Sketched Clustering via Approximate Message Passing

Phil Schniter

The Ohio State University

Joint work with **Evan Byrne** (Ohio State),
Antoine Chatalic, and **Rémi Gribonval** (INRIA)

Supported in part by NSF Grant 1716388 and MIT Lincoln Laboratory

SPARS Workshop (Toulouse, France)

July 3, 2019

Outline

1. Sketched Clustering
2. Cluster Recovery via EM and AMP
3. Numerical Experiments
   - Synthetic Data
   - Spectral MNIST
   - Spike Super-Resolution Recovery from Fourier Samples

Outline

1. Sketched Clustering
2. Cluster Recovery via EM and AMP
3. Numerical Experiments
   - Synthetic Data
   - Spectral MNIST
   - Spike Super-Resolution Recovery from Fourier Samples

Clustering with K-Means

- **Given**: T feature vectors $\{\boldsymbol{x}_t\}$ with $\boldsymbol{x}_t \in \mathbb{R}^N$
- **Goal**: Find K centroids $\{\boldsymbol{c}_k\}$ that minimize sum of squared errors:

$$\mathsf{SSE}(\boldsymbol{X}, \boldsymbol{C}) = \sum_{t=1}^{T} \min_k \|\boldsymbol{x}_t - \boldsymbol{c}_k\|_2^2$$

- Finding the SSE-minimizing centroids is NP-hard
- K-means++ is the standard heuristic approach:
  - Lloyd's algorithm plus a careful random initialization
  - Per-iteration complexity of $O(NKT)$
  - Challenge: Complexity and memory can be prohibitive for large T

Sketched Learning

- Sketched learning is an alternative framework:
  1. Compress data $\boldsymbol{X} \in \mathbb{R}^{N \times T}$ down to $\boldsymbol{y} \in \mathbb{C}^M$ (with $M \ll NT$).
  2. Learn parameters (e.g., centroids) from $\boldsymbol{y}$.

- We choose to build the sketch $\boldsymbol{y} = [y_1, \ldots, y_M]^\mathsf{T}$ using[1,4]

$$y_m = \frac{1}{T}\sum_{t=1}^{T} \exp(\mathrm{j}\boldsymbol{w}_m^\mathsf{T}\boldsymbol{x}_t) \quad \text{with random } \{\boldsymbol{w}_m\}_{m=1}^M$$

  - Well matched to distributed and/or streaming scenarios!
  - Complexity & memory of learning are invariant to T!

- Can interpret y_m as samples of the empirical characteristic function:

$$y_m = \phi(\boldsymbol{w}_m) = \int_{\boldsymbol{R}^N} p(\boldsymbol{x}) \exp(\mathrm{j}\boldsymbol{w}_m^\mathsf{T}\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \quad \text{with} \quad p(\boldsymbol{x}) = \frac{1}{T}\sum_{t=1}^{T}\delta(\boldsymbol{x} - \boldsymbol{x}_t)$$

[1] Keriven, Bourrier, Gribonval, Pérez'17, [4] Keriven, Tremblay, Traonmilin, Gribonval'17

Sketched Clustering

- How do we learn the centroids $\boldsymbol{C}$ from the sketch $\boldsymbol{y}$?

- The CL-OMPR algorithm[3][4] aims to solve
$$\{\widehat{\boldsymbol{C}}, \widehat{\boldsymbol{\alpha}}\} = \arg\min_{\boldsymbol{C},\boldsymbol{\alpha}} \sum_{m=1}^{M} \left| y_m - \sum_{k=1}^{K} \alpha_k \exp(\mathrm{j}\boldsymbol{w}_m^{\mathsf{T}} \boldsymbol{c}_k) \right|^2$$
using a greedy heuristic inspired by OMP.

- In practice, CL-OMPR ...
  - recovers accurate centroids with sketch length $M \approx 10KN$
  - has a per-iteration complexity of $O(MNK^2)$

- Can we do better in terms of sample complexity and computational complexity?

