Deterministic Independent Component Analysis (ICA)

Ruitong Huang   András György   Csaba Szepesvári

University of Alberta

July 17, 2015

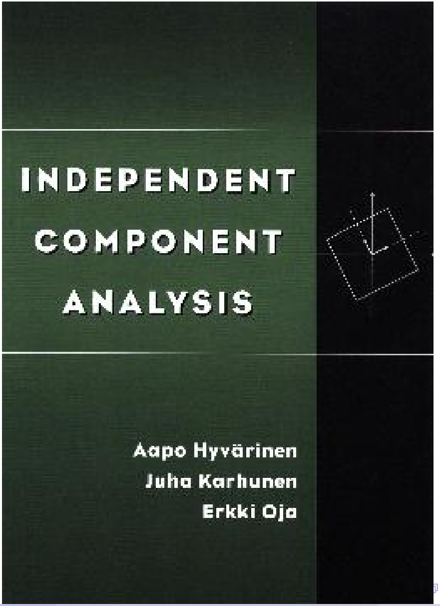
UNIVERSITY OF ALBERTA
DEPARTMENT OF COMPUTING SCIENCE

ALBERTA INNOVATES CENTRE FOR MACHINE LEARNING

Outline

1. Introduction
   - What is ICA, really?
   - Deterministic ICA
   - Has not this been done previously?
2. Results
   - Our algorithm: Deterministic ICA
3. Empirical Illustration
4. Conclusions

What is Independent Component Analysis (ICA)?



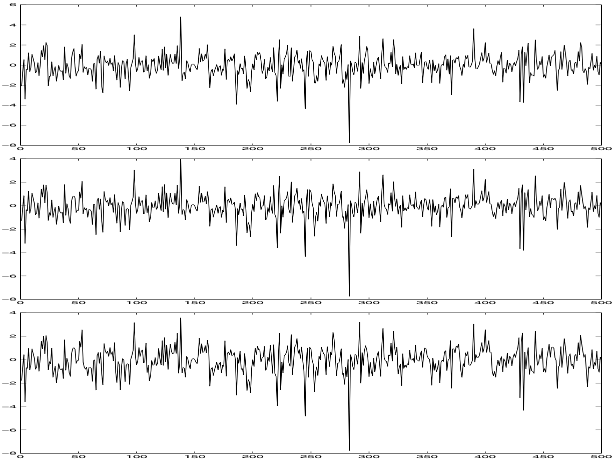


Fig. 1.2 The observed signals that are assumed to be mixtures of some underlying source signals.

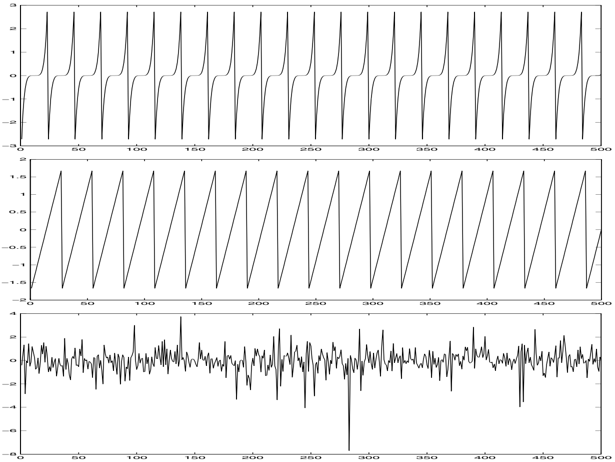


Fig. 1.3 The estimates of the original source signals, estimated using only the observed mixture signals in Fig. 1.2. The original signals were found very accurately.

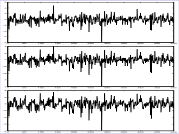


Fig. 1.2 The observed signals that are assumed to be mixtures of some underlying source signals.

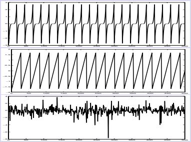


Fig. 1.3 The estimates of the original source signals, estimated using only the observed mixture signals in Fig. 1.2. The original signals were found very accurately.

"Using just this information on the statistical independence, we can in fact estimate the coefficient matrix W for the signals in Fig. 1.2.

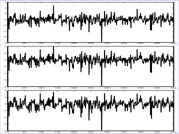


Fig. 1.2 The observed signals that are assumed to be mixtures of some underlying source signals.

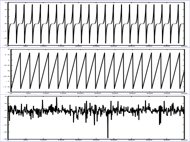


Fig. 1.3 The estimates of the original source signals, estimated using only the observed mixture signals in Fig. 1.2. The original signals were found very accurately.

"Using just this information on the statistical independence, we can in fact estimate the coefficient matrix W for the signals in Fig. 1.2. What we obtain are the source signals in Fig. 1.3. (These signals were estimated by the FastICA algorithm that we shall meet in several chapters of this book.)

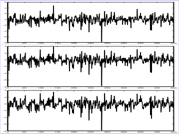


Fig. 1.2 The observed signals that are assumed to be mixtures of some underlying source signals.

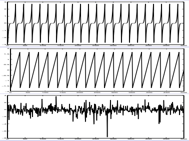


Fig. 1.3 The estimates of the original source signals, estimated using only the observed mixture signals in Fig. 1.2. The original signals were found very accurately.

"Using just this information on the statistical independence, we can in fact estimate the coefficient matrix W for the signals in Fig. 1.2. What we obtain are the source signals in Fig. 1.3. (These signals were estimated by the FastICA algorithm that we shall meet in several chapters of this book.) We see that from a data set that seemed to be just noise, we were able to estimate the original source signals, using an algorithm that used the information on the independence only.

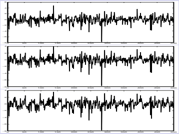


Fig. 1.2 The observed signals that are assumed to be mixtures of some underlying source signals.

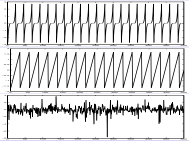


Fig. 1.3 The estimates of the original source signals, estimated using only the observed mixture signals in Fig. 1.2. The original signals were found very accurately.

"Using just this information on the statistical independence, we can in fact estimate the coefficient matrix W for the signals in Fig. 1.2. What we obtain are the source signals in Fig. 1.3. (These signals were estimated by the FastICA algorithm that we shall meet in several chapters of this book.) We see that from a data set that seemed to be just noise, we were able to estimate the original source signals, using an algorithm that used the information on the independence only. These estimated signals are indeed equal to those that were used in creating the mixtures in Fig. 1.2 (the original signals are not shown, but they really are virtually identical to what the algorithm found).

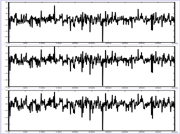


Fig. 1.2 The observed signals that are assumed to be mixtures of some underlying source signals.

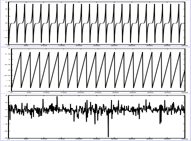


Fig. 1.3 The estimates of the original source signals, estimated using only the observed mixture signals in Fig. 1.2. The original signals were found very accurately.

