAEs are **not identifiable**.
  - By Gaussianity, we have equivalence to othogonal rotations

$$\mathbf{z}' = \mathbf{R}\mathbf{z}, \quad \mathbf{z} \sim p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{21}$$

$$p_{\mathbf{z}'}(\boldsymbol{\xi}) = p_{\mathbf{z}}(R^T\boldsymbol{\xi})|\det R| = \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}||R^T\boldsymbol{\xi}||^2\right\} \tag{22}$$

$$= \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}||\boldsymbol{\xi}||^2\right\} = p_{\mathbf{z}}(\boldsymbol{\xi}) \tag{23}$$

ICA in Deep Latent Variable Models (DLVM)

- VAE is **not identifiable**.
- Practically used for data compression $\rightarrow$ PCA.

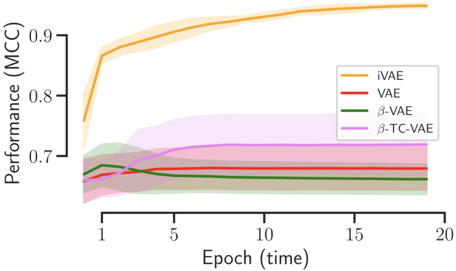
ICA in Deep Latent Variable Models (DLVM)

- VAE is **not identifiable**.
- Practically used for data compression $\rightarrow$ PCA.
- But conditioning makes the model identifiable (e.g. time segment, history, ...).

ICA in Deep Latent Variable Models (DLVM)

- VAE is **not identifiable**.
- Practically used for data compression $\rightarrow$ PCA.
- But conditioning makes the model identifiable (e.g. time segment, history, ...).
- **Solution**: Condition sources by some auxiliary observed variable $\mathbf{u}$ [Khemakhem et al., 2020].

- sources s_i **conditionally** independent given $\mathbf{u}$.
- Provably **identifiable**
- Estimated using identifiable VAE (iVAE).



ICA with unconditional priors

Nonlinear ICA with unconditional priors

- Recent identifiability proof with unconditional priors [Zheng et al., 2022].

$$\mathbf{x} = \mathbf{f}(\mathbf{s}), \quad p(\mathbf{s}) = \prod_{i=1}^{n} p_i(s_i) \tag{24}$$

Nonlinear ICA with unconditional priors

- Recent identifiability proof with unconditional priors [Zheng et al., 2022].

$$\mathbf{x} = \mathbf{f}(\mathbf{s}), \quad p(\mathbf{s}) = \prod_{i=1}^{n} p_i(s_i) \tag{24}$$

- Identifiability is achieved via restrictions on the mixing function $\mathbf{f}$.

Nonlinear ICA with unconditional priors

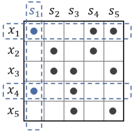
- Recent identifiability proof with unconditional priors [Zheng et al., 2022].

$$\mathbf{x} = \mathbf{f}(\mathbf{s}), \quad p(\mathbf{s}) = \prod_{i=1}^{n} p_i(s_i) \tag{24}$$

- Identifiability is achieved via restrictions on the mixing function $\mathbf{f}$.

Structural sparsity

1 - Given s_i, there exists a set of $\mathbf{x}$ such that s_i is the only latent source generating the set.



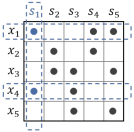
Nonlinear ICA with unconditional priors

- Recent identifiability proof with unconditional priors [Zheng et al., 2022].

$$\mathbf{x} = \mathbf{f}(\mathbf{s}), \quad p(\mathbf{s}) = \prod_{i=1}^{n} p_i(s_i) \tag{24}$$

- Identifiability is achieved via restrictions on the mixing function $\mathbf{f}$.

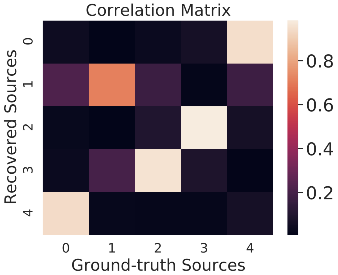
1. Structural sparsity
   - Given s_i, there exists a set of $\mathbf{x}$ such that s_i is the only latent source generating the set.

