Compressive Independent Component Analysis

Michael P. Sheehan
IDCOM, University of Edinburgh

Joint work with Mike E. Davies and Madeleine Kotzagiannidis
Outline

1. Compressive Learning
2. Compressive ICA
   i. Independent Component Analysis
   ii. Motivation
   iii. Theory
   iv. Inverse Problem
3. Results
   i. Phase Transition
   ii. Toy Example
4. Outlook
1. Compressive Learning
Learning

- Classification
- Regression
- Parameter estimation
Challenges

1. Dataset has to be stored in memory

2. Computation complexity may scale with the dataset dimensions

3. Amenable to online/distributed learning?

Large Scale Learning

What if $d$ or $n$ is large?
Compression Schemes

- e.g. Feature Selection, Random Projection
- Does not compress the number of data points $n$
Compression Schemes

a) Original
- \( n \) instances of \( d \) dimensions

b) Dimensionality Reduction
- Reduced to \( d' \) dimensions

- e.g. sampling, Nystrom, coresets
- Does not compress the feature space
- Could potentially discard important items
Linear Sketches

- The sketch has dimension $m \ll n d$
- The size $m$ typically scales independent of $n$ and $d$
- Amenable to online learning

**a) Original**

- $n$ elements
- $d$ dimensions

**b) Dimensionality Reduction**

- $n'$ elements
- $d'$ dimensions
- $m$ typically scales independent of $n$ and $d$

**c) Subsampling**

- $n'$ elements
- $d$ dimensions
Compressive learning: How do we form the sketch?

- We pass each data point $x_i$ through a feature function $\Phi: \mathbb{R}^d \rightarrow \mathbb{C}^m$.
Compressive learning: How do we form the sketch?

- We pass each data point $x_i$ through a feature function $\Phi: \mathbb{R}^d \rightarrow \mathbb{C}^m$
Compressive learning: How do we form the sketch?

- We pass each data point $x_i$ through a feature function $\Phi: \mathbb{R}^d \rightarrow \mathbb{C}^m$.
Compressive learning: How do we form the sketch?

- We pass each data point $x_i$ through a feature function $\Phi: \mathbb{R}^d \rightarrow \mathbb{C}^m$.
Compressive learning: How do we form the sketch?

\[ [\Phi(x_1), \Phi(x_2), \Phi(x_3), ..., \Phi(x_n)] \]

- We then pool and average the feature function of each data point to form the sketch

\[ z = \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i) \]
Compressive learning: How do we form the sketch?

Advantages

• Only the sketch of size $m$ has to be stored in memory

• Typically, $m \ll nd$

• Easily amenable to online and distributed learning
How do we learn from a sketch?

- We want to learn the parameters $\theta$ of the learning model $\pi_\theta \in \mathcal{P}$ where $x \sim \pi_\theta$
- Similar to moment matching, we match the sketch with it’s expectation

$$
\min_{\theta \in \Theta} \left\| z - \mathbb{E}_{x \sim \pi_\theta} \Phi(x) \right\|
$$

Sketch (empirical moment)  True moment
Compressive Learning: How is it possible?

• Sketch reformulation: $\mathcal{A}(\pi_\theta) := \mathbb{E}_{x \sim \pi_\theta} \Phi(x)$

• $\mathcal{A}: \mathcal{P} \rightarrow \mathbb{R}^m$ equivalently a linear operator acting the model

$$\min_{\pi_\theta \in \mathbb{S}} \|z - \mathcal{A}(\pi_\theta)\|$$

Reformulated

Constrained model set
Compressive Learning: How is it possible?

$$\min_{\pi_\theta \in \mathcal{S}} \| z - A(\pi_\theta) \|$$

- Low Rank
- Low dimensional Manifold
- Sparsity
What has compressive learning achieved so far?

1. Compressive k means

- **Model set** $\pi_\theta \colon k$ centres $c_1, c_2, \ldots, c_k$

- **Feature Function** $\Phi(x) = \left( \frac{\exp(i \omega_j^T x)}{w \omega_j} \right)_{j=1}^m$

- **Sketch Size** $m \approx O(kd)$
What CL has achieved so far?

