Nonlinear ICA using temporal structure: A principled framework for unsupervised deep learning

Aapo Hyvärinen
with Hiroshi Morioka

Gatsby Unit, University College London, UK
Dept of Computer Science, University of Helsinki, Finland

Abstract

- How to extract nonlinear features from multi-dimensional data when there are no labels (unsupervised)?

Abstract

- How to extract nonlinear features from multi-dimensional data when there are no labels (unsupervised)?
- We use temporal structure in time series
  - in two different ways, two different methods

Abstract

- How to extract nonlinear features from multi-dimensional data when there are no labels (unsupervised)?
- We use temporal structure in time series
  - in two different ways, two different methods
- First cases of provably identifiable (well-defined) nonlinear ICA.

Abstract

- How to extract nonlinear features from multi-dimensional data when there are no labels (unsupervised)?
- We use temporal structure in time series
  - in two different ways, two different methods
- First cases of provably identifiable (well-defined) nonlinear ICA.
- A new principled framework for unsupervised deep learning

Background: Towards principled unsupervised learning

- Unsupervised deep learning is a largely unsolved problem

Background: Towards principled unsupervised learning

- Unsupervised deep learning is a largely unsolved problem
- Important because often labelled data costly to obtain

Background: Towards principled unsupervised learning

- Unsupervised deep learning is a largely unsolved problem
- Important because often labelled data costly to obtain
- Probabilistic models with latent variables offer a powerful principled approach

Background: ICA as principled unsupervised learning

- Linear independent component analysis (ICA)

$$x_i(t) = \sum_{j=1}^{n} a_{ij} s_j(t) \quad \text{for all } i, j = 1 \dots n \tag{1}$$

  - $x_i(t)$ is i-th observed signal in time point t
  - a_{ij} constant parameters describing "mixing"
  - Assuming independent, non-Gaussian "sources" s_j
- ICA is identifiable, i.e. well-defined: (Darmois-Skitovich 1950; Comon, 1994)
  - Observing only x_i we can recover both a_{ij} and s_j
  - I.e. original sources can be recovered

BTW, what is the goal in unsupervised learning?

1) Accurate model of data distribution?
   - Evaluate by e.g. Kullback-Leibler divergence

BTW, what is the goal in unsupervised learning?

1) Accurate model of data distribution?
   - Evaluate by e.g. Kullback-Leibler divergence
2) Sampling points from data distribution?
   - Evaluate more or less visually, for images

BTW, what is the goal in unsupervised learning?

1) Accurate model of data distribution?
    - Evaluate by e.g. Kullback-Leibler divergence
2) Sampling points from data distribution?
    - Evaluate more or less visually, for images
3) Useful features for supervised learning?
    - Evaluate e.g. by classification accuracy in benchmark data

BTW, what is the goal in unsupervised learning?

1) Accurate model of data distribution?
   - Evaluate by e.g. Kullback-Leibler divergence
2) Sampling points from data distribution?
   - Evaluate more or less visually, for images
3) Useful features for supervised learning?
   - Evaluate e.g. by classification accuracy in benchmark data
4) Reveal underlying structure in data?
   - Evaluation difficult, e.g. expert opinion

BTW, what is the goal in unsupervised learning?

1) Accurate model of data distribution?
   - Evaluate by e.g. Kullback-Leibler divergence
2) Sampling points from data distribution?
   - Evaluate more or less visually, for images
3) Useful features for supervised learning?
   - Evaluate e.g. by classification accuracy in benchmark data
4) Reveal underlying structure in data?
   - Evaluation difficult, e.g. expert opinion

- These criteria are orthogonal, even contradictory!

BTW, what is the goal in unsupervised learning?