"Using just this information on the statistical independence, we can in fact estimate the coefficient matrix W for the signals in Fig. 1.2. What we obtain are the source signals in Fig. 1.3. (These signals were estimated by the FastICA algorithm that we shall meet in several chapters of this book.) We see that from a data set that seemed to be just noise, we were able to estimate the original source signals, using an algorithm that used the information on the independence only. These estimated signals are indeed equal to those that were used in creating the mixtures in Fig. 1.2 (the original signals are not shown, but they really are virtually identical to what the algorithm found). Thus, in the source separation problem, the original signals were the "independent components" of the data set."

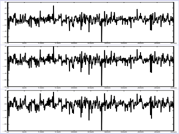


Fig. 1.2 The observed signals that are assumed to be mixtures of some underlying source signals.

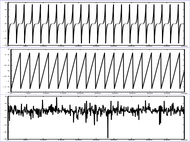


Fig. 1.3 The estimates of the original source signals, estimated using only the observed mixture signals in Fig. 1.2. The original signals were found very accurately.

This leads us to the following definition of ICA [...]

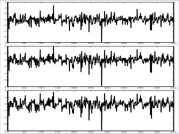


Fig. 1.2 The observed signals that are assumed to be mixtures of some underlying source signals.

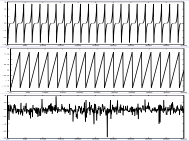


Fig. 1.3 The estimates of the original source signals, estimated using only the observed mixture signals in Fig. 1.2. The original signals were found very accurately.

This leads us to the following definition of ICA [...] Given a set of observations of random variables $(x_1(t), x_2(t), \dots, x_d(t))$*, where* t *is the time or sample index, assume that they are generated as a linear mixture of independent components:*

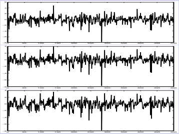


Fig. 1.2 The observed signals that are assumed to be mixtures of some underlying source signals.

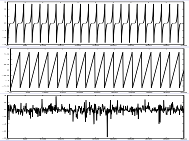


Fig. 1.3 The estimates of the original source signals, estimated using only the observed mixture signals in Fig. 1.2. The original signals were found very accurately.

This leads us to the following definition of ICA [...] Given a set of observations of random variables $(x_1(t), x_2(t), \ldots, x_d(t))$*, where t is the time or sample index, assume that they are generated as a linear mixture of independent components:*

$$\begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_d(t) \end{pmatrix} = A \begin{pmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_d(t) \end{pmatrix},$$

where A is some unknown matrix.

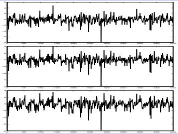


Fig. 1.2 The observed signals that are assumed to be mixtures of some underlying source signals.

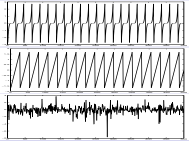


Fig. 1.3 The estimates of the original source signals, estimated using only the observed mixture signals in Fig. 1.2. The original signals were found very accurately.

This leads us to the following definition of ICA [...] Given a set of observations of random variables $(x_1(t), x_2(t), \ldots, x_d(t))$, where t is the time or sample index, assume that they are generated as a linear mixture of independent components:

$$\begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_d(t) \end{pmatrix} = A \begin{pmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_d(t) \end{pmatrix},$$

where A is some unknown matrix. Independent component analysis now consists of estimating both the matrix A and the $s_i(t)$, when we only observe the $x_i(t)$.

Good?

Independence?

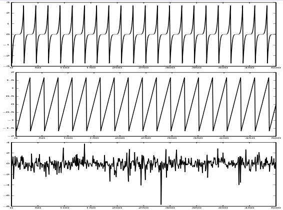


Fig. 1.3 The estimates of the original source signals, estimated using only the observed mixture signals in Fig. 1.2. The original signals were found very accurately.

- Is $s_1(t)$ independent of $s_2(t)$?

Independence?

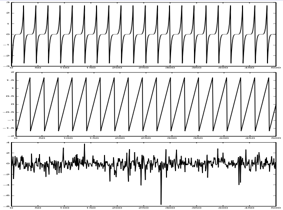


Fig. 1.3 The estimates of the original source signals, estimated using only the observed mixture signals in Fig. 1.2. The original signals were found very accurately.

- Is $s_1(t)$ independent of $s_2(t)$? Sure!

Independence?

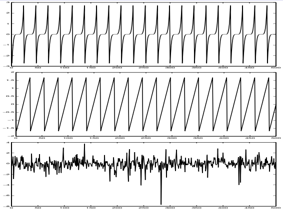


Fig. 1.3 The estimates of the original source signals, estimated using only the observed mixture signals in Fig. 1.2. The original signals were found very accurately.

- Is $s_1(t)$ independent of $s_2(t)$? Sure! How about $s_1(t)$ and $s_3(t)$?

Independence?

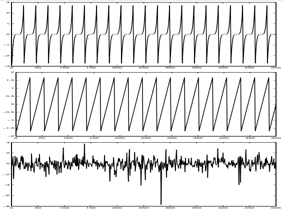


Fig. 1.3 The estimates of the original source signals, estimated using only the observed mixture signals in Fig. 1.2. The original signals were found very accurately.

- Is $s_1(t)$ independent of $s_2(t)$? Sure! How about $s_1(t)$ and $s_3(t)$?
- Any two numbers are independent of each other!

Independence?

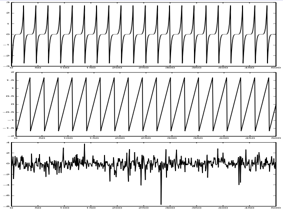


Fig. 1.3 The estimates of the original source signals, estimated using only the observed mixture signals in Fig. 1.2. The original signals were found very accurately.

- Is $s_1(t)$ independent of $s_2(t)$? Sure! How about $s_1(t)$ and $s_3(t)$?
- Any two numbers are independent of each other! All deterministic signal sources are fine then?

Independence?

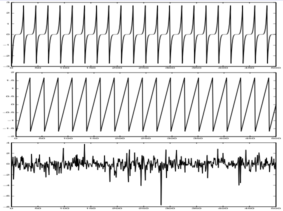


Fig. 1.3 The estimates of the original source signals, estimated using only the observed mixture signals in Fig. 1.2. The original signals were found very accurately.

- Is $s_1(t)$ independent of $s_2(t)$? Sure! How about $s_1(t)$ and $s_3(t)$?
- Any two numbers are independent of each other! All deterministic signal sources are fine then?
- Should we be worried about temporal dependencies?

Independence?

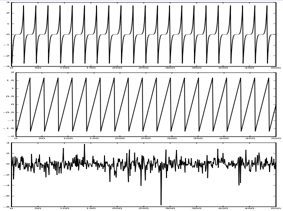


Fig. 1.3 The estimates of the original source signals, estimated using only the observed mixture signals in Fig. 1.2. The original signals were found very accurately.

- Is $s_1(t)$ independent of $s_2(t)$? Sure! How about $s_1(t)$ and $s_3(t)$?
- Any two numbers are independent of each other! All deterministic signal sources are fine then?
- Should we be worried about temporal dependencies? No?

Independence?

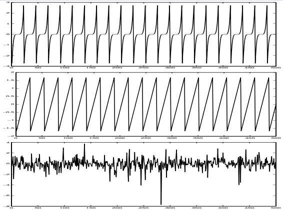


Fig. 1.3 The estimates of the original source signals, estimated using only the observed mixture signals in Fig. 1.2. The original signals were found very accurately.

