Compressive K-means

Nicolas Keriven, Nicolas Tremblay, Yann Traonmilin, Rémi Gribonval

COMPRESSIVE K-MEANS

Nicolas Keriven[†⋆] *Nicolas Tremblay*[‡] *Yann Traonmilin*[⋆] *Rémi Gribonval* [⋆]

[⋆] INRIA Rennes - Bretagne Atlantique, Campus de Beaulieu, FR-35042 Rennes Cedex, France
[†] Université Rennes 1, FR-35065 Rennes Cedex, France
[‡] CNRS, GIPSA-lab, FR-38402 Saint Martin d'Heres Cedex, France

ABSTRACT

The Lloyd-Max algorithm is a classical approach to perform K-means clustering. Unfortunately, its cost becomes prohibitive as the training dataset grows large. We propose a compressive version of K-means (CKM), that estimates cluster centers from a *sketch*, i.e. from a drastically compressed representation of the training dataset.

We demonstrate empirically that CKM performs similarly to Lloyd-Max, for a sketch size proportional to the number of centroids times the ambient dimension, and *independent of the size of the original dataset*. Given the sketch, the computational complexity of CKM is also independent of the size of the dataset. Unlike Lloyd-Max which requires several replicates, we further demonstrate that CKM is *almost insensitive to initialization*. For a large dataset of 10^7 data points, we show that CKM can run two orders of magnitude faster than five replicates of Lloyd-Max, with similar clustering performance on artificial data. Finally, CKM achieves lower classification errors on handwritten digits classification.

Index Terms— Compressive Sensing, K-means, Compressive Learning, Random Fourier Features

1. INTRODUCTION

Given a set of datapoints $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\} \subset \mathbb{R}^n$ (the dataset), the sum of squared errors (SSE) problem is to find K centroids $C = \{\boldsymbol{c}_1, \dots, \boldsymbol{c}_K\} \subset \mathbb{R}^n$ such that the SSE is minimized:

$$\mathrm{SSE}(X, C) = \sum_{i=1}^{N} \min_k \|\boldsymbol{x}_i - \boldsymbol{c}_k\|^2 . \tag{1}$$

Finding the global minimum of this cost function is NP-hard [1], and Lloyd [2] and Steinhaus [3] proposed 50 years ago a classical heuristics that is still commonly used today. Its complexity is $O(nNKI)$ with I the number of iterations until convergence. This becomes prohibitive when any of these factors become too large.

In this paper, we propose a heuristics to find the centroids C from a *sketch* of the dataset X which size m *does not depend* on N. More precisely, a sketching procedure Sk that converts a set of weighted vectors in $\mathbb{R}^n$ to a vector in $\mathbb{C}^m$ is defined, and centroids are derived by finding a set C of weighted points that has a sketch close to that of the dataset X with uniform weights:

$$\operatorname*{argmin}_{C, \boldsymbol{\alpha}} \|\mathsf{Sk}(X, \mathbf{1}/N) - \mathsf{Sk}(C, \boldsymbol{\alpha})\|_2^2 \tag{2}$$

with $\boldsymbol{\alpha} \geq 0$, $\sum_{k=1}^{K} \alpha_k = 1$. As we will see, computing the sketch $\mathsf{Sk}(X, \mathbf{1}/N) \in \mathbb{C}^m$ of the dataset requires only one pass over X. It can also benefit from distributed or online computing.

This work was partly funded by the European Research Council, PLEASE project (ERC-StG-2011-277906).

Centroids can be retrieved from the sketch with a variant of Compressive Learning Orthogonal Matching Pursuit (CLOMPR) [4, 5], an algorithm initially used for large-scale Gaussian Mixture Model (GMM) estimation. Its complexity reads $\mathcal{O}\left(nmK^2\right)$, thus eliminating entirely the dependence in N once the sketch has been computed. This complexity could be further reduced by leveraging fast transforms [6, 7] or embeddings in lower dimension as a preprocessing step [8].

We discuss related works in Section 2, and describe the proposed method in Section 3. Experiments on artificial and real data are performed in Section 4. Even though the cost function (2) bears no immediate connection with the SSE cost (1), we show empirically that the SSE obtained with the proposed method approaches that obtained with repeated runs of Lloyd-Max. Moreover the proposed algorithm is more stable to initialization, and although the sketch size is independent of N, the proposed algorithm performs much better than repeated runs of Lloyd-Max on large datasets, both in terms of clustering quality and computational complexity, with observed runtimes (*given the sketch*) two orders of magnitude faster for a large dataset of 10^7 data points.