1. Compressive k means

2. Compressive GMM

- **Model set** $\pi_\theta$: mixture of $k$ Gaussians

- **Feature Function** $\Phi(x) = \left(e^{i\omega^T x}\right)_j^m$

- **Sketch Size** $m \approx O(kd)$
What CL has achieved so far?

1. Compressive k means

2. Compressive GMM

3. Compressive PCA

- **Model set** \( \pi_\theta : k \) dimensional subspace - \( \text{rank}(\Sigma_{\pi_\theta}) \leq k \)

- **Feature Function** \( \Phi(x) = \langle a_j, xx^T \rangle_{j=1}^m \)

- **Sketch Size** \( m \approx O(kd) \)

“Compressive Statistical Learning with Random Feature Moments” Gribonval et. al. 2020
2. Compressive ICA
Independent Component Analysis

- ICA is a method to identify latent variables that are mutually independent to one another.
- Applications: Blind source separation, EEG recordings, financial modelling, telecommunications
- Given a dataset $X = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^{n \times d}$
- $S = (s_1, s_2, \ldots, s_n) \in \mathbb{R}^{n \times d}$
- Mixing matrix $Q \in \mathbb{R}^{d \times d}$

**Goal**

Estimate $Q$ from $X$
Independent Component Analysis

ICA assumptions
• Each source signal \( s = (s_1, s_2, \ldots, s_d) \) in time has components that are mutually independent:

\[
\pi(s) = \prod_{i=1}^{d} \pi_i(s_i)
\]

• The individual distributions are assumed non-Gaussian but left unspecified (Semi-parametric)
How do we estimate $Q$?

Maximize the independence $\hat{s}$

$\hat{s} = \hat{Q}^T x$

Measures of independence
- Mutual information
- KL divergence
- Kurtosis (cumulant based methods)

Optimise $\hat{Q}$
Cumulant Based ICA (Kurtosis)

- The 4\textsuperscript{th} order cumulant (kurtosis) of a $\mathbf{x}$ is defined as

$$
(\mathcal{X}_{\pi_\theta})_{ijkl} = \mathbb{E}_{\pi_\theta}[x_i x_j x_k x_l] - \mathbb{E}_{\pi_\theta}[x_i x_j] \mathbb{E}_{\pi_\theta}[x_k x_l] - \mathbb{E}_{\pi_\theta}[x_i x_k] \mathbb{E}_{\pi_\theta}[x_j x_l] - \mathbb{E}_{\pi_\theta}[x_i x_l] \mathbb{E}_{\pi_\theta}[x_j x_k]
$$

- In our multivariate setting, the 4\textsuperscript{th} order cumulant gives rise to a 4\textsuperscript{th} order tensor $\mathcal{X}_{\pi_\theta} \in \mathbb{R}^{d \times d \times d \times d}$

- Note
  1. The diagonal entries $(\mathcal{X}_{\pi_\theta})_{iiii}$ are auto-cumulants
  2. The off-diagonal entries $(\mathcal{X}_{\pi_\theta})_{ijkl}$ are the cross-cumulants

$$
\hat{s} = \hat{Q}^T \mathbf{x}
$$
Cumulant Based ICA (Kurtosis)

The 4\textsuperscript{th} order cumulant (kurtosis) of a $\mathbf{x}$ is defined as

$$ (\mathcal{X}_{\pi\theta})_{ijkl} = \mathbb{E}_{\pi\theta}[x_i x_j x_k x_l] - \mathbb{E}_{\pi\theta}[x_i x_j] \mathbb{E}_{\pi\theta}[x_k x_l] - \mathbb{E}_{\pi\theta}[x_i x_k] \mathbb{E}_{\pi\theta}[x_j x_l] - \mathbb{E}_{\pi\theta}[x_i x_l] \mathbb{E}_{\pi\theta}[x_j x_k] $$

In our multivariate setting, the 4\textsuperscript{th} order cumulant gives rise to a 4\textsuperscript{th} order tensor $\mathbf{X}_{\pi\theta} \in \mathbb{R}^{d \times d \times d \times d}$