1) Accurate model of data distribution?
   - Evaluate by e.g. Kullback-Leibler divergence
2) Sampling points from data distribution?
   - Evaluate more or less visually, for images
3) Useful features for supervised learning?
   - Evaluate e.g. by classification accuracy in benchmark data
4) Reveal underlying structure in data?
   - Evaluation difficult, e.g. expert opinion

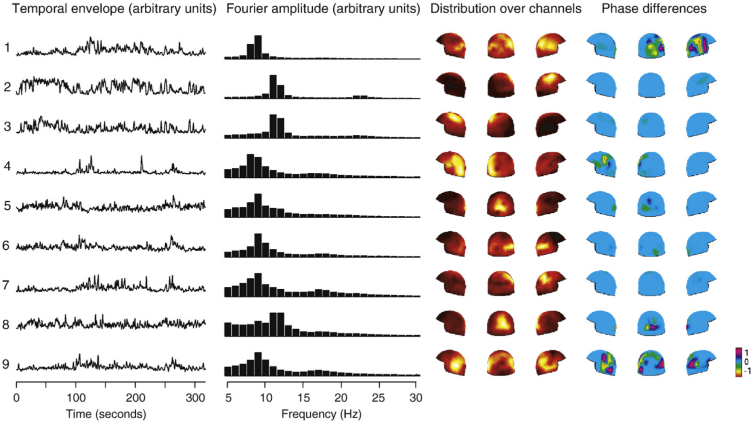
- These criteria are orthogonal, even contradictory!
- 1 & 2 essentially non-parametric problems, 3 & 4 parametric

BTW, what is the goal in unsupervised learning?

1) Accurate model of data distribution?
   - Evaluate by e.g. Kullback-Leibler divergence
2) Sampling points from data distribution?
   - Evaluate more or less visually, for images
3) Useful features for supervised learning?
   - Evaluate e.g. by classification accuracy in benchmark data
4) Reveal underlying structure in data?
   - Evaluation difficult, e.g. expert opinion

- These criteria are orthogonal, even contradictory!
- 1 & 2 essentially non-parametric problems, 3 & 4 parametric
- Goal in ICA (this talk) is 4)

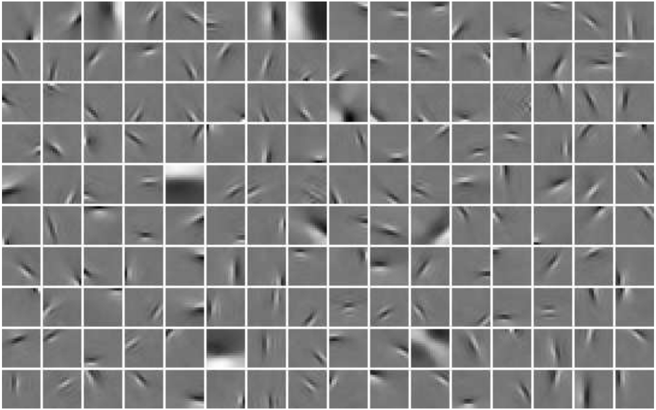
Applications of ICA: Brain source separation



(Hyvärinen, Ramkumar, Parkkonen, Hari, 2010)

Applications of ICA: Image features

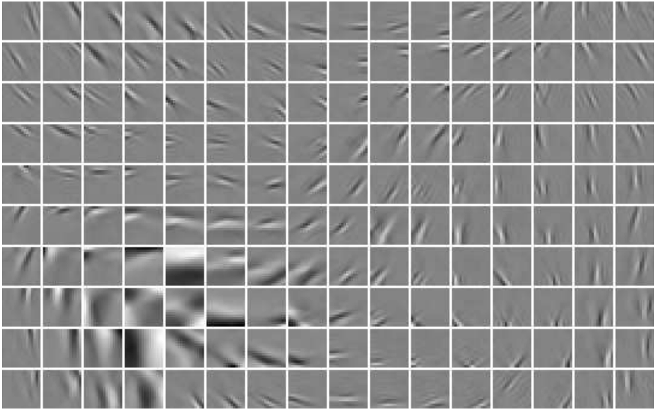
(Olshausen and Field, 1996; Bell and Sejnowski, 1997)



Features similar to wavelets, Gabor functions, simple cells.

Extension: Topographic ICA

(Hyvärinen and Hoyer, 2001)



Topography similar to what is found in the cortex.