2. RELATED WORK

Several lines of work tackle K-means on large datasets. The clever initialization of K-means++ [9] increases stability of the Lloyd-Max algorithm and decreases the number of iterations I until convergence. Some works reduce the ambient dimension n, either by selecting a limited number of features [10, 11], or by embedding all points in a lower dimension using, for instance, random projections [8].

Closer to our work, coresets methods [12, 13] aim at reducing the number of datapoints N by constructing intermediate structures that retain some properties of the SSE. Like the proposed sketches, coresets can also be constructed in a distributed/online fashion [13]. Unlike coreset methods our approach does not explicitly aim at approximating the SSE but uses a different objective function.

Our sketching structure bears connection with Random Fourier Features [14] in the context of Hilbert space embedding of probability distributions [15] (see [5] for further details). Similar embeddings have been used, for instance, in the context of classification of distributions [16]. Here we will see that the proposed approach can be formulated as an infinite-dimensional Compressive Sensing problem, in which a *probability distribution* is measured through a random linear operator, then decoded under the form of a "sparse" distribution, i.e. a finite combination of Diracs whose locations correspond to the desired centroids. This problem can be linked with the super-resolution problem where one aims at estimating combinations of Diracs from a low pass observation. In this case, in dimension one, stable recov-

ery in the low noise regime is possible based on the minimization of the total variation of probability measures [17]. However, the extension of these techniques to higher dimensions $n \gg 1$ does not yield practical results yet [18]. The CLOMPR heuristics empirically overcomes these limitations.

3. PROPOSED METHOD

In many important applications, one wishes to reconstruct a signal from incomplete samples of its discrete Fourier transform [19, 20]. A classical result from Compressive Sensing states that a few randomly selected Fourier samples contain "enough" information to reconstruct a sparse signal with high probability.

3.1. Sketching operator

Given m frequency vectors $\Omega = \{\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_m\}$ in $\mathbb{R}^n$, the sketch of a set of L points Y with weights $\boldsymbol{\beta}$ is formed as follows:

$$\mathsf{Sk}(Y, \boldsymbol{\beta}) = \Big[\sum_{l=1}^{L} \beta_l e^{-i\boldsymbol{\omega}_j^T \boldsymbol{y}_l} \Big]_{j=1}^{m} \in \mathbb{C}^m \tag{3}$$

This sketching procedure can be reformulated as an operator $\mathcal{A}$ which is linear *with respect to probability distributions*. Define this operator as a sampling of the characteristic function $\mathcal{A}p = \left[\mathbb{E}_{\boldsymbol{x}\sim p} e^{-i\boldsymbol{\omega}_j^T \boldsymbol{x}}\right]_{j=1}^{m}$ of a probability distribution p at frequencies $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_m$. Denoting $p_{Y,\boldsymbol{\beta}} = \sum_{l=1}^{L} \beta_l \boldsymbol{\delta}_{\boldsymbol{y}_l}$, the problem (2) can be reformulated as

$$\underset{C,\boldsymbol{\alpha}}{\operatorname{argmin}} \|\hat{\boldsymbol{z}} - \mathcal{A}p_{C,\boldsymbol{\alpha}}\|_2^2 \tag{4}$$

where $\hat{\boldsymbol{z}} = \mathcal{A}\hat{p}_X$ is the sketch of the dataset, with $\hat{p}_X = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\delta}_{\boldsymbol{x}_i}$ the empirical distribution of the data.

In the spirit of Random Fourier Sampling, the frequencies $\boldsymbol{\omega}_j$ are drawn $i.i.d.$ from a distribution Λ. In previous work [4, 5], we proposed a distribution Λ referred to as *Adapted radius* frequency distribution, based on a heuristics that maximizes the variation of the characteristic function at the selected frequencies when data are drawn from a GMM. In this paper we show empirically that this distribution is also adapted to a variety of scenarios even outside the context of GMMs. The Adapted radius distribution is parametrized by a scaling quantity σ^2. In [5] an algorithm to choose this scale parameter is proposed, by computing a small sketch of (a fraction of) the dataset X and performing an adapted regression on it.

The approach thus corresponds to a generalized Compressive Sensing problem, where we measure the probability distribution $\hat{p}_X$, which is the "signal" of interest, through a random linear operator $\mathcal{A}$, and reconstruct it under the form of a "sparse" distribution $p_{C,\boldsymbol{\alpha}}$ only supported on a few Diracs.

3.2. CLOMPR algorithm

The CLOMPR algorithm is a heuristic algorithm to seek a solution to problem (4) that has been proposed in previous work [4] for Gaussian Mixture Model estimation. It is a greedy algorithm inspired by Orthogonal Matching Pursuit (OMP) and its variant OMP with Replacement (OMPR), which comprises more iterations than OMP, with an additional Hard Thresholding step. As we recall below, CLOMPR involves several modifications to OMPR.

