Fourier PCA and Robust Tensor Decomposition

Navin Goyal, Santosh Vempala and Ying Xiao

Presented by Chicheng Zhang

Nov. 2015

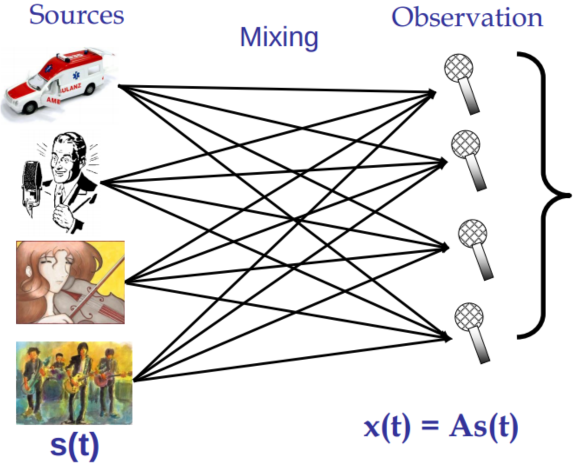
Outline

Introduction

- Problem: Linear Independent Component Analysis (ICA)
- $x = As$, $A \in \mathbb{R}^{n \times m}$ is a "mixing matrix" of full column rank, $s \in \mathbb{R}^m$ has independent entries
- Given iid samples $x_1, \ldots, x_N$
- Goal: (approximately) recover A.

Motivation: Blind Source Separation

- m people talk at a cocktail party
- n speakers receive voices with mixing weights A
- Find A in order to "de-mix" the signals

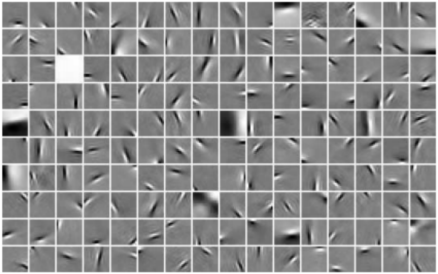


Motivation: Feature Extraction [Hoyer and Hyvärinen]

- Linear image synthesis model

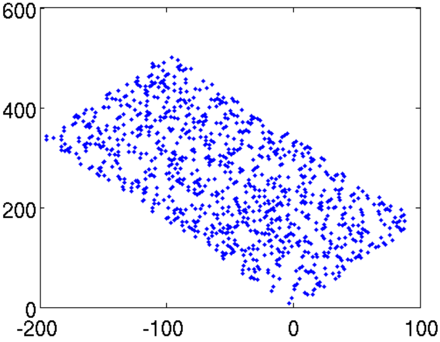
$$x = s_1 \cdot A_1 + s_2 \cdot A_2 + \ldots + s_n \cdot A_n$$

- ICA as feature extracton tool



Motivation: Learning a Parallelepiped [Frieze, Jerrum, Kannan]

- Given: random samples uniformly from a parallelepiped
- Goal: identify its edges (columns of A)
- $s_i \sim U([a_i, b_i])$ independent



Comparison with Principal Component Analysis(PCA)

- PCA: Find linear transformation W, such that Wx is a set of *uncorrelated* random variables that minimize the reconstruction error $\min_U \mathbb{E}\|x - UWx\|^2$
- ICA: Find linear transformation W, such that Wx is a set of *independent* random variables.
- As we will see PCA will be a preprocessing step of ICA

Outline

Preprocessing: Centering

Lemma

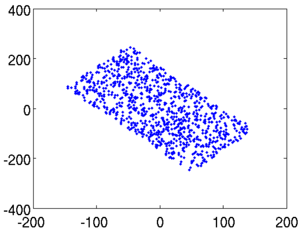
We can assume that $\mathbb{E}s = 0$.

Proof.

Since $x - \mathbb{E}x = A(s - \mathbb{E}s)$, let $\tilde{x} := x - \mathbb{E}x$, $\tilde{s} := s - \mathbb{E}s$, we have that $\tilde{s}$ still has independent entries and

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

□



Preprocessing: Whitening

Lemma

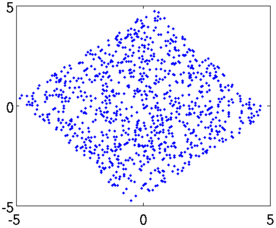
We can further assume that A is an $m \times m$ orthogonal matrix, and each entry of s is of unit variance.