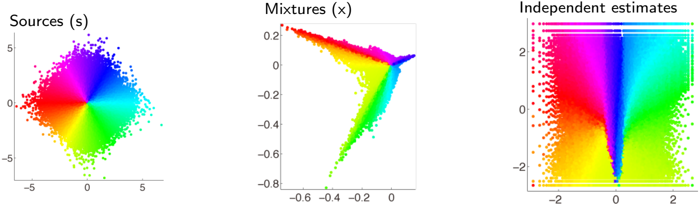
Background: Nonlinear ICA is an unsolved problem

- Extend ICA to nonlinear case to get deep learning?
- Unfortunately, “basic” nonlinear ICA is not identifiable:
- If we define nonlinear ICA model simply as

$$x_i(t) = f_i(s_1(t), \ldots, s_n(t)) \quad \text{for all } i, j = 1 \ldots n \tag{2}$$

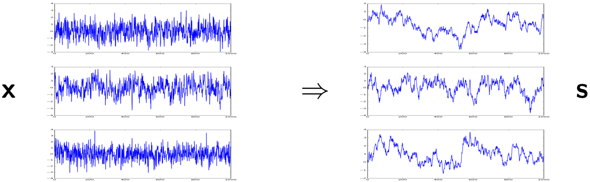
we cannot recover original sources (Darmois, 1952; Hyvärinen & Pajunen, 1999)

  - For any x_1, x_2, we can always find $g(x_1, x_2)$ independent of x_1.
  - Assuming we only consider marginal distribution over time



Background: Temporal correlations help in ICA

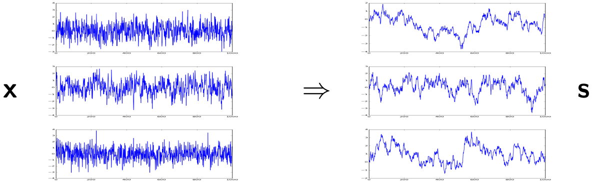
- Harmeling et al (2003) suggested using temporal structure



- Related to finding "slow" features (Földiák, 1991; Wiskott and Sejnowski, 2002)

Background: Temporal correlations help in ICA

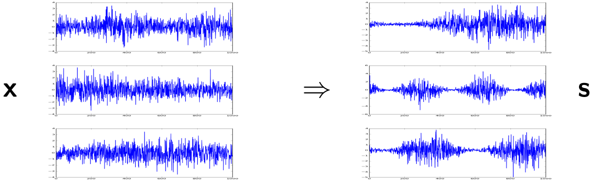
- Harmeling et al (2003) suggested using temporal structure



- Related to finding "slow" features (Földiák, 1991; Wiskott and Sejnowski, 2002)
- Identifiability?
  - Linear: Yes, if autocorrelations distinct for different sources (Tong et al 1991; Belouchrani et al, 1997)
  - Nonlinear: Unknown, although encouraging simulations

Background: Temporal structure as nonstationarity

- An alternative principle for ICA:
  Sources are nonstationary (Matsuoka et al, 2005)



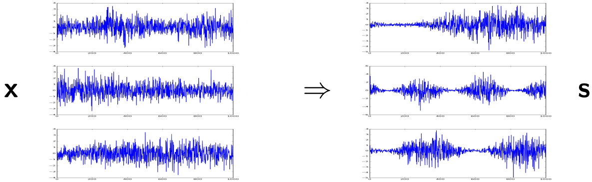
- E.g. variances of the sources can be nonstationary

$$s_i(t) \sim \mathcal{N}(0, \sigma_i^2(t)) \tag{3}$$

- Many data sets have such nonstationarity
  - Video, speech, EEG/MEG, financial time series

Background: Temporal structure as nonstationarity

- An alternative principle for ICA:
  Sources are nonstationary (Matsuoka et al, 2005)



- E.g. variances of the sources can be nonstationary

$$s_i(t) \sim \mathcal{N}(0, \sigma_i^2(t)) \tag{3}$$

- Many data sets have such nonstationarity
  - Video, speech, EEG/MEG, financial time series
- Identifiability?
  - Linear: Yes, no problem (Pham and Cardoso, 2001)
  - Nonlinear: Unknown, almost never attempted

Contributions in this talk

- We present two methods for nonlinear ICA

Contributions in this talk

- We present two methods for nonlinear ICA
- Methods extend linear separation principles above
  - Temporal dependencies
  - Nonstationarity