[3] Keriven, Bourrier, Gribonval, Pérez'17, [4] Keriven, Tremblay, Traonmilin, Gribonval'17

Outline

Formulation as a Generalized Linear Model

- Suppose we model the data $\boldsymbol{x}_t$ using a Gaussian mixture model (GMM):
$$\boldsymbol{x}_t \sim \textstyle\sum_{k=1}^K \alpha_k \mathcal{N}(\boldsymbol{c}_k, \boldsymbol{\Phi}_k) \quad \text{with } \textstyle\sum_{k=1}^K \alpha_k = 1,\ \alpha_k \geq 0,\ \boldsymbol{\Phi}_k > 0.$$

- As $T \to \infty$, have
$$\begin{aligned} y_m &= \tfrac{1}{T} \textstyle\sum_{t=1}^T \exp(\mathrm{j}\boldsymbol{w}_m^\mathsf{T} \boldsymbol{x}_t) \to \mathbb{E}\left\{ \exp(\mathrm{j}\boldsymbol{w}_m^\mathsf{T}\boldsymbol{x}_t)\right\} \\ &= \textstyle\sum_{k=1}^K \alpha_k \exp\left(\mathrm{j} g_m \underbrace{\boldsymbol{a}_m^\mathsf{T}\boldsymbol{c}_k}_{\triangleq\, z_{mk}} - g_m^2 \underbrace{\boldsymbol{a}_m^\mathsf{T}\boldsymbol{\Phi}_k\boldsymbol{a}_m}_{\triangleq\, \tau_{mk}}/2\right), \end{aligned}$$
where $g_m \triangleq \|\boldsymbol{w}_m\|$ and $\boldsymbol{a}_m \triangleq \boldsymbol{w}_m/g_m$.

- As $N \to \infty$, with isotropic $\boldsymbol{a}_m$, we have $\tau_{mk} \to \mathrm{tr}(\boldsymbol{\Phi}_k)/N \triangleq \tau_k$.

- Thus for large T and N we have the generalized linear model (GLM)
$$p(y_m|\boldsymbol{z}_m; \boldsymbol{\alpha}, \boldsymbol{\tau}) \approx \delta\Big(y_m - \textstyle\sum_{k=1}^K \alpha_k \exp\left(\mathrm{j} g_m z_{mk} - g_m^2 \tau_k/2\right)\Big)$$
with transformed centroids $\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{C}$ & random $\boldsymbol{A}$ w/ isotropic columns

Sketched Clustering via EM

- Objective: Recover the centroids $\boldsymbol{C}$ from the sketch $\boldsymbol{y}$ under the GLM

$$p(\boldsymbol{y}|\boldsymbol{Z};\boldsymbol{\alpha},\boldsymbol{\tau}) = \prod_{m=1}^{M} p(y_m|\boldsymbol{z}_m;\boldsymbol{\alpha},\boldsymbol{\tau}), \quad \boldsymbol{Z} = \boldsymbol{A}\boldsymbol{C}$$

- Challenge: GMM weights $\boldsymbol{\alpha}$ and variances $\boldsymbol{\tau}$ are unknown!

- Approach: Expectation Maximization (EM): Iterate ...

$$\begin{aligned}(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\tau}})^{\mathsf{new}} &= \underset{(\boldsymbol{\alpha},\boldsymbol{\tau}):\ \boldsymbol{\alpha}^{\mathsf{T}}\mathbf{1}=1,\ \boldsymbol{\alpha}\geq\mathbf{0},\ \boldsymbol{\tau}>\mathbf{0}}{\arg\max} \mathbb{E}\left\{ \ln p(\boldsymbol{y},\boldsymbol{Z};\boldsymbol{\alpha},\boldsymbol{\tau}) \,\middle|\, \boldsymbol{y},\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\tau}}\right\} \\ &= \underset{(\boldsymbol{\alpha},\boldsymbol{\tau}):\ \boldsymbol{\alpha}^{\mathsf{T}}\mathbf{1}=1,\ \boldsymbol{\alpha}\geq\mathbf{0},\ \boldsymbol{\tau}>\mathbf{0}}{\arg\max} \sum_{m=1}^{M}\int_{\mathbb{R}^K} \mathcal{N}(\boldsymbol{z}_m;\widehat{\boldsymbol{z}}_m,\boldsymbol{Q}^{\mathsf{z}}_m)\ln p(y_m|\boldsymbol{z}_m;\boldsymbol{\alpha},\boldsymbol{\tau})\end{aligned}$$