- **Non-negativity.** The compressive mixture estimation framework imposes a non-negativity constraint on the weights $\boldsymbol{\alpha}$. Thus **step 1** maximizes the real part of the correlation instead of its modulus. Similarly, in **step 4** a Non-Negative Least-Squares minimization is performed instead of a classical Least-Squares minimization.
- **Continuous dictionary.** The dictionary $\{\mathcal{A}\boldsymbol{\delta}_{\boldsymbol{c}}\}$ is continuously indexed and cannot be exhaustively searched. The maximization in **step 1** is thus done with a gradient ascent $\texttt{maximize}_{\boldsymbol{c}}$, leading to a –local– maximum of the correlation between atom and residual.
- Compared to OMP, CLOMPR involves an additional gradient descent $\texttt{minimize}_{C,\boldsymbol{\alpha}}$ *initialized with the current parameters* (**step 5**), to further reduce the cost function (4) (but also leading to a *local* minimum of (4)).

We also bring some modifications to the original CLOMPR [4, 5]:

- **Initialization strategies.** We test several initialization strategies for **step 1**, each being somehow similar to usual initialization strategies for K-means, see Section 4.2.
- **Additional constraints.** We add constraints to the gradient descents. During the computation of the sketch $\hat{\boldsymbol{z}}$ we also compute bounds $\boldsymbol{l},\ \boldsymbol{u} \in \mathbb{R}^n$ such that all data are comprised in these bounds: denoting $\boldsymbol{a} \leq \boldsymbol{b}$ the element-by-element comparison of vectors in $\mathbb{R}^n$, $\boldsymbol{l}$ and $\boldsymbol{u}$ are such that $\boldsymbol{l} \leq \boldsymbol{x}_i \leq \boldsymbol{u}$ for all i's. Note that the computation of these bounds is also done in one pass over X. Then we enforce $\boldsymbol{l} \leq \boldsymbol{c} \leq \boldsymbol{u}$ in $\texttt{maximize}_{\boldsymbol{c}}$ and $\boldsymbol{l} \leq \boldsymbol{c}_k \leq \boldsymbol{u}$ for all k's in $\texttt{minimize}_{C,\boldsymbol{\alpha}}$.

We denote the resulting algorithm Compressive K-means (CKM).

Algorithm 1: CLOMPR for K-means (CKM)

Data: Sketch $\hat{\boldsymbol{z}}$, frequencies Ω, parameter K , bounds $\boldsymbol{l},\ \boldsymbol{u}$
Result: Centroids C, weights $\boldsymbol{\alpha}$
$\hat{\boldsymbol{r}} \leftarrow \hat{\boldsymbol{z}}$; $C \leftarrow \emptyset$;
for $t \leftarrow 1$ **to** $2K$ **do**
  Step 1: Find a new centroid
    $\boldsymbol{c} \leftarrow \texttt{maximize}_{\boldsymbol{c}} \left(\mathrm{Re} \left\langle \frac{\mathcal{A}\boldsymbol{\delta}_{\boldsymbol{c}}}{\|\mathcal{A}\boldsymbol{\delta}_{\boldsymbol{c}}\|}, \hat{\boldsymbol{r}} \right\rangle,\ \boldsymbol{l},\ \boldsymbol{u} \right)$
  end
  Step 2: Expand support
    $C \leftarrow C \cup \{\boldsymbol{c}\}$
  end
  Step 3: Enforce sparsity by Hard Thresholding if $t > K$
    if $|C| > K$ **then**
      $\boldsymbol{\beta} \leftarrow \arg\min_{\boldsymbol{\beta}\geq 0} \left\| \hat{\boldsymbol{z}} - \sum_{k=1}^{|C|} \beta_k \frac{\mathcal{A}\boldsymbol{\delta}_{\boldsymbol{c}_k}}{\|\mathcal{A}\boldsymbol{\delta}_{\boldsymbol{c}_k}\|} \right\|$
      Select K largest entries $\beta_{i_1}, ..., \beta_{i_K}$
      Reduce the support $C \leftarrow \{\boldsymbol{c}_{i_1}, ..., \boldsymbol{c}_{i_K}\}$
    end
  end
  Step 4: Project to find $\boldsymbol{\alpha}$
    $\boldsymbol{\alpha} \leftarrow \arg\min_{\boldsymbol{\alpha}\geq 0} \left\| \hat{\boldsymbol{z}} - \sum_{k=1}^{|C|} \alpha_k \mathcal{A}\boldsymbol{\delta}_{\boldsymbol{c}_k} \right\|$
  end
  Step 5: Global gradient descent
    $C, \boldsymbol{\alpha} \leftarrow$
    $\texttt{minimize}_{C,\boldsymbol{\alpha}} \left(\left\| \hat{\boldsymbol{z}} - \sum_{k=1}^{|C|} \alpha_k \mathcal{A}\boldsymbol{\delta}_{\boldsymbol{c}_k} \right\|,\ \boldsymbol{l},\ \boldsymbol{u} \right)$
  end
  Update residual: $\hat{\boldsymbol{r}} \leftarrow \hat{\boldsymbol{z}} - \sum_{k=1}^{|C|} \alpha_k \mathcal{A}\boldsymbol{\delta}_{\boldsymbol{c}_k}$
end