Contributions in this talk

- We present two methods for nonlinear ICA
- Methods extend linear separation principles above
  - Temporal dependencies
  - Nonstationarity
- We use logistic regression in NN with artificially defined labels
  - Turning unsupervised learning into supervised
  - Cf. noise-contrastive learning, GAN

Contributions in this talk

- We present two methods for nonlinear ICA
- Methods extend linear separation principles above
  - Temporal dependencies
  - Nonstationarity
- We use logistic regression in NN with artificially defined labels
  - Turning unsupervised learning into supervised
  - Cf. noise-contrastive learning, GAN
- Both methods proven to separate nonlinearly mixed sources
- We have constructive proofs of identifiability for nonlinear ICA

Method I: Time-contrastive learning (NIPS2016)

Outline

- We learn features that enable discriminating data points from different time segments

Method I: Time-contrastive learning (NIPS2016)

Outline

- We learn features that enable discriminating data points from different time segments
- We use ordinary neural network training:
  Last hidden layer gives the features

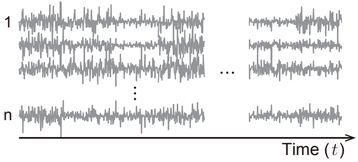
Method I: Time-contrastive learning (NIPS2016)

Outline

- We learn features that enable discriminating data points from different time segments
- We use ordinary neural network training:
  Last hidden layer gives the features
- Surprising theoretical result:
  Estimates a nonlinear ICA model
  - with general nonlinear mixing $\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t))$.
  - nonstationary components $s_i(t)$

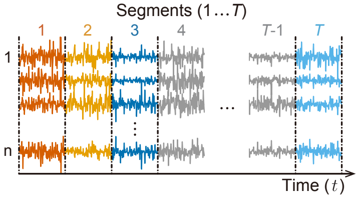
Time-contrastive learning: Definition

- Observe n-dim time series $\mathbf{x}(t)$



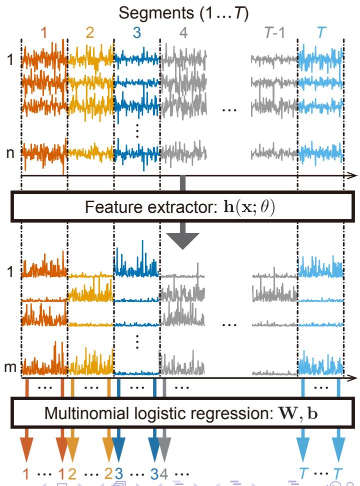
Time-contrastive learning: Definition

- Observe n-dim time series $\mathbf{x}(t)$
- Divide $\mathbf{x}(t)$ into T segments (e.g. bins with equal sizes)



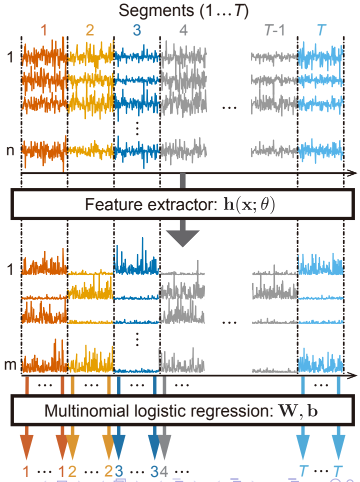
Time-contrastive learning: Definition

- Observe n-dim time series $\mathbf{x}(t)$
- Divide $\mathbf{x}(t)$ into T segments (e.g. bins with equal sizes)
- Train MLP to tell which segment *a single data point* comes from
  - Number of classes is T, labels given by index of segment
  - Multinomial logistic regression



Time-contrastive learning: Definition

- Observe n-dim time series $\mathbf{x}(t)$
- Divide $\mathbf{x}(t)$ into T segments (e.g. bins with equal sizes)
- Train MLP to tell which segment *a single data point* comes from
  - Number of classes is T, labels given by index of segment
  - Multinomial logistic regression
- In hidden layer $\mathbf{h}$, MLP should learn to represent nonstationarity (= differences between segments)



Theorem: TCL estimates nonlinear nonstationary ICA

- Assume data follows nonlinear ICA model $\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t))$ with
  - smooth, invertible nonlinear mixing $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$
  - independent sources $s_i(t)$ with nonstationary variances