  with conditional mean $\widehat{\boldsymbol{z}}_m = \mathbb{E}\{\boldsymbol{z}_m \,|\, \boldsymbol{y};\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\tau}}\}$ and conditional covariance $\boldsymbol{Q}^{\mathsf{z}}_m$.

- Thus we aim to compute MMSE centroid estimates $\widehat{\boldsymbol{C}} = \mathbb{E}\{\boldsymbol{C}|\boldsymbol{Y};\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\tau}}\}$, since 1) they provide $\widehat{\boldsymbol{Z}} = \boldsymbol{A}\widehat{\boldsymbol{C}}$ for EM and 2) solve our sketched clustering problem.

MMSE Inference for Sketched Clustering

- Objective: Compute MMSE centroid estimate $\widehat{\boldsymbol{C}}$ from $\boldsymbol{y}$ under GLM

$$p(\boldsymbol{y}|\boldsymbol{Z}) = \prod_{m=1}^{M} p_{\mathsf{y}|\mathbf{z}}(y_m|\boldsymbol{z}_m;\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\tau}}), \quad \boldsymbol{Z} = \boldsymbol{A}\boldsymbol{C}.$$

- Note that the posterior centroid density is

$$p(\boldsymbol{C}|\boldsymbol{y}) \propto \prod_{m=1}^{M} p_{\mathsf{y}|\mathbf{z}}(y_m|\boldsymbol{a}_m^{\mathsf{T}}\boldsymbol{C};\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\tau}}) \prod_{n=1}^{N} p_{\mathbf{c}}(\boldsymbol{c}_n)$$

  - We assume the trivial centroid prior $p_{\mathbf{c}}(\boldsymbol{c}_n) \propto 1$, but other priors are possible

- We can approximately compute $\widehat{\boldsymbol{C}}$ using approximate message passing
  - Due to the form of the likelihood, we use the "HyGAMP" algorithm[5]

[5] Rangan, Fletcher, Goyal, Schniter'12

Lineage of HyGAMP

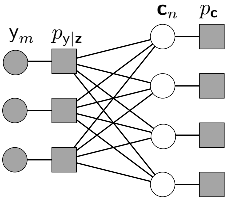
- **Approximate Message Passing** (AMP) [Donoho,Maleki,Montanari'09]
  - Estimate $\boldsymbol{c}$ under the standard linear model $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{c} + \boldsymbol{w}$ with known iid $\boldsymbol{A}$
  - Assumes separable prior $p_{\mathsf{c}}(\boldsymbol{c}) = \prod_n p_{\mathsf{c}}(c_n)$ and AWGN $\boldsymbol{w}$

- **Generalized AMP** (GAMP) [Rangan'11]
  - Estimate $\boldsymbol{c}$ under generalized linear model $\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{z})$ with $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{c}$
  - Assumes separable prior and likelihood $p(\boldsymbol{y}|\boldsymbol{z}) = \prod_m p_{\mathsf{y}|\mathsf{z}}(y_m|z_m)$

- **Hybrid GAMP** (HyGAMP) [Rangan,Fletcher,Goyal,Schniter'12]
  - GAMP with vector-valued variables $\boldsymbol{z}_m, \boldsymbol{c}_n \in \mathbb{R}^K$
  - Separable likelihood: $\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{Z}) = \prod_m p_{\mathsf{y}|\mathbf{z}}(y_m|\boldsymbol{z}_m)$ with $\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{C}$
  - Separable prior: $p(\boldsymbol{C}) = \prod_n p_{\mathbf{c}}(\boldsymbol{c}_n)$