Proof.

Consider $\Sigma = \mathbb{E}xx^T$ that has reduced SVD $\Sigma = UDU^T$, and $\Lambda = \mathbb{E}ss^T$. Then let $\tilde{x} := D^{-1/2}U^T x$ and $\tilde{s} := \Lambda^{-1/2}s$, we have that

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

where $\tilde{A} = D^{-1/2}U^T A\Lambda^{1/2}$ is a $m \times m$ orthogonal matrix. □



Identifiability Problem

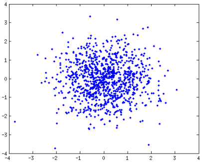
- Example:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix}$$

- Observation: If $s_1, s_2 \sim N(0,1)$, then the plausible A's may not be unique!
- An altenative explanation would be:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$$

where $z_1, z_2 \sim N(0,1)$



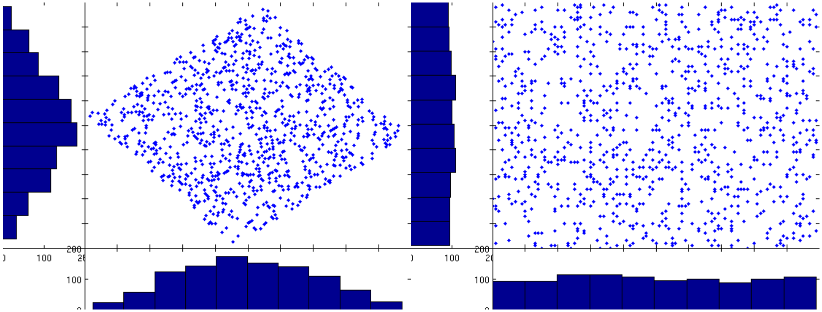
- Claim: so long as there are two Gaussian independent components, cannot hope to recover the columns of A

Outline

Previous Work [Hyvärinen, Oja; Frieze, Jerrum, Kannan]

- CLT implies that sums of independent random variables will be Gaussian like
- Intuition: find transformation W such that each coordinate of Wx is as far from Gaussian as possible
- e.g. Find w maximizing(minimizing) kurtosis of $w^T x$:

$$\max_{w:\|w\|=1} \mathbb{E}(w^T x)^4 - 3$$



Previous Work: Method of Moments [Cardoso]

- Suppose the skewness of s_i, i.e. $\text{skew}(s_i) = \mathbb{E}s_i^3$ are all nonzero
- Then
$$\hat{\mathbb{E}}(x^{\otimes 3}) \to \mathbb{E}(x^{\otimes 3}) = \sum_i \text{skew}(s_i) A_i^{\otimes 3}$$
- Decompose tensor $\hat{\mathbb{E}}(x^{\otimes 3})$ to recover A

Previous Work: Method of Moments [Cardoso]

- Suppose the kurtosis of s_i, i.e. $\text{kurt}(s_i) = \mathbb{E}s_i^4 - 3$ are all nonzero
- Then some statistic of x converges to

$$\sum_i \text{kurt}(s_i) A_i^{\otimes 4}$$

- Decompose the tensor to recover A
- Sanity check: $\text{skew}(s) = 0$, $\text{kurt}(s) = 0$ if s is Gaussian

Outline

Algorithm Description

Fourier PCA

- Input: samples $x_1, \ldots, x_N$.
- Output: columns of mixing matrix $\hat{A}_1, \ldots, \hat{A}_m$
- **1. Fourier Weights:**

  Draw $u \sim N(0, \sigma^2 I_m)$, let $w_i = \frac{e^{ju^T x_i}}{\frac{1}{N}\sum_i e^{ju^T x_i}}$, where $j = \sqrt{-1}$ is the imaginary unit, for $i = 1, 2, \ldots, N$.

Algorithm Description (Cont'd)

- **2. Fourier Covariance:**
  Let $\hat{M}_{ju} = \frac{1}{N}\sum_i w_i(x_i - \hat{m}_{ju})(x_i - \hat{m}_{ju})^T$, where $\hat{m}_{ju} = \frac{1}{N}\sum_i w_i x_i$.
- **3. Eigendecomposition:**
  Let $E_1, \ldots, E_m$ be the unit eigenvectors of $\hat{M}_{ju}$.
- **4. Postprocessing:**
  For each E_i, find $\theta_i \in [0, 2\pi)$ such that $\|\mathrm{Re}(E_i e^{j\theta_i})\|$ is maximized. Let $\hat{A}_i = \mathrm{Re}(E_i e^{j\theta_i})$.