Theorem: TCL estimates nonlinear nonstationary ICA

- Assume data follows nonlinear ICA model $\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t))$ with
  - smooth, invertible nonlinear mixing $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$
  - independent sources $s_i(t)$ with nonstationary variances
- Assume we apply time-contrastive learning on $\mathbf{x}(t)$
  - i.e. logistic regression to discriminate between time segments
  - using MLP with hidden layer in $\mathbf{h}(\mathbf{x}(t))$ with $\dim(\mathbf{h}) = \dim(\mathbf{x})$

Theorem: TCL estimates nonlinear nonstationary ICA

- Assume data follows nonlinear ICA model $\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t))$ with
  - smooth, invertible nonlinear mixing $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$
  - independent sources $s_i(t)$ with nonstationary variances
- Assume we apply time-contrastive learning on $\mathbf{x}(t)$
  - i.e. logistic regression to discriminate between time segments
  - using MLP with hidden layer in $\mathbf{h}(\mathbf{x}(t))$ with $\dim(\mathbf{h}) = \dim(\mathbf{x})$
- Then, $\mathbf{s}(t)^2 = \mathbf{A}\mathbf{h}(\mathbf{x}(t))$ for some linear mixing matrix $\mathbf{A}$. (Squaring is element-wise)

Theorem: TCL estimates nonlinear nonstationary ICA

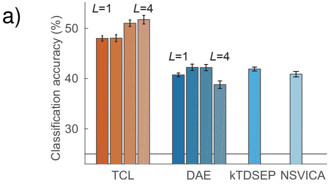
- Assume data follows nonlinear ICA model $\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t))$ with
  - smooth, invertible nonlinear mixing $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^n$
  - independent sources $s_i(t)$ with nonstationary variances
- Assume we apply time-contrastive learning on $\mathbf{x}(t)$
  - i.e. logistic regression to discriminate between time segments
  - using MLP with hidden layer in $\mathbf{h}(\mathbf{x}(t))$ with $\dim(\mathbf{h}) = \dim(\mathbf{x})$
- Then, $\mathbf{s}(t)^2 = \mathbf{A}\mathbf{h}(\mathbf{x}(t))$ for some linear mixing matrix $\mathbf{A}$. (Squaring is element-wise)
- I.e.: TCL demixes nonlinear ICA model up to linear mixing (which can be estimated by linear ICA) and up to squaring.
- This is a constructive proof of identifiability

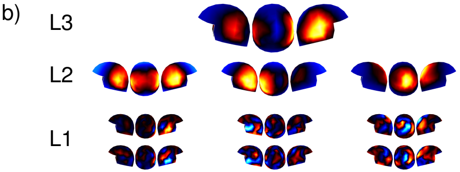
Theorem: TCL estimates nonlinear nonstationary ICA

- Assume data follows nonlinear ICA model $\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t))$ with
  - smooth, invertible nonlinear mixing $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$
  - independent sources $s_i(t)$ with nonstationary variances
- Assume we apply time-contrastive learning on $\mathbf{x}(t)$
  - i.e. logistic regression to discriminate between time segments
  - using MLP with hidden layer in $\mathbf{h}(\mathbf{x}(t))$ with $\dim(\mathbf{h}) = \dim(\mathbf{x})$
- Then, $\mathbf{s}(t)^2 = \mathbf{A}\mathbf{h}(\mathbf{x}(t))$ for some linear mixing matrix $\mathbf{A}$. (Squaring is element-wise)
- I.e.: TCL demixes nonlinear ICA model up to linear mixing (which can be estimated by linear ICA) and up to squaring.
- This is a constructive proof of identifiability
- Generalizations: exponential families, dimension reduction

Experiments with brain imaging data

- MEG data (like EEG but better)
- Sources estimated from resting data (no stimulation)
- a) Validation by classifying another data set with four stimulation modalities: visual, auditory, tactile, rest.
  - Trained a linear SVM on estimated sources
  - Number of layers in MLP ranging from 1 to 4
- b) Attempt to visualize nonlinear processing





Method II: Permutation-contrastive learning (AISTATS2017)

Outline

- We learn features that enable discriminating between short time windows of real data vs. time-shuffled (permuted) data