Message-Passing View of HyGAMP

- HyGAMP can be derived by approximating belief propagation (either sum-product or max-product algorithm) on a factor graph with the form:



- Messages are approximated as K-dimensional Gaussian pdfs assuming $N \to \infty$
- HyGAMP tackles the (NK)-dimensional inference problem by iteratively solving $M+N$ inference problems of dimension K

HyGAMP Inference Steps

- HyGAMP's K-dimensional inference steps compute the posterior mean and covariance of the random vectors $\{\mathbf{c}_n\}$ and $\{\mathbf{z}_m\}$ under the posterior pdfs

$$p(\boldsymbol{c}_n|\boldsymbol{r}_n;\boldsymbol{Q}^{\mathsf{r}}) \propto p_{\mathsf{c}}(\boldsymbol{c}_n)\mathcal{N}(\boldsymbol{c}_n;\boldsymbol{r}_n,\boldsymbol{Q}^{\mathsf{r}})$$

$$p(\boldsymbol{z}_m|y_m,\boldsymbol{p}_m;\boldsymbol{Q}^{\mathsf{p}}) \propto p_{\mathsf{y|z}}(y_m|\boldsymbol{z}_m)\mathcal{N}(\boldsymbol{z}_m;\boldsymbol{p}_m,\boldsymbol{Q}^{\mathsf{p}})$$

  - The correctness of these posteriors can be argued, under large i.i.d. Gaussian $\boldsymbol{A}$, using the analysis in [Javanmard,Montanari'13]

- To reduce computational complexity, we use the Simplified HyGAMP (SHyGAMP) algorithm,[6] which approximates covariance matrices as diagonal
  - The per-iteration complexity of SHyGAMP is only $O(MNK)$.

- For the sketched-clustering likelihood $p_{\mathsf{y|z}}(y_m|\boldsymbol{z}_m)$, the computation of $\widehat{\boldsymbol{z}}_m$ and $\operatorname{diag}(\boldsymbol{Q}^{\mathsf{z}}_m)$ uses generalized von Mises functions, and is somewhat involved.[7]

[6] Byrne,Schniter'15, [7] Byrne,Chatalic,Gribonval,Schniter'19

Outline

1. Sketched Clustering
2. Cluster Recovery via EM and AMP
3. Numerical Experiments
   - Synthetic Data
   - Spectral MNIST
   - Spike Super-Resolution Recovery from Fourier Samples

Experiment 1: Synthetic Data

Data generation:

- $\{\boldsymbol{x}_t\}$ drawn i.i.d. from a GMM with
  - centroids $\boldsymbol{c}_k$ drawn $\sim \mathcal{N}(\mathbf{0}, 1.5^2 K^{2/N} \boldsymbol{I}_N)$
  - equal weights $\alpha_k = 1/K$
  - covariances $\boldsymbol{\Phi}_k = \boldsymbol{I}$
- $N = 100$ dimensional, $K = 10$ classes, $T = 10^7$ samples

Sketching:[8]

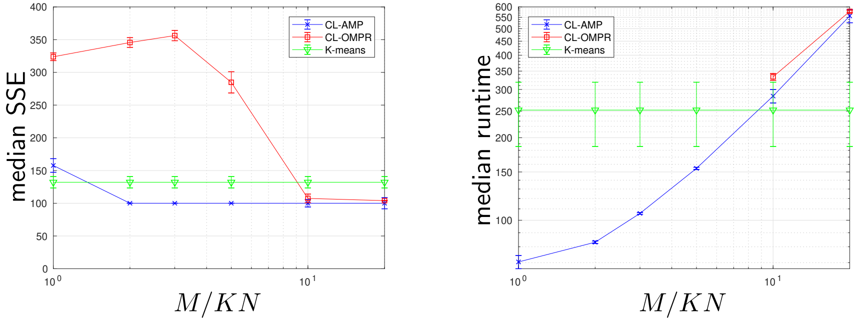
- frequencies $\boldsymbol{w}_m = g_m \boldsymbol{a}_m$ with unit-norm isotropic $\boldsymbol{a}_m$
- $g_m \sim p(g) = 1_{[0,\infty)} \sqrt{g^2\sigma^2 + \frac{g^4\sigma^4}{4}} \exp(-g^2\sigma^2/2)$ with $\sigma^2 = \frac{1}{NT}\|\boldsymbol{X}\|_F^2$