- Is $s_1(t)$ independent of $s_2(t)$? Sure! How about $s_1(t)$ and $s_3(t)$?
- Any two numbers are independent of each other! All deterministic signal sources are fine then?
- Should we be worried about temporal dependencies? No? What if $s_1(t) = s_1(t+1) = \ldots$?

Independence?

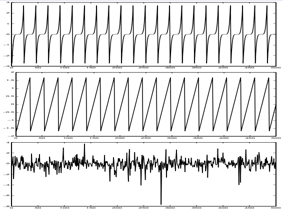


Fig. 1.3 The estimates of the original source signals, estimated using only the observed mixture signals in Fig. 1.2. The original signals were found very accurately.

- Is $s_1(t)$ independent of $s_2(t)$? Sure! How about $s_1(t)$ and $s_3(t)$?
- Any two numbers are independent of each other! All deterministic signal sources are fine then?
- Should we be worried about temporal dependencies? No? What if $s_1(t) = s_1(t+1) = \dots$?
- What is ICA then???

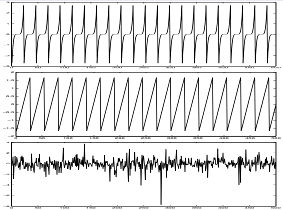


Fig. 1.3 The estimates of the original source signals, estimated using only the observed mixture signals in Fig. 1.2. The original signals were found very accurately.

- Is $s_1(t)$ independent of $s_2(t)$? Sure! How about $s_1(t)$ and $s_3(t)$?
- Any two numbers are independent of each other! All deterministic signal sources are fine then?
- Should we be worried about temporal dependencies? No? What if $s_1(t) = s_1(t+1) = \ldots$?
- What is ICA then???
- Leon Bottou: Can't reduce everything to statistics!

Independence?

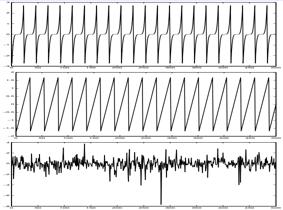


Fig. 1.3 The estimates of the original source signals, estimated using only the observed mixture signals in Fig. 1.2. The original signals were found very accurately.

- Is $s_1(t)$ independent of $s_2(t)$? Sure! How about $s_1(t)$ and $s_3(t)$?
- Any two numbers are independent of each other! All deterministic signal sources are fine then?
- Should we be worried about temporal dependencies? No? What if $s_1(t) = s_1(t+1) = \dots$?
- What is ICA then???
- Leon Bottou: Can't reduce everything to statistics!
- Here: Let's go beyond statistics!

How to go beyond statistical analysis?

1. Perform a deterministic analysis of the algorithm, reducing the problem to perturbation analysis

How to go beyond statistical analysis?

1. Perform a deterministic analysis of the algorithm, reducing the problem to perturbation analysis
2. Perform statistical analysis on the size of perturbations when necessary or desired

How to go beyond statistical analysis?

1. Perform a deterministic analysis of the algorithm, reducing the problem to perturbation analysis
2. Perform statistical analysis on the size of perturbations when necessary or desired

History

How to go beyond statistical analysis?

1. Perform a deterministic analysis of the algorithm, reducing the problem to perturbation analysis
2. Perform statistical analysis on the size of perturbations when necessary or desired

History

- Online learning (adversarial, regret analysis of learning algorithms) Cesa-Bianchi and Lugosi [2006]

How to go beyond statistical analysis?

1. Perform a deterministic analysis of the algorithm, reducing the problem to perturbation analysis
2. Perform statistical analysis on the size of perturbations when necessary or desired

History

- Online learning (adversarial, regret analysis of learning algorithms) Cesa-Bianchi and Lugosi [2006]
- Regression analysis: Vito et al. [2006]

How to go beyond statistical analysis?

1. Perform a deterministic analysis of the algorithm, reducing the problem to perturbation analysis
2. Perform statistical analysis on the size of perturbations when necessary or desired

History

- Online learning (adversarial, regret analysis of learning algorithms) Cesa-Bianchi and Lugosi [2006]
- Regression analysis: Vito et al. [2006]
- Value function estimation in RL: Pires and Szepesvári [2012]

How to go beyond statistical analysis?

1. Perform a deterministic analysis of the algorithm, reducing the problem to perturbation analysis
2. Perform statistical analysis on the size of perturbations when necessary or desired

History

- Online learning (adversarial, regret analysis of learning algorithms) Cesa-Bianchi and Lugosi [2006]
- Regression analysis: Vito et al. [2006]
- Value function estimation in RL: Pires and Szepesvári [2012]
- ICA: This work

Outline

1. Introduction
   - What is ICA, really?
   - Deterministic ICA
   - Has not this been done previously?
2. Results
   - Our algorithm: Deterministic ICA
3. Empirical Illustration
4. Conclusions

What signals can be separated?

What signals can be separated?

- Wrong question!

What signals can be separated?

- Wrong question!
- Better: To what extent can we separate the mixture of some signals?

What signals can be separated?

- Wrong question!
- Better: To what extent can we separate the mixture of some signals?

Let $[T] = \{1, \dots, T\}$. Sources: $s : [T] \to [-C, C]^d$.

What signals can be separated?

- Wrong question!
- Better: To what extent can we separate the mixture of some signals?

Let $[T] = \{1, \ldots, T\}$. Sources: $s : [T] \to [-C, C]^d$.
Let $\nu_T^{(s)}$ be the empirical distribution induced by s; for $B \subset [-C, C]^d$,

$$\nu_T^{(s)}(B) = \frac{1}{T}|\{t \in [T] : s(\tau) \in B\}|.$$

What signals can be separated?

- Wrong question!
- Better: To what extent can we separate the mixture of some signals?

Let $[T] = \{1, \dots, T\}$. Sources: $s : [T] \to [-C, C]^d$.
Let $\nu_T^{(s)}$ be the empirical distribution induced by s; for $B \subset [-C, C]^d$,

$$\nu_T^{(s)}(B) = \frac{1}{T}|\{t \in [T] : s(\tau) \in B\}|.$$

Measure of independence

$$D_4 = \inf_{\mu} \sup_{f \in \mathcal{F}} \left| \int f(s) d\nu_T^{(s)} - \int f(s) d\mu(s) \right|,$$

$\mathcal{F}$: the set of all monomials up to degree 4;
μ: any product measure.

What signals can be separated?

- Wrong question!
- Better: To what extent can we separate the mixture of some signals?

Let $[T] = \{1, \ldots, T\}$. Sources: $s : [T] \to [-C, C]^d$.
Let $\nu_T^{(s)}$ be the empirical distribution induced by s; for $B \subset [-C, C]^d$,

$$\nu_T^{(s)}(B) = \tfrac{1}{T}|\{t \in [T] : s(\tau) \in B\}|.$$

Measure of independence

$$D_4 = \inf_{\mu} \sup_{f \in \mathcal{F}} \left| \int f(s) d\nu_T^{(s)} - \int f(s) d\mu(s) \right|,$$

$\mathcal{F}$: the set of all monomials up to degree 4;
μ: any product measure.

Note: When $s(t)$ has independent components and $s(1), \ldots, s(T)$ are iid, $D_4 = O(1/\sqrt{T})$.