Method II: Permutation-contrastive learning (AISTATS2017)

Outline

- We learn features that enable discriminating between short time windows of real data vs. time-shuffled (permuted) data
- Again, ordinary NN training:
  Last hidden layer gives the features

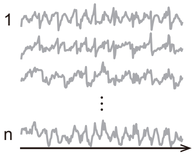
Method II: Permutation-contrastive learning (AISTATS2017)

Outline

- We learn features that enable discriminating between short time windows of real data vs. time-shuffled (permuted) data
- Again, ordinary NN training:
  Last hidden layer gives the features
- Surprising (again!) theoretical result:
  Estimates a nonlinear ICA model
  - with general nonlinear mixing $\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t))$.
  - stationary components $s_i(t)$ with temporal dependencies

Permutation-contrastive learning: Definition

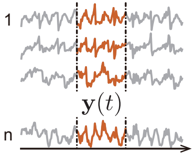
- Observe n-dim time series $\mathbf{x}(t)$



Permutation-contrastive learning: Definition

- Observe n-dim time series $\mathbf{x}(t)$
- Take short time windows as new data

$$\mathbf{y}(t) = \big(\mathbf{x}(t), \mathbf{x}(t-1)\big)$$



Permutation-contrastive learning: Definition

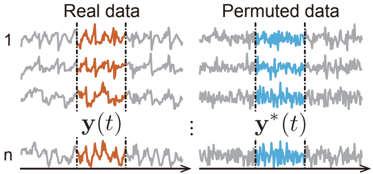
- Observe n-dim time series $\mathbf{x}(t)$
- Take short time windows as new data

$$\mathbf{y}(t) = \big(\mathbf{x}(t), \mathbf{x}(t-1)\big)$$

- Create randomly time-permuted data

$$\mathbf{y}^*(t) = \big(\mathbf{x}(t), \mathbf{x}(t^*)\big)$$

with t^* a random time point.



Permutation-contrastive learning: Definition

- Observe n-dim time series $\mathbf{x}(t)$
- Take short time windows as new data

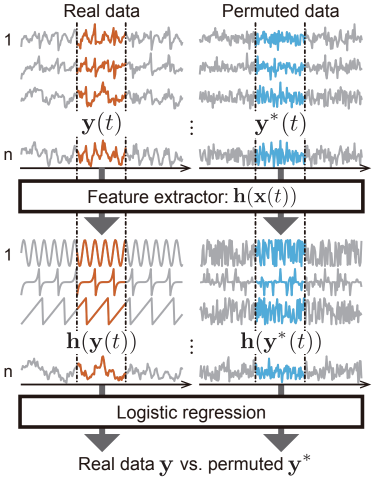
$$\mathbf{y}(t) = (\mathbf{x}(t), \mathbf{x}(t-1))$$

- Create randomly time-permuted data

$$\mathbf{y}^*(t) = (\mathbf{x}(t), \mathbf{x}(t^*))$$

with t^* a random time point.

- Train MLP to discriminate $\mathbf{y}$ from $\mathbf{y}^*$
  - Ordinary nonlinear logistic regression with two classes
  - "Siamese" structure over time



Definitions for convergence theory

- Denote $x = s_i(t)$ and $y = s_i(t-1)$, and

$$q_{x,y}(x,y) := \frac{\partial^2 \log p_{x,y}(x,y)}{\partial x \partial y}$$

- Define (x,y) is quasi-Gaussian if

$$q_{x,y}(x,y) = c\,\alpha(x)\alpha(y)$$

  - Intuitively, dependency is "similar" to Gaussian. Equivalent to

$$\log p(x,y) = \beta_1(x) + \beta_2(y) + c\bar{\alpha}(x)\bar{\alpha}(y)$$

- Define (x,y) is uniformly dependent if $q \neq 0$ for any x, y
  - Basically, a stronger form of dependence. Not necessary (?)

