# Sketching for Large-Scale Learning of Mixture Model

N. Keriven[⋆§]   A. Bourrier[†]   R. Gribonval[§]   P. Pérez[‡]

⋆ Université Rennes 1, France
§ INRIA Rennes-Bretagne Atlantique, France
† Gipsa-Lab, St-Martin-d'Hères, France
‡ Technicolor, Cesson Sévigné, France

ICASSP 2016

# Outline

# Paths to Compressive Learning

**Objective**

Learn parameters $\Theta$ from a **large** database $(\mathbf{x}_1, ..., \mathbf{x}_N) \in \mathbb{R}^n$.

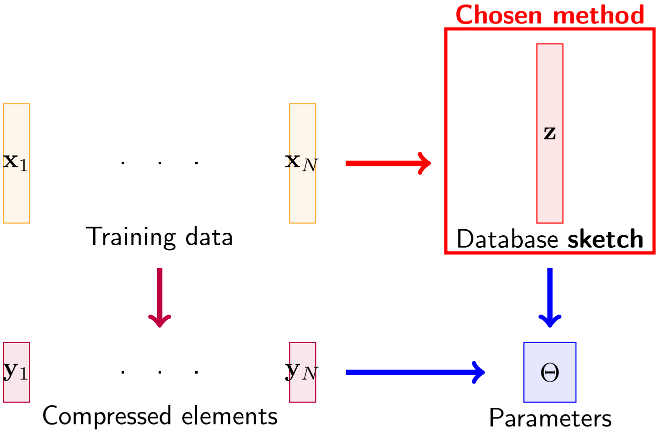
Examples:

- Learn subspace $V_\Theta$ of principal components
- Learn parameters of a classifier $f_\Theta$
- Fit a probability distribution $p_\Theta$
- ...

# Paths to Compressive Learning

## Objective

Learn parameters $\Theta$ from a **large** database $(\mathbf{x}_1, ..., \mathbf{x}_N) \in \mathbb{R}^n$.



# In this paper:

**In this paper**

Efficient method for Gaussian Mixture Model (GMM) estimation from a sketch.

Ex : Estimation of a $20$-GMM from a database of $10^6$ vectors in $\mathbb{R}^{10}$

- $5000$-fold compression of the database
  - Can be performed efficiently on GPU / clusters
- Estimation process $70\times$ faster than EM
- Same precision than EM in the result

# Approach : Generalized Compressive Sensing

## Traditional Compressive Sensing (CS)

From $\mathbf{y} \approx \mathbf{M}\mathbf{x} \in \mathbb{R}^m$ recover vector $\mathbf{x} \in \mathbb{R}^n$

- Linear $\mathbf{M} \in \mathbb{R}^{m \times n}$ with $m < n$
- Typical assumption: $\mathbf{x}$ sparse, etc.

## Generalized Compressive Sensing

From $\mathbf{z} \approx \mathcal{A}p \in \mathbb{C}^m$ recover probability distribution $p \in L^1(\mathbb{R}^n)$

Must define:

- Linear operator $\mathcal{A} : L^1(\mathbb{R}^n) \mapsto \mathbb{C}^m$
- Generalized "sparsity" in $L^1(\mathbb{R}^n)$

# Sparse probability distributions: Mixture Models

- Set of parametric probability distributions: $\mathcal{G} = \{p_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \mathcal{T}\}$
- "$K$-sparse" probability distributions :

$$p_{\Theta,\boldsymbol{\alpha}} = \sum_{k=1}^{K} \alpha_k p_{\boldsymbol{\theta}_k}$$

- Sketch $\mathbf{z} = \sum_{k=1}^{K} \alpha_k \mathcal{A} p_{\boldsymbol{\theta}_k}$ as a combination of atoms in the dictionary:

$$\mathcal{D} = \{\mathcal{A} p_{\boldsymbol{\theta}};\ \boldsymbol{\theta} \in \mathcal{T}\}$$