Accuracy metric:

- median of SSE$(\widehat{\boldsymbol{C}}) = \sum_{t=1}^{T} \min_k \|\boldsymbol{x}_t - \widehat{\boldsymbol{c}}_k\|_2^2$ over 10 trials

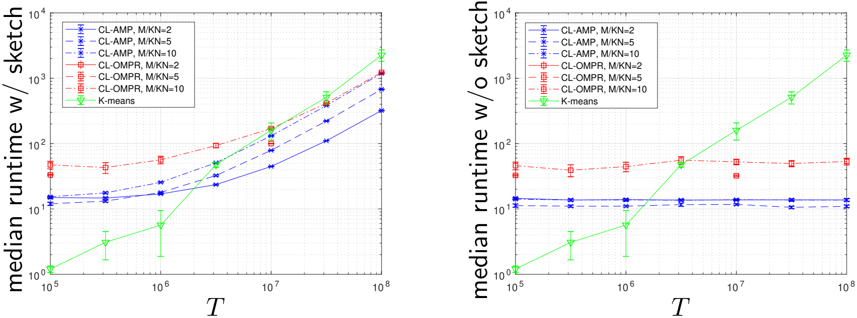
[8] Keriven, Bourier, Gribonval, Perez'17

Accuracy & Runtime vs Sketch Length M



- Sample complexity:
  - CL-AMP needs only $M \approx 2KN$ samples
  - CL-OMPR needs $M \approx 10KN$
- Computational complexity (including sketch):
  - CL-AMP $3\times$ faster than K-means++ for similar accuracy

Runtime vs Data Size T



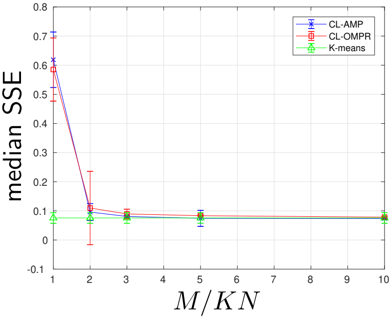
When $T > 2 \times 10^6 \ldots$

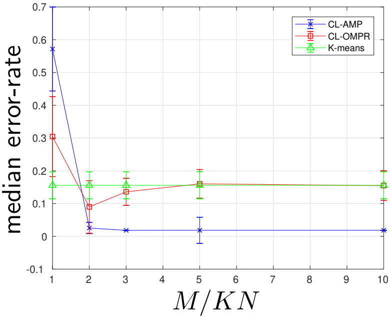
- sketching+CL-AMP is faster than K-means++
- sketching is more expensive than CL-AMP

Experiment 2: Spectral Clustering of MNIST

- We repeat an experiment from [Keriven,Tremblay,Traonmilin,Gribonval'17]
- Original MNIST data:
  - $T = 70,000$ samples of handwritten digits from $K = 10$ classes
- Preprocessing used to extract features of dimension $N = 10$
  - Compute SIFT descriptors
  - Compute k-NN adjacency matrix (for $k=10$) using FLANN
  - Compute $K=10$ principle eigenvectors of normalized Laplacian matrix
- Dataset partitioned into equal-sized training and test sets (10 trials)
- Kmeans++, CL-OMPR, and CL-AMP estimate $K = 10$ centroids from training set
- Accuracy metrics: 1) SSE on training set
  2) error of minimum-distance classifier on test set

Accuracy vs Sketch Length M





For $M \geq 2KN$ ...