Theorem: PCL estimates nonlinear ICA with time dependencies

- Assume data follows nonlinear ICA model $\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t))$ with
  - smooth, invertible nonlinear mixing $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$
  - Sources $s_i(t)$ are independent (over i) and stationary
  - All $(s_i(t), s_i(t-1))$ non-quasi-Gaussian & uniformly dependent

Theorem: PCL estimates nonlinear ICA with time dependencies

- Assume data follows nonlinear ICA model $\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t))$ with
  - smooth, invertible nonlinear mixing $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$
  - Sources $s_i(t)$ are independent (over i) and stationary
  - All $(s_i(t), s_i(t-1))$ non-quasi-Gaussian & uniformly dependent
- Assume we apply permutation-contrastive learning on $\mathbf{x}(t)$
  - i.e. logistic regression to discriminate between real time windows and time-permuted
  - using MLP with hidden layer in $\mathbf{h}(\mathbf{x}(t))$ with $\dim(\mathbf{h}) = \dim(\mathbf{x})$

Theorem: PCL estimates nonlinear ICA with time dependencies

- Assume data follows nonlinear ICA model $\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t))$ with
  - smooth, invertible nonlinear mixing $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$
  - Sources $s_i(t)$ are independent (over i) and stationary
  - All $(s_i(t), s_i(t-1))$ non-quasi-Gaussian & uniformly dependent
- Assume we apply permutation-contrastive learning on $\mathbf{x}(t)$
  - i.e. logistic regression to discriminate between real time windows and time-permuted
  - using MLP with hidden layer in $\mathbf{h}(\mathbf{x}(t))$ with $\dim(\mathbf{h}) = \dim(\mathbf{x})$
- Then, for all $s_i(t) = k_i(h_j(\mathbf{x}(t)))$ for some ordering of the j, and some scalar nonlinearities $k_i : \mathbb{R} \to \mathbb{R}$.

Theorem: PCL estimates nonlinear ICA with time dependencies

- Assume data follows nonlinear ICA model $\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t))$ with
  - smooth, invertible nonlinear mixing $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^n$
  - Sources $s_i(t)$ are independent (over i) and stationary
  - All $(s_i(t), s_i(t-1))$ non-quasi-Gaussian & uniformly dependent
- Assume we apply permutation-contrastive learning on $\mathbf{x}(t)$
  - i.e. logistic regression to discriminate between real time windows and time-permuted
  - using MLP with hidden layer in $\mathbf{h}(\mathbf{x}(t))$ with $\dim(\mathbf{h}) = \dim(\mathbf{x})$
- Then, for all $s_i(t) = k_i(h_j(\mathbf{x}(t)))$ for some ordering of the j, and some scalar nonlinearities $k_i : \mathbb{R} \rightarrow \mathbb{R}$.
- I.e.: PCL demixes nonlinear ICA
- This is a constructive proof of identifiability of (second) model

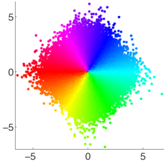
Theorem: PCL estimates nonlinear ICA with time dependencies

- Assume data follows nonlinear ICA model $\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t))$ with
  - smooth, invertible nonlinear mixing $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^n$
  - Sources $s_i(t)$ are independent (over i) and stationary
  - All $(s_i(t), s_i(t-1))$ non-quasi-Gaussian & uniformly dependent
- Assume we apply permutation-contrastive learning on $\mathbf{x}(t)$
  - i.e. logistic regression to discriminate between real time windows and time-permuted
  - using MLP with hidden layer in $\mathbf{h}(\mathbf{x}(t))$ with $\dim(\mathbf{h}) = \dim(\mathbf{x})$
- Then, for all $s_i(t) = k_i(h_j(\mathbf{x}(t)))$ for some ordering of the j, and some scalar nonlinearities $k_i : \mathbb{R} \rightarrow \mathbb{R}$.
- I.e.: PCL demixes nonlinear ICA
- This is a constructive proof of identifiability of (second) model
- For quasi-Gaussian sources, demixes up to linear mixing

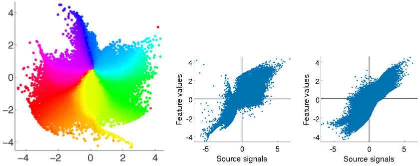
Illustration of demixing capability

- AR Model with Laplacian innovations, $n = 2$
  $\log p(s(t)|s(t-1)) = |s(t) - \rho s(t-1)|$
- Nonlinearity is MLP. Mixing: leaky ReLU's; Demixing: maxout