Goal

Assuming $x(t) = As(t) + \epsilon(t),\ t = 1, \ldots, T$;
$\epsilon(t)$ "noise".

Can we find a method with the following characteristics?

- Universal: No free parameters!

Goal

Assuming $x(t) = As(t) + \epsilon(t)$, $t = 1, \ldots, T$;
$\epsilon(t)$ "noise".

Can we find a method with the following characteristics?

- ♢ Universal: No free parameters!
- ♢ Efficient: $\mathrm{poly}(T, d)$ runtime (no dependence on C, A, $\ldots$);

Goal

Assuming $x(t) = As(t) + \epsilon(t)$, $t = 1, \ldots, T$;
$\epsilon(t)$ "noise".

Can we find a method with the following characteristics?

- Universal: No free parameters!
- Efficient: $\mathrm{poly}(T, d)$ runtime (no dependence on C, A, $\ldots$);
- Robust & noise-tolerant: $\mathrm{lin}(D_4 + 1/\sqrt{T})$ accurate in recovering A (with i.i.d. observation noise)?

Goal

Assuming $x(t) = As(t) + \epsilon(t)$, $t = 1, \dots, T$;
$\epsilon(t)$ "noise".

Can we find a method with the following characteristics?

- ♢ Universal: No free parameters!
- ♢ Efficient: $\mathrm{poly}(T, d)$ runtime (no dependence on C, A, $\dots$);
- ♢ Robust & noise-tolerant: $\mathrm{lin}(D_4 + 1/\sqrt{T})$ accurate in recovering A (with i.i.d. observation noise)?

Accuracy measure:

$$d(\hat{A}, A) = \inf_{\substack{\pi \in \mathrm{Perm}([d]) \\ c \in \mathbb{R}^d}} \max_k ||c_k A_{:\pi(k)} - A_{:k}||_2 \,.$$

Outline

Previous works with theoretical guarantees

Samarov and Tsybakov [2004], Chen and Bickel [2006] and others:

Previous works with theoretical guarantees

Samarov and Tsybakov [2004], Chen and Bickel [2006] and others:

- No noise

Previous works with theoretical guarantees

Samarov and Tsybakov [2004], Chen and Bickel [2006] and others:

- No noise
- Semi-parametrics, "average derivative estimation";

Previous works with theoretical guarantees

Samarov and Tsybakov [2004], Chen and Bickel [2006] and others:

- No noise
- Semi-parametrics, "average derivative estimation";
- Asymptotics: $\sqrt{n}$ consistency and efficiency;

Previous works with theoretical guarantees

Samarov and Tsybakov [2004], Chen and Bickel [2006] and others:

- No noise
- Semi-parametrics, "average derivative estimation";
- Asymptotics: $\sqrt{n}$ consistency and efficiency;
- Why estimate something that is then thrown away? (Also: We'll need conditions on these. . .)

Previous works with theoretical guarantees

Samarov and Tsybakov [2004], Chen and Bickel [2006] and others:

- No noise
- Semi-parametrics, "average derivative estimation";
- Asymptotics: $\sqrt{n}$ consistency and efficiency;
- Why estimate something that is then thrown away? (Also: We'll need conditions on these. . .)

FastICA by Hyvärinen [1999]:

Previous works with theoretical guarantees

Samarov and Tsybakov [2004], Chen and Bickel [2006] and others:

- No noise
- Semi-parametrics, "average derivative estimation";
- Asymptotics: $\sqrt{n}$ consistency and efficiency;
- Why estimate something that is then thrown away? (Also: We'll need conditions on these. . .)

FastICA by Hyvärinen [1999]:

- Perhaps the most popular ICA algorithm.

Previous works with theoretical guarantees

Samarov and Tsybakov [2004], Chen and Bickel [2006] and others:

- No noise
- Semi-parametrics, "average derivative estimation";
- Asymptotics: $\sqrt{n}$ consistency and efficiency;
- Why estimate something that is then thrown away? (Also: We'll need conditions on these...)

FastICA by Hyvärinen [1999]:

- Perhaps the most popular ICA algorithm.
- With probability 1, all local optimizers are desired solutions [Wei, 2014], given:
  - infinitely many noiseless samples;
  - using kurtosis as the scoring function.

Previous works with theoretical guarantees

Samarov and Tsybakov [2004], Chen and Bickel [2006] and others:

- No noise
- Semi-parametrics, "average derivative estimation";
- Asymptotics: $\sqrt{n}$ consistency and efficiency;
- Why estimate something that is then thrown away? (Also: We'll need conditions on these. . .)

FastICA by Hyvärinen [1999]:

- Perhaps the most popular ICA algorithm.
- With probability 1, all local optimizers are desired solutions [Wei, 2014], given:
  - infinitely many noiseless samples;
  - using kurtosis as the scoring function.
- Weakness: noisy periodic signals;

Previous works with theoretical guarantees

Samarov and Tsybakov [2004], Chen and Bickel [2006] and others:

- No noise
- Semi-parametrics, "average derivative estimation";
- Asymptotics: $\sqrt{n}$ consistency and efficiency;
- Why estimate something that is then thrown away? (Also: We'll need conditions on these...)

FastICA by Hyvärinen [1999]:

- Perhaps the most popular ICA algorithm.
- With probability 1, all local optimizers are desired solutions [Wei, 2014], given:
  - infinitely many noiseless samples;
  - using kurtosis as the scoring function.
- Weakness: noisy periodic signals;

Moment methods: [Frieze et al., 1996, Hsu and Kakade, 2012, Arora et al., 2012, Goyal et al., 2014].

Moment methods

- Frieze et al. [1996]:

Moment methods

- Frieze et al. [1996]:
  - No noise allowed (also: minor gap fixed by Arora et al. [2012]);

Moment methods

- Frieze et al. [1996]:
  - No noise allowed (also: minor gap fixed by Arora et al. [2012]);
- Hsu and Kakade [2012] ("HKICA"):

Moment methods

- Frieze et al. [1996]:
  - No noise allowed (also: minor gap fixed by Arora et al. [2012]);
- Hsu and Kakade [2012] ("HKICA"):
  - No theoretical guarantee stated;

Moment methods

- Frieze et al. [1996]:
  - No noise allowed (also: minor gap fixed by Arora et al. [2012]);
- Hsu and Kakade [2012] ("HKICA"):
  - No theoretical guarantee stated;
  - If we do the analysis, accuracy will depend on γ_A; an uncontrolled parameter (see later).

Moment methods

- Frieze et al. [1996]:
  - No noise allowed (also: minor gap fixed by Arora et al. [2012]);
- Hsu and Kakade [2012] ("HKICA"):
  - No theoretical guarantee stated;
  - If we do the analysis, accuracy will depend on γ_A; an uncontrolled parameter (see later).
- Arora et al. [2012]:

Moment methods

- Frieze et al. [1996]:
  - No noise allowed (also: minor gap fixed by Arora et al. [2012]);
- Hsu and Kakade [2012] ("HKICA"):
  - No theoretical guarantee stated;
  - If we do the analysis, accuracy will depend on γ_A; an uncontrolled parameter (see later).
- Arora et al. [2012]:
  - Free parameter "β", whose choice depends on $||A_{:i}||_2$, while $x \in \{-1, +1\}^d$.