Key Observation: Cumulant Generating Function

Definition

The cumulant generating function (c.g.f.) of m-dimensional random variable X is $\psi_X : \mathbb{C}^m \to \mathbb{C}$

$$\psi_X(t) = \ln \mathbb{E} e^{t^T X}$$

Observation: in ICA problem, the c.g.f. of x is decomposable.

$$\begin{aligned}
\psi_x(t) &= \ln \mathbb{E} e^{t^T x} \\
&= \ln \mathbb{E} e^{t^T A s} \\
&= \ln(\mathbb{E} e^{t^T A_1 s_1} \cdot \cdot \mathbb{E} e^{t^T A_m s_m}) \\
&= \sum_{i=1}^{m} \ln \mathbb{E} e^{t^T A_i s_i} \\
&= \sum_{i=1}^{m} \psi_{s_i}(A_i^T t)
\end{aligned}$$

Key Observation: Cumulant Generating Function

Consider the Hessian of $\psi_x(t)$:

$$
\begin{aligned}
D^2\psi_x(t) &= \sum_{i=1}^{m} D^2\phi_{s_i}(A_i^T t) \\
&= \sum_{i=1}^{m} \psi_{s_i}''(A_i^T t) A_i A_i^T \\
&= A\,\mathrm{diag}(\psi_{s_1}''(A_1^T t), \ldots, \phi_{s_m}''(A_m^T t)) A^T \\
&= A\,\mathrm{diag}(\psi_{s_1}''(A_1^T t), \ldots, \phi_{s_m}''(A_m^T t)) A^{-1}
\end{aligned}
$$

Observation: $D^2\psi_X(t)$'s eigenvectors are precisely columns of A

The Hessian

Lemma

The Hessian $D^2\psi_X(t)$ *can be written as*

$$M_t = \mathbb{E} w_t(X)(X - m_t)(X - m_t)^T$$

where $w_t(x) = \frac{e^{t^T x}}{\mathbb{E} e^{t^T X}}$ *is the "exponential" weight,* $m_t = \mathbb{E} w_t(X) X$ *is the "exponential" weighted mean.*

Proof.

By standard calculus. □

Note that M_t can be estimated by $\hat{M}_t$ using random samples.

Why Complex Numbers?

Key idea: Concentration of $\hat{M}_t$ towards M_t

- i.e. $\hat{\mathbb{E}} w_t(x)(x - \hat{m}_t)(x - \hat{m}_t)^T \to \mathbb{E} w_t(x)(x - m_t)(x - m_t)^T$
- We would like the concentration applicable to a broad family of distributions
- For heavy tailed x, $\mathbb{E} e^{t^T x}$ may even be undefined for any real t
- Solution: take $t = ju$, where $u \in \mathbb{R}^m$ and j is the imaginary unit

Additional Remarks

- Random choice of u: affect M_{ju}'s eigenvalue spacings
- If all s_i's are non-Gaussian, then with proabability 1, $(\psi''_{s_1}(jA_1^T u), \ldots, \psi''_{s_m}(jA_m^T u))$, the eigenvalues of M_{ju}, are distinct
- This is crucial to ensure the eigenvector recovery
- More general results: tensor decomposition of $D^d \psi_x(t)|_{t=ju}$ for $d > 2$.

Outline

Consistency Theorem

Theorem (Informal)

*Suppose we have N iid samples drawn from model $x = As$, where $A \in \mathbb{R}^{m\times m}$ is an orthogonal matrix and s_i's are independent. Moreover, each s_i is far from Gaussian. Then with high probability (over the random draw of u and the samples), Algorithm **Fourier PCA** recovers the columns of A such that*

$$\|\hat{A}_i - A_i\| = o(1)$$

for all $i = 1, 2, \ldots, m$, as $N \to \infty$.

Discussion

- Provides a systematic way of utilizing non-Gaussianity in ICA problem
- Cumulant generating function viewpoint unifies method of moments approaches
- New computationally efficient algorithm using only second-order moments
- Open problem: independent subspace analysis: subsets of $\{s_i\}$ are independent, recovering the respective subspaces of A.

Thank you! Questions?