# Application to Compressive Learning

## Structure of the sketching operator $\mathcal{A}$

Collection of generalized moments $M_j : \mathbb{R}^n \mapsto \mathbb{C}$:

$$\mathcal{A}p = [\mathbb{E}_{\mathbf{x}\sim p} M_j(\mathbf{x})]_{j=1...m}$$

## Compressive Learning procedure

Given a database $(\mathbf{x}_1, ..., \mathbf{x}_N) \overset{i.i.d.}{\sim} p$:

- Compute empirical sketch $\hat{\mathbf{z}} = [\hat{\mathbb{E}} M_j(\mathbf{x})]_{j=1...m} \approx \mathcal{A}p$
- Recover $p_{\Theta,\boldsymbol{\alpha}}$ from $\hat{\mathbf{z}}$ using (generalized) CS techniques

**Questions:**

- Reconstruction algorithm ? (Section 2)
- Choice of sketching operator $\mathcal{A}$ ? (Section 3)

# Outline

# Approach

## Cost function

$$\min_{\Theta,\boldsymbol{\alpha}} \|\hat{\mathbf{z}} - \mathcal{A}p_{\Theta,\boldsymbol{\alpha}}\|_2$$

- Similar to $\min_{\mathbf{x}:\|\mathbf{x}\|_0 \le s} \|\mathbf{y} - \mathbf{M}\mathbf{x}\|_2$ in CS.

- **Pros:** Under some hypothesis on $\mathcal{G}$ and $\mathcal{A}$, yields provably good solutions with high probability (upcoming paper)
- **Cons:** Generally highly non-convex / intractable
  - Convex relaxation[1]: seems difficult because of infinite / continuous dictionary
  - Greedy approaches: approach retained here

[1] Florentina Bunea et al. **SPADES and mixture models**. *The Annals of Statistics* (2010)

# Orthogonal Matching Pursuit with Replacement

- OMP: add an atom to the support by maximizing its correlation to the residual, update the residual, repeat.
- **OMP with Replacement**[2]
  - Perform potentially more iterations than OMP, add a Hard Thresholding step.

*Similar to CoSAMP or Subspace Pursuit.*

- **Compressive Learning OMPR** *(proposed)*
  - Enforce non-negativity on weights $\boldsymbol{\alpha}$
  - Deal with continuous dictionary using gradient descents
  - Add a global optimization step at each iteration.

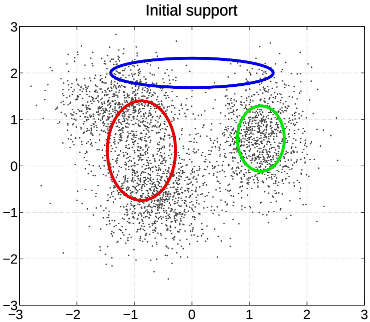
| Number of iterations | Compressive Sensing | Compressive Learning |
|---|---|---|
| $K$ | OMP | CLOMP |
| $2K$ | OMPR | CLOMPR |

[2] Prateek Jain, Ambuj Tewari, and Inderjit S. Dhillon. **Orthogonal matching pursuit with replacement**. *NIPS* (2011)

# Compressive Learning OMPR

Example : iteration $4$ of CLOMPR, searching for a $3$-GMM

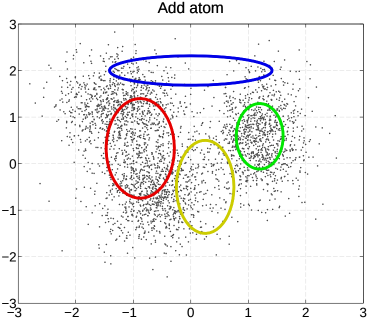
- Current support



# Compressive Learning OMPR

Example : iteration $4$ of CLOMPR, searching for a $3$-GMM

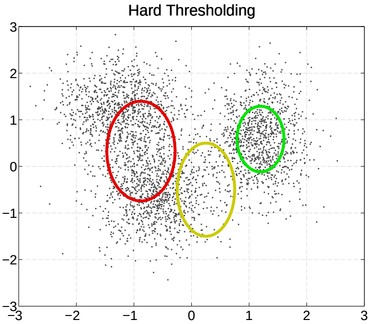
- Add an atom to the support with a gradient descent: $\arg\max_{\boldsymbol{\theta}} Re \left\langle \mathbf{r}, \frac{\mathcal{A}p_{\boldsymbol{\theta}}}{\|\mathcal{A}p_{\boldsymbol{\theta}}\|_2} \right\rangle$