Moment methods

- Frieze et al. [1996]:
  - No noise allowed (also: minor gap fixed by Arora et al. [2012]);
- Hsu and Kakade [2012] ("HKICA"):
  - No theoretical guarantee stated;
  - If we do the analysis, accuracy will depend on γ_A; an uncontrolled parameter (see later).
- Arora et al. [2012]:
  - Free parameter "β", whose choice depends on $||A_{:i}||_2$, while $x \in \{-1, +1\}^d$.
  - Choose either the scale of sources or the scale of the columns of A, not both!

Moment methods

- Frieze et al. [1996]:
  - No noise allowed (also: minor gap fixed by Arora et al. [2012]);
- Hsu and Kakade [2012] ("HKICA"):
  - No theoretical guarantee stated;
  - If we do the analysis, accuracy will depend on γ_A; an uncontrolled parameter (see later).
- Arora et al. [2012]:
  - Free parameter "β", whose choice depends on $||A_{:i}||_2$, while $x \in \{-1, +1\}^d$.
  - Choose either the scale of sources or the scale of the columns of A, not both!
- Fourier PCA [Goyal et al., 2014] (FPCA):

Moment methods

- Frieze et al. [1996]:
  - No noise allowed (also: minor gap fixed by Arora et al. [2012]);
- Hsu and Kakade [2012] ("HKICA"):
  - No theoretical guarantee stated;
  - If we do the analysis, accuracy will depend on γ_A; an uncontrolled parameter (see later).
- Arora et al. [2012]:
  - Free parameter "β", whose choice depends on $||A_{:i}||_2$, while $x \in \{-1, +1\}^d$.
  - Choose either the scale of sources or the scale of the columns of A, not both!
- Fourier PCA [Goyal et al., 2014] (FPCA):
  - Free parameter: Same problem as with Arora et al. [2012]'s approach.

Moment methods

- Frieze et al. [1996]:
  - No noise allowed (also: minor gap fixed by Arora et al. [2012]);
- Hsu and Kakade [2012] ("HKICA"):
  - No theoretical guarantee stated;
  - If we do the analysis, accuracy will depend on γ_A; an uncontrolled parameter (see later).
- Arora et al. [2012]:
  - Free parameter "β", whose choice depends on $||A_{:i}||_2$, while $x \in \{-1, +1\}^d$.
  - Choose either the scale of sources or the scale of the columns of A, not both!
- Fourier PCA [Goyal et al., 2014] (FPCA):
  - Free parameter: Same problem as with Arora et al. [2012]'s approach.

Why care about a free parameters?

Moment methods

- Frieze et al. [1996]:
  - No noise allowed (also: minor gap fixed by Arora et al. [2012]);
- Hsu and Kakade [2012] ("HKICA"):
  - No theoretical guarantee stated;
  - If we do the analysis, accuracy will depend on γ_A; an uncontrolled parameter (see later).
- Arora et al. [2012]:
  - Free parameter "β", whose choice depends on $||A_{:i}||_2$, while $x \in \{-1, +1\}^d$.
  - Choose either the scale of sources or the scale of the columns of A, not both!
- Fourier PCA [Goyal et al., 2014] (FPCA):
  - Free parameter: Same problem as with Arora et al. [2012]'s approach.

Why care about a free parameters? ... unsupervised learning

Outline

Result

There exist a randomized method to estimate A from $(x(t))_{t=1}^{T}$, with $x(t) = As(t) + \epsilon(t),\ t = 1, \ldots, T$ s.t.:

- The computational complexity is $O(d^3 T)$;

Result

There exist a randomized method to estimate A from $(x(t))_{t=1}^{T}$, with $x(t) = As(t) + \epsilon(t),\ t = 1, \ldots, T$ s.t.:

- The computational complexity is $O(d^3 T)$;
- With high probability, the output $\hat{A}$ satisfies

$$d(\hat{A}, A) \leq \theta_1 \min(D_4 + 1/\sqrt{T}, \theta_2)\,.$$

Result

There exist a randomized method to estimate A from $(x(t))_{t=1}^{T}$, with $x(t) = As(t) + \epsilon(t)$, $t = 1, \ldots, T$ s.t.:

- The computational complexity is $O(d^3 T)$;
- With high probability, the output $\hat{A}$ satisfies

$$d(\hat{A}, A) \leq \theta_1 \min(D_4 + 1/\sqrt{T}, \theta_2)\,.$$

- Here, θ_1, θ_2 problem dependent, polynomial in the parameters.

Outline

Hsu-Kakade method (HKICA)

For $x \sim \mu = \mu_1 \otimes \cdots \otimes \mu_d$, $\eta \in \mathbb{R}^d$, let

$$f(\eta) = \mathbb{E}\left[(\eta^\top x)^4\right] - 3\mathbb{E}\left[(\eta^\top x)^2\right]^2 .$$

Hsu-Kakade method (HKICA)

For $x \sim \mu = \mu_1 \otimes \cdots \otimes \mu_d$, $\eta \in \mathbb{R}^d$, let

$$f(\eta) = \mathbb{E}\left[(\eta^\top x)^4\right] - 3\mathbb{E}\left[(\eta^\top x)^2\right]^2 .$$

- Choose ϕ and ψ. *(How?)*

Hsu-Kakade method (HKICA)

For $x \sim \mu = \mu_1 \otimes \cdots \otimes \mu_d$, $\eta \in \mathbb{R}^d$, let

$$f(\eta) = \mathbb{E}\left[(\eta^\top x)^4\right] - 3\mathbb{E}\left[(\eta^\top x)^2\right]^2 .$$

- Choose ϕ and ψ. *(How?)*
- Let $T(\phi) = \nabla^2 f(\phi)$. Then

$$T(\phi) = AK \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_d \end{pmatrix} A^\top,$$

where $\sigma_i = \left(\phi^\top A_i\right)^2$ and K is some diagonal matrix.

Hsu-Kakade method (HKICA)

For $x \sim \mu = \mu_1 \otimes \cdots \otimes \mu_d$, $\eta \in \mathbb{R}^d$, let

$$f(\eta) = \mathbb{E}\left[(\eta^\top x)^4\right] - 3\mathbb{E}\left[(\eta^\top x)^2\right]^2 .$$

- Choose ϕ and ψ. *(How?)*
- Let $T(\phi) = \nabla^2 f(\phi)$. Then

$$T(\phi) = AK \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_d \end{pmatrix} A^\top,$$

  where $\sigma_i = \left(\phi^\top A_i\right)^2$ and K is some diagonal matrix.
- Let $M = T(\phi)(T(\psi))^{-1}$.

Hsu-Kakade method (HKICA)

For $x \sim \mu = \mu_1 \otimes \cdots \otimes \mu_d$, $\eta \in \mathbb{R}^d$, let

$$f(\eta) = \mathbb{E}\left[(\eta^\top x)^4\right] - 3\mathbb{E}\left[(\eta^\top x)^2\right]^2 .$$

- Choose ϕ and ψ. *(How?)*
- Let $T(\phi) = \nabla^2 f(\phi)$. Then

$$T(\phi) = AK \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_d \end{pmatrix} A^\top,$$

  where $\sigma_i = \left(\phi^\top A_i\right)^2$ and K is some diagonal matrix.
- Let $M = T(\phi)(T(\psi))^{-1}$.
- Then

$$M = A \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{pmatrix} A^{-1},$$

  where $\lambda_i = \left(\frac{\phi^\top A_i}{\psi^\top A_i}\right)^2$.