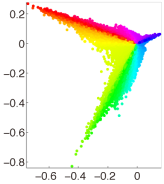
Sources (s)



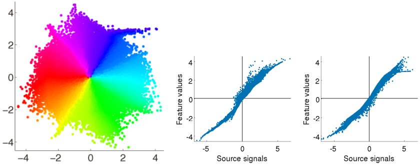
Estimates by kTDSEP (Harmeling et al 2003)



Mixtures (x)

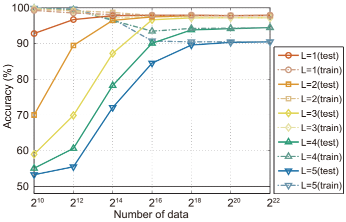


Estimates by our PCL

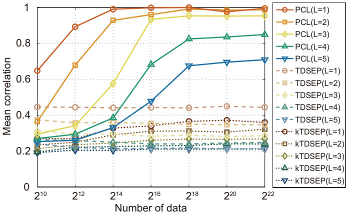


Simulations

- AR Model with Laplacian innovations, $n = 20$
- Nonlinearity is MLP. Mixing: leaky ReLU's; Demixing: maxout



Classification accuracies.
L: number of layers.
Solid lines: test data.
Dash-dotted line: training data.
Chance level is 50%.



Rank correlation coefficients between sources and estimates.
Solid lines: PCL.
Dashed line: TDSEP.
Dotted line: kTDSEP.

Conclusion

- Two new methods for unsupervised learning
  - In time-contrastive learning, divide time series into segments, learn to discriminate data points in them

Conclusion

- Two new methods for unsupervised learning
  - In time-contrastive learning, divide time series into segments, learn to discriminate data points in them
  - In permutation-contrastive learning, discriminate between time windows of real data vs. of permuted (shuffled) data.

Conclusion

- Two new methods for unsupervised learning
  - In time-contrastive learning, divide time series into segments, learn to discriminate data points in them
  - In permutation-contrastive learning, discriminate between time windows of real data vs. of permuted (shuffled) data.
- Training uses ordinary deep learning algorithms and software

Conclusion

- Two new methods for unsupervised learning
  - In time-contrastive learning, divide time series into segments, learn to discriminate data points in them
  - In permutation-contrastive learning, discriminate between time windows of real data vs. of permuted (shuffled) data.
- Training uses ordinary deep learning algorithms and software
- We proved that TCL and PCL solve nonlinear ICA
  - with general (smooth) nonlinear mixing function
  - nonstationary (TCL) or time-dependent (PCL) sources

Conclusion

- Two new methods for unsupervised learning
  - In time-contrastive learning, divide time series into segments, learn to discriminate data points in them
  - In permutation-contrastive learning, discriminate between time windows of real data vs. of permuted (shuffled) data.
- Training uses ordinary deep learning algorithms and software
- We proved that TCL and PCL solve nonlinear ICA
  - with general (smooth) nonlinear mixing function
  - nonstationary (TCL) or time-dependent (PCL) sources
- First cases of identifiable nonlinear ICA

Conclusion

- Two new methods for unsupervised learning
  - In time-contrastive learning, divide time series into segments, learn to discriminate data points in them
  - In permutation-contrastive learning, discriminate between time windows of real data vs. of permuted (shuffled) data.
- Training uses ordinary deep learning algorithms and software
- We proved that TCL and PCL solve nonlinear ICA
  - with general (smooth) nonlinear mixing function
  - nonstationary (TCL) or time-dependent (PCL) sources
- First cases of identifiable nonlinear ICA
- A new principled framework for unsupervised deep learning

Conclusion

- Two new methods for unsupervised learning
  - In time-contrastive learning, divide time series into segments, learn to discriminate data points in them
  - In permutation-contrastive learning, discriminate between time windows of real data vs. of permuted (shuffled) data.
- Training uses ordinary deep learning algorithms and software
- We proved that TCL and PCL solve nonlinear ICA
  - with general (smooth) nonlinear mixing function
  - nonstationary (TCL) or time-dependent (PCL) sources
- First cases of identifiable nonlinear ICA
- A new principled framework for unsupervised deep learning
- Future work:
  - Application on image/video data
  - Combine the two methods