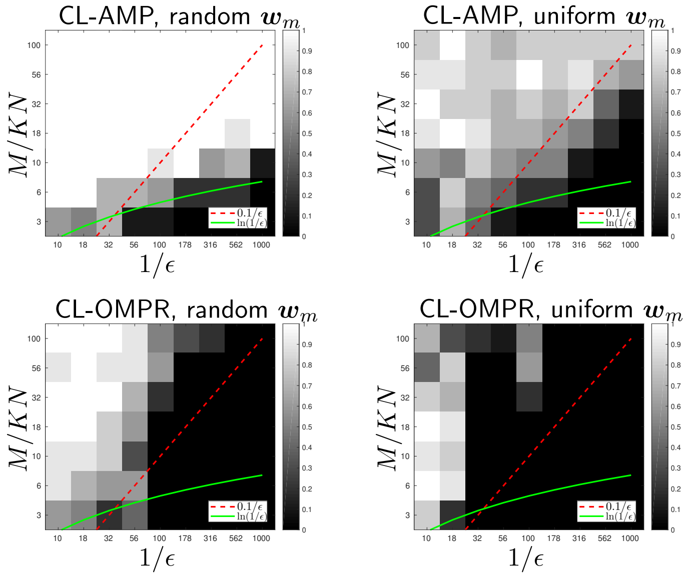
- CL-OMPR and CL-AMP give SSE similar to that of k-means++
- CL-AMP gives error rate much better than CL-OMPR and k-means++

Experiment 3: Spike Super-Resolution w/ Fourier Samples

- Sum-of-spikes signal: $\sum_{k=1}^{K} \alpha_k \delta(\boldsymbol{t} - \boldsymbol{c}_k)$ with time $\boldsymbol{t} \in \mathbb{R}^N$
- Fourier transform: $y(\boldsymbol{w}) = \sum_{k=1}^{K} \alpha_k \exp(\mathrm{j}\boldsymbol{w}^\mathsf{T} \boldsymbol{c}_k)$ with freq $\boldsymbol{w} \in \mathbb{R}^N$
- Goal: Recover $\{\boldsymbol{c}_k\}_{k=1}^{K}$ from Fourier samples $\{y(\boldsymbol{w}_m)\}_{m=1}^{M}$
- Experiment:
  - Generate frequency pairs $\{(\boldsymbol{c}_{2i-1}, \boldsymbol{c}_{2i})\}_{i=1}^{K/2}$ with $\|\boldsymbol{c}_{2i-1} - \boldsymbol{c}_{2i}\| = \epsilon \ \forall i$
  - "Success" if $\max_k \|\widehat{\boldsymbol{c}}_k - \boldsymbol{c}_{i_k}\| < \epsilon/2$ for some $\{i_1, \ldots, i_K\} = \{1, \ldots, K\}$
  - Theoretical analysis[9] says that
    - $M \geq O(\log(1/\epsilon))$ samples suffice for random frequencies $\{\boldsymbol{w}_m\}$
    - $M \geq O(1/\epsilon)$ samples suffice for uniformly spaced frequencies $\{\boldsymbol{w}_m\}$

[9] Traonmilin,Keriven,Gribonval,Blanchard'17

Frequency Estimation Results ($K = 4, N = 2$)



Conclusion

- Sketched clustering is an alternative to traditional clustering that
  1. compresses the dataset down to a sketch (of generalized moments)
  2. extracts centroids from that sketch

  and is well matched to distributed and/or streamed scenarios

- We formulated sketched clustering as a GLM inference problem, and applied EM-SHyGAMP.

- Numerical results suggest that has CL-AMP has good sample & computational complexity

- Ongoing work to analyze the AMP state evolution in the large-system limit $(N, M \to \infty)$

Full paper

E. Byrne, A. Chatalic, R. Gribonval, and P. Schniter, “Sketched Clustering via Hybrid Approximate Message Passing,” *IEEE Trans. Signal Processing*, to appear 2019 (see also `https://arxiv.org/abs/1712.02849`).