Hsu-Kakade method (HKICA)

For $x \sim \mu = \mu_1 \otimes \cdots \otimes \mu_d$, $\eta \in \mathbb{R}^d$, let

$$f(\eta) = \mathbb{E}\left[(\eta^\top x)^4\right] - 3\mathbb{E}\left[(\eta^\top x)^2\right]^2 .$$

- Choose ϕ and ψ. *(How?)*
- Let $T(\phi) = \nabla^2 f(\phi)$. Then

$$T(\phi) = AK \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_d \end{pmatrix} A^\top,$$

  where $\sigma_i = \left(\phi^\top A_i\right)^2$ and K is some diagonal matrix.
- Let $M = T(\phi)(T(\psi))^{-1}$.
- Then

$$M = A \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{pmatrix} A^{-1},$$

  where $\lambda_i = \left(\frac{\phi^\top A_i}{\psi^\top A_i}\right)^2$.
- Do an eigen-decomposition of M to recover A, assuming all λ_i's are distinct.

Hsu-Kakade method (HKICA)

For $x \sim \mu = \mu_1 \otimes \cdots \otimes \mu_d$, $\eta \in \mathbb{R}^d$, let

$$f(\eta) = \mathbb{E}\left[(\eta^\top x)^4\right] - 3\mathbb{E}\left[(\eta^\top x)^2\right]^2 .$$

- Choose ϕ and ψ. *(How?)*
- Let $T(\phi) = \nabla^2 f(\phi)$. Then

$$T(\phi) = AK \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_d \end{pmatrix} A^\top,$$

  where $\sigma_i = \left(\phi^\top A_i\right)^2$ and K is some diagonal matrix.
- Let $M = T(\phi)(T(\psi))^{-1}$.
- Then

$$M = A \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{pmatrix} A^{-1},$$

  where $\lambda_i = \left(\frac{\phi^\top A_i}{\psi^\top A_i}\right)^2$.
- Do an eigen-decomposition of M to recover A, assuming all λ_i's are distinct.

Given finite sample $(x(t))_{t=1}^T$, replace $\mathbb{E}[\cdot]$ with $\mathbb{E}_n[\cdot]$.

Our issue with the HK method

Problem: Minimal gap of the eigenvalues.

- Theoretical analysis shows that the performance depends on γ_A^{-1}, where
$$\gamma_A = \min_{i \neq j} |\lambda_i - \lambda_j| \,.$$

[1] Is γ_A^{-1} polynomially bounded in d?

Our issue with the HK method

Problem: Minimal gap of the eigenvalues.

- Theoretical analysis shows that the performance depends on γ_A^{-1}, where

$$\gamma_A = \min_{i \neq j} |\lambda_i - \lambda_j| .$$

- γ_A is not yet well understood[1]

[1] Is γ_A^{-1} polynomially bounded in d?

Our issue with the HK method

Problem: Minimal gap of the eigenvalues.

- Theoretical analysis shows that the performance depends on γ_A^{-1}, where

$$\gamma_A = \min_{i \neq j} |\lambda_i - \lambda_j| .$$

- γ_A is not yet well understood[1]

Goal: Avoid dependency on γ_A^{-1}!

[1] Is γ_A^{-1} polynomially bounded in d?

Our method: Deterministic ICA (DICA)

– Inspired by Arora et al. [2012] and Frieze et al. [1996];

2

Our method: Deterministic ICA (DICA)

- Inspired by Arora et al. [2012] and Frieze et al. [1996];
- Sample ψ, ϕ_1 and ϕ_2 independently from standard normal distribution.

2

Our method: Deterministic ICA (DICA)

- Inspired by Arora et al. [2012] and Frieze et al. [1996];
- Sample ψ, ϕ_1 and ϕ_2 independently from standard normal distribution.
- Calculate $\nabla^2 f(\psi)$ and B such that $\nabla^2 f(\psi) = BB^\top$.[2]

2

Our method: Deterministic ICA (DICA)

- Inspired by Arora et al. [2012] and Frieze et al. [1996];
- Sample ψ, ϕ_1 and ϕ_2 independently from standard normal distribution.
- Calculate $\nabla^2 f(\psi)$ and B such that $\nabla^2 f(\psi) = BB^\top$.[2]

[2] Recall: $f(\eta) = \mathbb{E}\left[(\eta^\top x)^4\right] - 3\mathbb{E}\left[(\eta^\top x)^2\right]^2$

Our method: Deterministic ICA (DICA)

- Inspired by Arora et al. [2012] and Frieze et al. [1996];
- Sample ψ, ϕ_1 and ϕ_2 independently from standard normal distribution.
- Calculate $\nabla^2 f(\psi)$ and B such that $\nabla^2 f(\psi) = BB^\top$.[2]
- Calculate $T(\phi_1) = \nabla^2 f(B^{-\top}\phi_1)$ and $T(\phi_2) = \nabla^2 f(B^{-\top}\phi_2)$.

[2] Recall: $f(\eta) = \mathbb{E}\left[(\eta^\top x)^4\right] - 3\mathbb{E}\left[(\eta^\top x)^2\right]^2$

Our method: Deterministic ICA (DICA)

- Inspired by Arora et al. [2012] and Frieze et al. [1996];
- Sample ψ, ϕ_1 and ϕ_2 independently from standard normal distribution.
- Calculate $\nabla^2 f(\psi)$ and B such that $\nabla^2 f(\psi) = BB^\top$.[2]
- Calculate $T(\phi_1) = \nabla^2 f(B^{-\top}\phi_1)$ and $T(\phi_2) = \nabla^2 f(B^{-\top}\phi_2)$.
- Calculate $M = T(\phi_1)(T(\phi_2))^{-1}$.

$$M = R\,\mathrm{diag}\left(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_d\right) R^\top,$$

where $\tilde{\lambda}_i = \left(\frac{\phi_1^\top R_i}{\phi_2^\top R_i}\right)^2$ and R is some orthonormal matrix s.t. $A = BR$.

[2]Recall: $f(\eta) = \mathbb{E}\left[(\eta^\top x)^4\right] - 3\mathbb{E}\left[(\eta^\top x)^2\right]^2$

Our method: Deterministic ICA (DICA)

- Inspired by Arora et al. [2012] and Frieze et al. [1996];
- Sample ψ, ϕ_1 and ϕ_2 independently from standard normal distribution.
- Calculate $\nabla^2 f(\psi)$ and B such that $\nabla^2 f(\psi) = BB^\top$.[2]
- Calculate $T(\phi_1) = \nabla^2 f(B^{-\top}\phi_1)$ and $T(\phi_2) = \nabla^2 f(B^{-\top}\phi_2)$.
- Calculate $M = T(\phi_1)(T(\phi_2))^{-1}$.

$$M = R\,\mathrm{diag}\left(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_d\right) R^\top,$$

where $\tilde{\lambda}_i = \left(\frac{\phi_1^\top R_i}{\phi_2^\top R_i}\right)^2$ and R is some orthonormal matrix s.t. $A = BR$.