# Compressive Learning OMPR

Example : iteration $4$ of CLOMPR, searching for a $3$-GMM

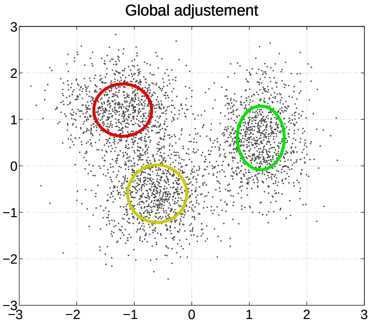
- **Hard Thresholding** to reduce the support
- Solve a **Non-negative** Least Squares to find the weights $\alpha$.



# Compressive Learning OMPR

Example : iteration $4$ of CLOMPR, searching for a $3$-GMM

- **New step:** global gradient descent initialized with the current parameters to further reduce $\|\hat{\mathbf{z}} - \mathcal{A}p_{\Theta,\boldsymbol{\alpha}}\|_2$
- Update residual.



# Outline

# Sketching operator, Gaussian Mixture Model (GMM)

*Recover* $p_{\Theta,\boldsymbol{\alpha}} = \sum_{k=1}^{K} \alpha_k p_{\boldsymbol{\theta}_k}$ *from* $\hat{\mathbf{z}} \approx \mathcal{A}p$.

## Gaussian Mixture Model

$$p_{\boldsymbol{\theta}} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ with diagonal } \boldsymbol{\Sigma}$$

## Random Sampling of the characteristic function[3]

Denote $\psi_p(\boldsymbol{\omega}) = \mathbb{E}_{\mathbf{x}\sim p}(e^{\mathrm{i}\boldsymbol{\omega}^T\mathbf{x}})$. Given $(\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_m) \in \mathbb{R}^n$, define

$$\mathcal{A}p = [\psi_p(\boldsymbol{\omega}_j)]_{j=1,...,m}$$

- Analog to Random Fourier Sampling: $(\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_m) \overset{i.i.d.}{\sim} \Lambda$

[3] Anthony Bourrier, Rémi Gribonval, and Patrick Pérez. **Compressive gaussian mixture estimation**. *ICASSP* (2013)

# Designing the frequency distribution

***The frequency distribution must "scale" with the variances of the GMM.***

Approach 1 Optimize the variance of a Gaussian frequency distribution

- Requires training data with known distribution
- Classical choice[4]

[4] Dougal J Sutherland et al. **Linear-time Learning on Distributions with Approximate Kernel Embeddings**. *arXiv:1509.07553* (2015)

# Designing the frequency distribution

***The frequency distribution must "scale" with the variances of the GMM.***

Approach 1 Optimize the variance of a Gaussian frequency distribution

Approach 2 Proposed:

- Partial preprocessing to compute the appropriate "scaling"
- Distribution that aims at maximizing $\|\nabla_{\boldsymbol{\theta}} \psi_{p_{\boldsymbol{\theta}}}\|_2$

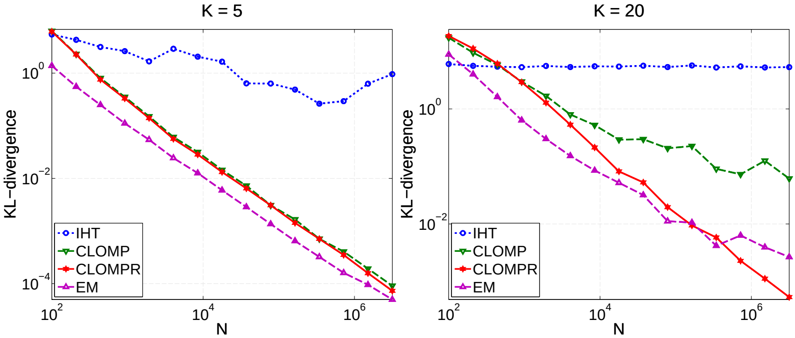
## The proposed distribution

- Yields better precision in the reconstruction
- Is $20\times$ to $100\times$ faster to design