- Note
  1. The diagonal entries $(\mathcal{X}_{\pi\theta})_{iiii}$ are auto-cumulants
  2. The off-diagonal entries $(\mathcal{X}_{\pi\theta})_{ijkl}$ are the cross-cumulants

\[ \hat{\mathbf{s}} = \hat{Q}^T \mathbf{x} \]
Cumulant Based ICA (Kurtosis)

**GOAL**

Optimize $\hat{Q}$ such that

$$\hat{S} = X_{\pi\theta} \times_1 \hat{Q}^T \times_2 \hat{Q}^T \times_3 \hat{Q}^T \times_4 \hat{Q}^T$$

is **diagonal**

$$\hat{s} = \hat{Q}^T x$$
Compressive ICA

• The geometry of the cumulant tenors gives rise to a natural model set:

$$\mathcal{S} = \{\pi_{\theta} \mid X_{\pi_{\theta}} = S \times_1 Q \times_2 Q \times_3 Q \times_4 Q\}$$
Compressive ICA

• The geometry of the cumulant tenors gives rise to a natural model set:

\[ \mathcal{S} = \{ \pi_{\theta} \mid X_{\pi_{\theta}} = S \times_1 Q \times_2 Q \times_3 Q \times_4 Q \} \]

Diagonal Tensor with \( d \) degrees of freedom
Compressive ICA

• The geometry of the cumulant tenors gives rise to a natural model set:

$$\mathcal{G} = \{ \pi_\theta \mid X_{\pi_\theta} = S \times_1 Q \times_2 Q \times_3 Q \times_4 Q \}$$

Diagonal Tensor with $d$ degrees of freedom
Orthogonal matrix with $\frac{d(d-1)}{2}$ degrees of freedom
Compressive ICA

- The geometry of the cumulant tenors gives rise to a natural model set:

\[ \mathcal{S} = \{ \pi_{\theta} \mid X_{\pi_{\theta}} = S \times_1 Q \times_2 Q \times_3 Q \times_4 Q \} \]

- Diagonal Tensor with \( d \) degrees of freedom totaling \( \frac{d(d+1)}{2} \) degrees of freedom
- Orthogonal matrix with \( \frac{d(d-1)}{2} \) degrees of freedom
Compressive ICA

• The geometry of the cumulant tenors gives rises to a natural model set:

$$\mathcal{G} = \{ \pi_\theta \mid X_{\pi_\theta} = S \times_1 Q \times_2 Q \times_3 Q \times_4 Q \}$$

• $\mathcal{G}$ is a low dimension model set residing in $\mathbb{R}^{d \times d \times d \times d}$

Total of $d(d + 1)$ degrees of freedom
Compressive ICA

• The geometry of the cumulant tenors gives rises to a natural model set:

$$\mathcal{S} = \{\pi_\theta \mid X_{\pi_\theta} = S \times_1 Q \times_2 Q \times_3 Q \times_4 Q\}$$

Total of $d(d+1)$ degrees of freedom

• $\mathcal{S}$ is a low dimension model set residing in $\mathbb{R}^{d \times d \times d \times d}$

• Can we exploit the structure/sparsity of $X_{\pi_\theta}$ to from a sketch?
Compressive ICA: Forming the Sketch

- Feature function: $\Phi(x) = \langle a_j, x \otimes x \otimes x \otimes x \rangle_{j=1}^m$

- Equivalent Sketching Operator: $\mathcal{A}(X_{\pi\theta}) = A \text{vec}(X_{\pi\theta})$
  where $A \in \mathbb{R}^{m \times d^4}$ is a random (sub) Gaussian matrix
Compressive ICA: Forming the Sketch

- Feature function: $\Phi(x) = \langle a_j, x \otimes x \otimes x \otimes x \rangle_{j=1}^m$

- Equivalent Sketching Operator: $A(X_{\pi\theta}) = A\text{vec}(X_{\pi\theta})$

  where $A \in \mathbb{R}^{m \times d^4}$ is a random (sub) Gaussian matrix

Recall: $z = \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i)$
Compressible ICA: Forming the Sketch