- Do an eigen-decomposition of M to recover R.

[2] Recall: $f(\eta) = \mathbb{E}\left[(\eta^\top x)^4\right] - 3\mathbb{E}\left[(\eta^\top x)^2\right]^2$

Our method: Deterministic ICA (DICA)

- Inspired by Arora et al. [2012] and Frieze et al. [1996];
- Sample ψ, ϕ_1 and ϕ_2 independently from standard normal distribution.
- Calculate $\nabla^2 f(\psi)$ and B such that $\nabla^2 f(\psi) = BB^\top$.[2]
- Calculate $T(\phi_1) = \nabla^2 f(B^{-\top}\phi_1)$ and $T(\phi_2) = \nabla^2 f(B^{-\top}\phi_2)$.
- Calculate $M = T(\phi_1)(T(\phi_2))^{-1}$.

$$M = R \operatorname{diag}\left(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_d\right) R^\top,$$

where $\tilde{\lambda}_i = \left(\frac{\phi_1^\top R_i}{\phi_2^\top R_i}\right)^2$ and R is some orthonormal matrix s.t. $A = BR$.
- Do an eigen-decomposition of M to recover R.
- Return $\hat{A} = BR$ as an estimate of A.

[2] Recall: $f(\eta) = \mathbb{E}\left[(\eta^\top x)^4\right] - 3\mathbb{E}\left[(\eta^\top x)^2\right]^2$

Our method: Deterministic ICA (DICA)

- The reconstruction error is proportional to γ_R^{-1}, where

$$\gamma_R = \min_{i \neq j} \left| \tilde{\lambda}_i - \tilde{\lambda}_j \right| = \min_{i \neq j} \left| \left(\frac{\phi_1^\top R_i}{\phi_2^\top R_i} \right)^2 - \left(\frac{\phi_1^\top R_j}{\phi_2^\top R_j} \right)^2 \right| .$$

Our method: Deterministic ICA (DICA)

- The reconstruction error is proportional to γ_R^{-1}, where

$$\gamma_R = \min_{i \neq j} \left| \tilde{\lambda}_i - \tilde{\lambda}_j \right| = \min_{i \neq j} \left| \left(\frac{\phi_1^\top R_i}{\phi_2^\top R_i} \right)^2 - \left(\frac{\phi_1^\top R_j}{\phi_2^\top R_j} \right)^2 \right| .$$

  - $\{\phi_1^\top R_1, \ldots, \phi_1^\top R_d, \phi_2^\top R_1, \ldots, \phi_2^\top R_d\}$ are independent standard normal variables.

Our method: Deterministic ICA (DICA)

- The reconstruction error is proportional to γ_R^{-1}, where

$$\gamma_R = \min_{i \neq j} \left| \tilde{\lambda}_i - \tilde{\lambda}_j \right| = \min_{i \neq j} \left| \left(\frac{\phi_1^\top R_i}{\phi_2^\top R_i} \right)^2 - \left(\frac{\phi_1^\top R_j}{\phi_2^\top R_j} \right)^2 \right|.$$

  - $\{\phi_1^\top R_1, \ldots, \phi_1^\top R_d, \phi_2^\top R_1, \ldots, \phi_2^\top R_d\}$ are independent standard normal variables.
  - No dependency on A.

Our method: Deterministic ICA (DICA)

- The reconstruction error is proportional to γ_R^{-1}, where

$$\gamma_R = \min_{i \neq j} \left| \tilde{\lambda}_i - \tilde{\lambda}_j \right| = \min_{i \neq j} \left| \left(\frac{\phi_1^\top R_i}{\phi_2^\top R_i} \right)^2 - \left(\frac{\phi_1^\top R_j}{\phi_2^\top R_j} \right)^2 \right|.$$

  - $\{\phi_1^\top R_1, \ldots, \phi_1^\top R_d, \phi_2^\top R_1, \ldots, \phi_2^\top R_d\}$ are independent standard normal variables.
  - No dependency on A.
  - Can be reduced to the minimal spacing of Cauchy random variables.

Our method: Deterministic ICA (DICA)

- The reconstruction error is proportional to γ_R^{-1}, where

$$\gamma_R = \min_{i \neq j} \left| \tilde{\lambda}_i - \tilde{\lambda}_j \right| = \min_{i \neq j} \left| \left(\frac{\phi_1^\top R_i}{\phi_2^\top R_i} \right)^2 - \left(\frac{\phi_1^\top R_j}{\phi_2^\top R_j} \right)^2 \right|.$$

  - $\{\phi_1^\top R_1, \ldots, \phi_1^\top R_d, \phi_2^\top R_1, \ldots, \phi_2^\top R_d\}$ are independent standard normal variables.
  - No dependency on A.
  - Can be reduced to the minimal spacing of Cauchy random variables.
- Recursive version is also developed in the paper, based on the idea of Vempala and Xiao [2014].

Outline

Setting

Questions:

Setting

Questions:

- Does noise matter?

Setting

Questions:

- Does noise matter?
  - Additive Gaussian noise with std $= 0.3$.

Setting

Questions:

- Does noise matter?
  - Additive Gaussian noise with std $= 0.3$.
- Does coherence matter?

Setting

Questions:

- Does noise matter?
  - Additive Gaussian noise with std $= 0.3$.
- Does coherence matter?
  - A_1: Gaussian matrix; A_4: more coherent matrix.

Setting

Questions:

- Does noise matter?
  - Additive Gaussian noise with std $= 0.3$.
- Does coherence matter?
  - A_1: Gaussian matrix; A_4: more coherent matrix.
- Do the recursive versions improve the results?

Setting

Questions:

- Does noise matter?
  - Additive Gaussian noise with std $= 0.3$.
- Does coherence matter?
  - A_1: Gaussian matrix; A_4: more coherent matrix.
- Do the recursive versions improve the results?

Algorithms:

Setting

Questions:

- Does noise matter?
  - Additive Gaussian noise with std $= 0.3$.
- Does coherence matter?
  - A_1: Gaussian matrix; A_4: more coherent matrix.
- Do the recursive versions improve the results?

Algorithms:

- FastICA [Szabó et al., 2012];

Setting

Questions:

- Does noise matter?
  - Additive Gaussian noise with std $= 0.3$.
- Does coherence matter?
  - A_1: Gaussian matrix; A_4: more coherent matrix.
- Do the recursive versions improve the results?

Algorithms:

- FastICA [Szabó et al., 2012];
- HKICA/HKICA.R;

Setting

Questions:

- Does noise matter?
  - Additive Gaussian noise with std $= 0.3$.
- Does coherence matter?
  - A_1: Gaussian matrix; A_4: more coherent matrix.
- Do the recursive versions improve the results?

Algorithms:

- FastICA [Szabó et al., 2012];
- HKICA/HKICA.R;
- (recursive) FPCA due to Vempala and Xiao [2014];

Setting

Questions:

- Does noise matter?
  - Additive Gaussian noise with std $= 0.3$.
- Does coherence matter?
  - A_1: Gaussian matrix; A_4: more coherent matrix.
- Do the recursive versions improve the results?