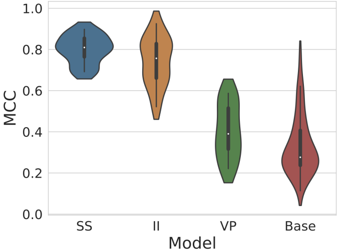


2. Factorial change of volume
   - E.g. volume-preserving transformation
   - It helps weakening the requirement of auxiliary variables.

Nonlinear ICA with unconditional priors

- Identifiable up to component-wise invertible transofrmation and permutation





Conclusions

Conclusions

- Identifiability in ICA is a fundamental issue, where for the linear case is solved.

Conclusions

- Identifiability in ICA is a fundamental issue, where for the linear case is solved.
- Nonlinear ICA is generally ill-defined as proven **unidentibiable**.

Conclusions

- Identifiability in ICA is a fundamental issue, where for the linear case is solved.
- Nonlinear ICA is generally ill-defined as proven **unidentibiable**.
- Additional assumptions are required for identifiability
  1. Structured data
  2. Auxiliary conditioning variables
  3. Restrictions on the nonlinear mixing

Conclusions

- Identifiability in ICA is a fundamental issue, where for the linear case is solved.
- Nonlinear ICA is generally ill-defined as proven **unidentibiable**.
- Additional assumptions are required for identifiability
  1. Structured data
  2. Auxiliary conditioning variables
  3. Restrictions on the nonlinear mixing
- ICA has connections to DLVMs $\rightarrow$ iVAE, Δ-SNICA

Conclusions

- Identifiability in ICA is a fundamental issue, where for the linear case is solved.
- Nonlinear ICA is generally ill-defined as proven **unidentibiable**.
- Additional assumptions are required for identifiability
  1. Structured data
  2. Auxiliary conditioning variables
  3. Restrictions on the nonlinear mixing
- ICA has connections to DLVMs $\rightarrow$ iVAE, Δ-SNICA
- ICA is a principled framework for "disentanglement"

References I

Comon, P. (1994).
Independent component analysis, a new concept?
Signal processing, 36(3):287–314.

Darmois, G. (1951).
Analyse des liaisons de probabilité.
In *Proc. Int. Stat. Conferences 1947*, page 231.

Gassiat, É., Cleynen, A., and Robin, S. (2016).
Inference in finite state space non parametric hidden markov models and applications.
Statistics and Computing, 26(1):61–71.

References II

Hälvä, H. and Hyvarinen, A. (2020).
Hidden markov nonlinear ica: Unsupervised learning from nonstationary time series.
In *Conference on Uncertainty in Artificial Intelligence*, pages 939–948. PMLR.

Hälvä, H., Le Corff, S., Lehéricy, L., So, J., Zhu, Y., Gassiat, E., and Hyvarinen, A. (2021).
Disentangling identifiable features from noisy data with structured nonlinear ica.
Advances in Neural Information Processing Systems, 34:1624–1633.

Hyvarinen, A. and Morioka, H. (2016).
Unsupervised feature extraction by time-contrastive learning and nonlinear ica.
Advances in neural information processing systems, 29.

References III

Hyvärinen, A. and Pajunen, P. (1999).
Nonlinear independent component analysis: Existence and uniqueness results.
Neural networks, 12(3):429–439.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. (2020).
Variational autoencoders and nonlinear ica: A unifying framework.
In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR.

Kingma, D. P. and Welling, M. (2013).
Auto-encoding variational bayes.
arXiv preprint arXiv:1312.6114.

References IV

Zheng, Y., Ng, I., and Zhang, K. (2022).
On the identifiability of nonlinear ica: Sparsity and beyond.
arXiv preprint arXiv:2206.07751.