# Outline

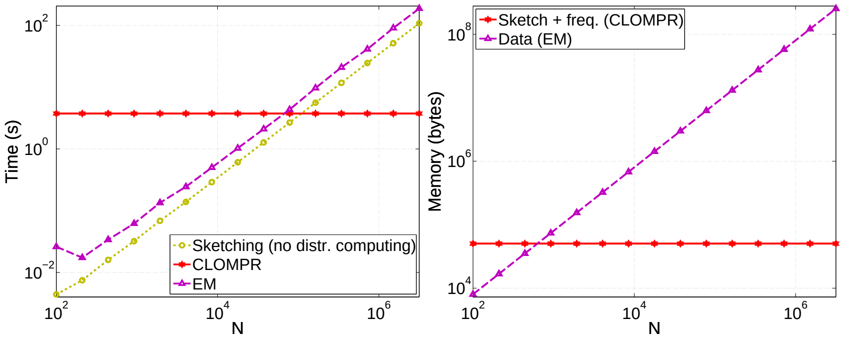
# Reconstruction results

Comparison with EM (VLFeat toolbox) and previous Compressive Learning IHT[4] (originally designed for GMM with fixed covariance). KL-div (lower is better), $n = 10$, $m = 5(2n + 1)K$.




[4]Anthony Bourrier, Rémi Gribonval, and Patrick Pérez. **Compressive gaussian mixture estimation**. *ICASSP* (2013)

# Memory usage and computation time




- Sketching easily done on GPU

# Application : speaker verification

- *NIST2005 database with MFCCs*
- *Classical method*[5]*, far from state-of-the-art but serves as a proof of concept*

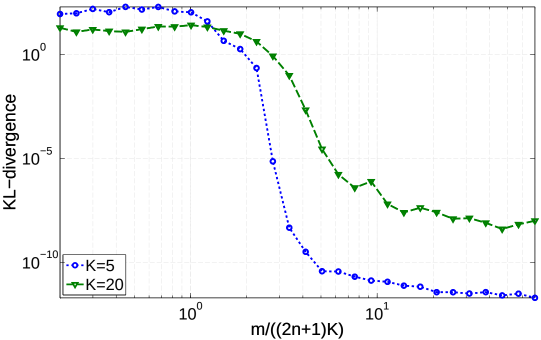
| | CLOMPR | | | EM |
|---|---|---|---|---|
| | $m = 10^3$ | $m = 10^4$ | $m = 10^5$ | |
| $N = 3.10^5$ | 37.15 | 30.24 | 29.77 | 29.53 |
| $\mathbf{N = 2.10^8}$ | 36.57 | **28.96** | **28.59** | **N/A** |

- A large database enhances the quality of the sketch

[5] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. **Speaker Verification Using Adapted Gaussian Mixture Models**. *Digital Signal Processing* 10.1-3 (Jan. 2000)

# Phase transition

*Synthetic data*




*Though we have preliminary theoretical guarantees (upcoming paper), not fully explained yet.*

# Outline

# Conclusion

## Summary

Effective method to learn GMMs from a sketch, using greedy algorithms and an efficient heuristic to design the sketching operator.

## Upcoming paper

- Faster algorithm for GMM with large $K$
- Preliminary theoretical guarantees

## Future Work

- Application to other Mixture Models ($\alpha$-stable distributions...)
- Generalized theoretical guarantees
- Application to other kernel methods[6] (classification...)

[6] Dougal J Sutherland et al. **Linear-time Learning on Distributions with Approximate Kernel Embeddings**. *arXiv:1509.07553* (2015)

Questions ?

Nicolas Keriven et al. **Sketching for Large-Scale Learning of Mixture Models**. *hal-01208027v3, ICASSP* (2016)