Algorithms:

- FastICA [Szabó et al., 2012];
- HKICA/HKICA.R;
- (recursive) FPCA due to Vempala and Xiao [2014];
- DICA/DICA.R;

Setting

Questions:

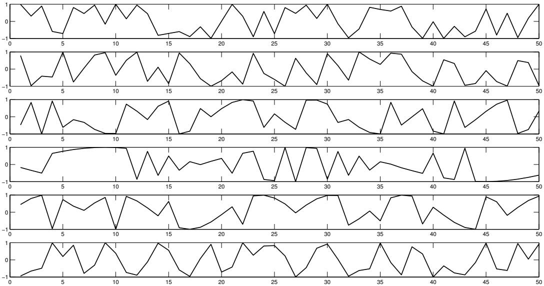
- Does noise matter?
  - Additive Gaussian noise with std $= 0.3$.
- Does coherence matter?
  - A_1: Gaussian matrix; A_4: more coherent matrix.
- Do the recursive versions improve the results?

Algorithms:

- FastICA [Szabó et al., 2012];
- HKICA/HKICA.R;
- (recursive) FPCA due to Vempala and Xiao [2014];
- DICA/DICA.R;
- MDICA/MDICA.R (heuristic version of DICA)

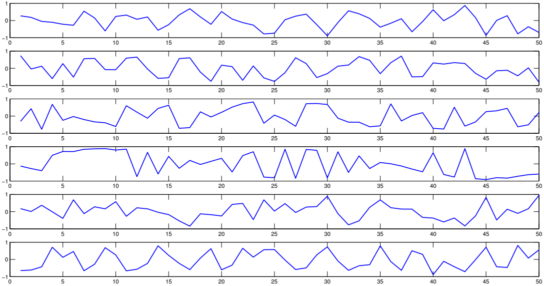
Sources

BPSK sources, $T = 20,000$:

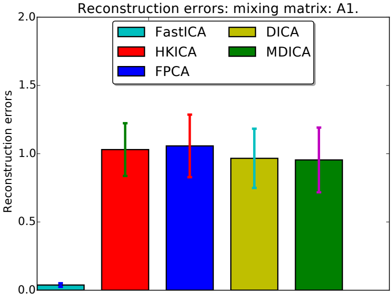


Example reconstructions: Noise-free case

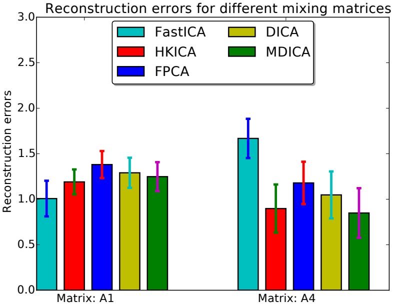
Reconstruction:



Turf: Noise and coherence



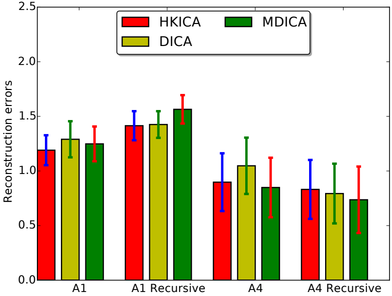
Noise-free



Noisy

Reconstruction error of “random” matrix: 2.2.
Error bars are based on 150 runs.

Recursive versions



Outline

Conclusions and Open Questions

- Independent Component Analysis without probabilities!

Conclusions and Open Questions

- Independent Component Analysis without probabilities!
- Deterministic analysis: Cleaner, more general, should do it more often! Limits?

Conclusions and Open Questions

- Independent Component Analysis without probabilities!
- Deterministic analysis: Cleaner, more general, should do it more often! Limits?
- New method: DICA. Universal, strong guarantees.

Conclusions and Open Questions

- ♢ Independent Component Analysis without probabilities!
- ♢ Deterministic analysis: Cleaner, more general, should do it more often! Limits?
- ♢ New method: DICA. Universal, strong guarantees.
- ♢ In practice, moment-methods are indeed more robust to noise.

Conclusions and Open Questions

- Independent Component Analysis without probabilities!
- Deterministic analysis: Cleaner, more general, should do it more often! Limits?
- New method: DICA. Universal, strong guarantees.
- In practice, moment-methods are indeed more robust to noise.
- Why does coherence help?

Conclusions and Open Questions

- Independent Component Analysis without probabilities!
- Deterministic analysis: Cleaner, more general, should do it more often! Limits?
- New method: DICA. Universal, strong guarantees.
- In practice, moment-methods are indeed more robust to noise.
- Why does coherence help?
- HKICA is great! Why? MDICA is also good. Why?

Conclusions and Open Questions

- ◇ Independent Component Analysis without probabilities!
- ◇ Deterministic analysis: Cleaner, more general, should do it more often! Limits?
- ◇ New method: DICA. Universal, strong guarantees.
- ◇ In practice, moment-methods are indeed more robust to noise.
- ◇ Why does coherence help?
- ◇ HKICA is great! Why? MDICA is also good. Why?
- ◇ "Recursive versions": No significant advantage was found (limited study?).

Conclusions and Open Questions

- Independent Component Analysis without probabilities!
- Deterministic analysis: Cleaner, more general, should do it more often! Limits?
- New method: DICA. Universal, strong guarantees.
- In practice, moment-methods are indeed more robust to noise.
- Why does coherence help?
- HKICA is great! Why? MDICA is also good. Why?
- "Recursive versions": No significant advantage was found (limited study?).

Questions?

Code: `https://github.com/Armstring/Deterministic-ICA`

References

S. Arora, R. Ge, A. Moitra, and S. Sachdeva. Provable ica with unknown gaussian noise, with implications for gaussian mixtures and autoencoders. In *Advances in Neural Information Processing Systems*, pages 2375–2383, 2012.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

A. Chen and P. J. Bickel. Efficient independent component analysis. *Annals of Statistics*, 34(6):2825–2855, 12 2006.

A. Frieze, M. Jerrum, and R. Kannan. Learning linear transformations. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 359–359. IEEE Computer Society, 1996.

N. Goyal, S. Vempala, and Y. Xiao. Fourier pca and robust tensor decomposition. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 584–593. ACM, 2014.

D. Hsu and S. Kakade. Learning Gaussian mixture models: Moment methods and spectral decompositions. *CoRR*, abs/1206.5766, 2012.

A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

B. Pires and C. Szepesvári. Statistical linear estimation with penalized estimators: an application to reinforcement learning. In *ICML*, pages 1535–1542, June 2012.

A. Samarov and A. Tsybakov. Nonparametric independent component analysis. *Bernoulli*, 10(4):565–582, 08 2004.

Z. Szabó, B. Póczos, and A. Lőrincz. Separation theorem for independent subspace analysis and its consequences. *Pattern Recognition*, 45:1782–1791, 2012.

S. S. Vempala and Y. F. Xiao. Max vs min: Independent component analysis with nearly linear sample complexity. *arXiv preprint arXiv:1412.2954*, 2014.

E. D. Vito, L. Rosasco, A. Caponnetto, U. D. Giovannini, and F. Odone. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6(1):883, 2006.

T. Wei. A study of the fixed points and spurious solutions of the fastica algorithm. *arXiv preprint arXiv:1408.6693*, 2014.