- Feature function: $\Phi(x) = \langle a_j, x \otimes x \otimes x \otimes x \rangle_{j=1}^m$

- Equivalent Sketching Operator: $A(X_{\pi_\theta}) = A \text{vec}(X_{\pi_\theta})$
  where $A \in \mathbb{R}^{m \times d^4}$ is a random (sub) Gaussian matrix

\[ \text{Inverse Problem} \]
\[ \min_{X_{\pi_\theta} \in \mathcal{S}} \| z - A(X_{\pi_\theta}) \| \]

Recall: $z = \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i)$
Compressive ICA Restricted Isometry Property

Let $A_{ij} \sim N(0, \sqrt{m^{-1}})$, then for any $\xi, \delta \in (0,1)$ and $X_{\pi\theta} \in \mathcal{S}$, we have

$$(1 - \delta) \left\| X_{\pi\theta_1} - X_{\pi\theta_2} \right\|_F^2 \leq \left\| A \left( X_{\pi\theta_1} - X_{\pi\theta_2} \right) \right\|_2^2$$
Compressive ICA Restricted Isometry Property

Let $A_{ij} \sim N(0, \sqrt{m^{-1}})$, then for any $\xi, \delta \in (0,1)$ and $x_{\pi \theta} \in \mathcal{S}$, we have

$$(1 - \delta) \left\| x_{\pi \theta_1} - x_{\pi \theta_2} \right\|_F^2 \leq \left\| A \left( x_{\pi \theta_1} - x_{\pi \theta_2} \right) \right\|_2^2$$

with prob. $1 - \xi$ provided that the sketch size

$$m \geq \frac{C}{\delta^2} \max \left\{ 4d(d + 1) \log(6), \log\left(\frac{6}{\xi}\right) \right\}$$
Compressive ICA Restricted Isometry Property

\[ m \approx O(d(d + 1)) \]

\[ m \geq \frac{C}{\delta^2} \max \left\{ 4d(d + 1) \log(6) , \log\left(\frac{6}{\xi}\right) \right\} \]
Constrained Optimization

- We propose an iterative projection gradient descent scheme to solve the OP

\[
\min_{\mathbf{X}_{\pi\theta} \in \mathcal{G}} \| \mathbf{z} - \mathcal{A}(\mathbf{X}_{\pi\theta}) \|
\]

Constrained Optimization

\[
\min_{\mathcal{X}_{\pi\theta} \in \mathcal{S}} \| z - \mathcal{A}(\mathcal{X}_{\pi\theta}) \|
\]

- We propose an iterative projection gradient descent scheme to solve the OP.

- After each gradient step, we project onto the model set \( \mathcal{S} \).
Constrained Optimization

\[
\min_{X_{\pi\theta} \in \mathcal{G}} \| z - \mathcal{A}(X_{\pi\theta}) \|
\]

- We propose an iterative projection gradient descent scheme to solve the OP
- After each gradient step, we project onto the model set \( \mathcal{G} \)
- Akin to Compressive Sensing (hard thresholding onto the sparse set)
Constrained Optimization

\[
\min_{\mathcal{X} \pi \theta \in \mathcal{G}} \| \mathbf{z} - \mathcal{A}(\mathbf{X}_{\pi \theta}) \|
\]

- We propose an iterative projection gradient descent scheme to solve the OP
- After each gradient step, we project onto the model set \( \mathcal{G} \)
- Akin to Compressive Sensing (hard thresholding onto the sparse set)
Constrained Optimization

\[
\min_{\frak{x}_{\pi\theta} \in \mathcal{S}} \| \mathbf{z} - \mathcal{A}(\frak{x}_{\pi\theta}) \|
\]

- We propose an iterative projection gradient descent scheme to solve the OP
- After each gradient step, we project onto the model set \( \mathcal{S} \)
- Akin to Compressive Sensing (hard thresholding onto the sparse set)
Constrained Optimization

\[
\min_{\mathbf{X}_{\pi \theta} \in \mathbb{S}} \| \mathbf{z} - \mathcal{A}(\mathbf{X}_{\pi \theta}) \|
\]

- We propose an iterative projection gradient descent scheme to solve the OP
- After each gradient step, we project onto the model set \( \mathbb{S} \)
- Akin to Compressive Sensing (hard thresholding onto the sparse set)
Constrained Optimization

\[ \min_{x_{\pi \theta} \in \mathcal{G}} \| z - A(x_{\pi \theta}) \| \]

- We propose an iterative projection gradient descent scheme to solve the OP
- After each gradient step, we project onto the model set \( \mathcal{G} \)
- Akin to Compressive Sensing (hard thresholding onto the sparse set)
3. Results
Compressive ICA RIP in Practice

Experiment 1

• We mix the cumulant tensor of \( d = 7 \) sources signals by a known mixing matrix \( Q \) to construct \( X_{\pi_\theta} \).

• For varying sketch sizes \( m \), we compute the sketch and obtain the estimate \( \hat{Q} \) using the IPG algorithm.

• Consider the estimation successful if \( D(Q, \hat{Q}) \leq 10^{-7} \), where \( D \) is the Amari error/distance.

Compressive ICA RIP: \( m \approx O(d(d + 1)) \)
Compressive ICA RIP in Practice

Compressive ICA RIP:
\[ m \approx O(d(d + 1)) \]
Compressive ICA RIP in Practice

Compressive ICA RIP: \( m \approx \mathcal{O}(d(d + 1)) \)

Phase transition
Compressive ICA RIP in Practice

Compressive ICA RIP:
\[ m \approx O(d(d + 1)) \]
Compressive ICA RIP in Practice

Phase transition occurs around $m = 2d(d + 1)$

Compressive ICA RIP: $m \approx O(d(d + 1))$
Compressive ICA RIP in Practice

Compressive ICA RIP: $m \approx \mathcal{O}(d(d + 1))$

In practice the order is $\approx 2.5$

Phase transition occurs around $m = 2d(d + 1)$
How efficient are sketches?

Experiment 2

- $d = 7$ source signals mixed by $Q$
- Measure the average Amari error $D(Q, \hat{Q})$ over 1000 trials
- Plotted as function of the number of data points $n$
How efficient are sketches?

Sketch size is not sufficient so fails.
How efficient are sketches?

The mean Amari error of the sketch converges toward the full data error.
Toy Example
Toy Example
Toy Example

TOP SECRET!!!
True Source Signal

Compressive ICA Estimate

Cumulant Based ICA (Comon) Estimate

Fast ICA Estimate
<table>
<thead>
<tr>
<th>Method</th>
<th>Amari Error</th>
<th>Space Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fast ICA (recourse to data)</td>
<td>0.4087</td>
<td>$5 \times 16000$ Data matrix</td>
</tr>
<tr>
<td>Cumulant based ICA (no Compression)</td>
<td>0.4129</td>
<td>$5 \times 5 \times 5 \times 5$ cumulant Tensor (70 DoF)</td>
</tr>
<tr>
<td>Compressive ICA</td>
<td>0.4156</td>
<td>$m = 38$ size sketch</td>
</tr>
</tbody>
</table>
Limitations

• In general it is difficult to find closed form projections onto model sets

• Here we use a proxy projection, where we first partially diagonalise the cumulant tensor using existing techniques and then threshold the cross cumulants to zero.

• As a result, the computational complexity is equivalent to other cumulant based method
Summary

• We have shown that a low dimensional model set exists in the space of cumulant tensors for the ICA problem

• As a result, we can form sketches that are of the order of the model set to estimate the parameters of the ICA model

• The memory complexity is reduced from $O(d^4)$ to $O(d(d + 1))$
Outlook

- Seek a cheaper projection operator or proxy that exhibits a computational complexity that scales with $m$

- Quantify theoretically the controlled loss of information/efficiency of taking a sketch of size $m$

- Can we leverage other sufficient statistics to produce sketches from when the distribution, like ICA, is left unspecified?
Thank you for your attention!

Any questions?