3.3. Complexity of the method

The method can be summarized as follows. Given a dataset X, a number of frequencies m and a number of clusters K,

1. Use the algorithm in [5] on a small fraction of X to choose a frequency distribution Λ;
2. Draw m frequencies $\boldsymbol{\omega}_j$ $i.i.d.$ from Λ;
3. compute the sketch $\mathcal{A}\hat{p} = \left[\frac{1}{N}\sum_{i=1}^{N} e^{-i\boldsymbol{\omega}_j^T \boldsymbol{x}_i}\right]_{j=1}^{m}$;
4. Retrieve C from the sketch using the CKM algorithm.

The CKM algorithm scales in $\mathcal{O}\left(K^2 mn\right)$, which is potentially far lower than the $\mathcal{O}\left(nNKI\right)$ of Lloyd-Max for large N.

To compute the sketch, one has to perform the multiplication $\mathsf{W}^T\mathsf{X}$, where $\mathsf{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]$ and $\mathsf{W} = [\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_m]$ are the matrices of data and frequencies. It theoretically scales in $\mathcal{O}\left(nmN\right)$, but can be done in a distributed manner by splitting the dataset over several computing units and averaging the obtained sketches, such that the full data need never be stored in one single location. One can also exploit GPU computing for very large-scale matrix multiplication [21]. The proposed sketch can also be maintained online, which is another crucial property of typical dataset sketches [22].

Some techniques might further reduce these complexities. As detailed in [23], most operations in CKM can be narrowed down to performing multiplications by W and W^T. Therefore, both computing the sketch and performing CKM could benefit from the replacement of W by a suitably randomized fast transform [6, 7].

It is also possible to reduce the dimension n to $\mathcal{O}\left(\log K\right)$ with random projections [8], as a preprocessing step.

Finally, empirical results (see Sec. 4.3) suggest that the size of the sketch only needs to scale linearly with the number of parameters, i.e. $m \approx \mathcal{O}\left(nK\right)$. Combining all these results, it would be potentially possible to compute the sketch in $\mathcal{O}\left(KNT^{-1}(\log K)^2\right)$, where T is the number of parallel computing units, and perform CKM in $\mathcal{O}\left(K^3(\log K)^2\right)$. We also mention that a hierarchical adaptation of CLOMPR which scales in $\mathcal{O}\left(K^2(\log K)^3\right)$ has been proposed for GMM estimation [5], and that a variant for the K-means setting considered here might be implementable. We leave possible integration of those techniques to future work.

4. EXPERIMENTS

4.1. Setup

We compare our Matlab implementation of CKM, available at [23], with Matlab's `kmeans` function that implements Lloyd-Max.

We first use artificial clustered data drawn from a mixture of K unit Gaussians in dimension n with uniform weights, with means $\boldsymbol{\mu_k}$ drawn according to a centered Gaussian with covariance $cK^{1/n}\mathsf{Id}$. The constant $c = 1.5$ is chosen so that clusters are sufficiently separated with high probability. Unless indicated otherwise, $N = 3 \cdot 10^5$ points are generated from $K = 10$ clusters with $n = 10$.

The second problem consists in performing spectral clustering [24] on the MNIST dataset [25]. In fact, to test our method's performance on a large dataset, we use the original $7 \cdot 10^4$ images, that we complete with images artificially created by distortion of the original ones using the toolbox infMNIST proposed in [26]. We thereby test on three dataset sizes: the original one with $N_1 = 7 \cdot 10^4$ and two augmented ones with $N_2 = 3 \cdot 10^5$, $N_3 = 10^6$. For each dataset, we extract SIFT [27] descriptors of each image, and compute the K-nearest neighbours adjacency matrix (with $K = 10$) using FLANN [28]. As we know there are ten classes, we compute the first ten eigenvectors of the associated normalized Laplacian matrix, and run CKM on these N 10-dimensional feature vectors. Note that spectral clustering requires the first few eigenvectors of the global Laplacian matrix, of size N^2, which becomes prohibitive for large N. Replacing the K-means step by CKM in compressive versions of spectral clustering [29, 30] or in efficient kernel methods such as in [31] are left for future investigations.

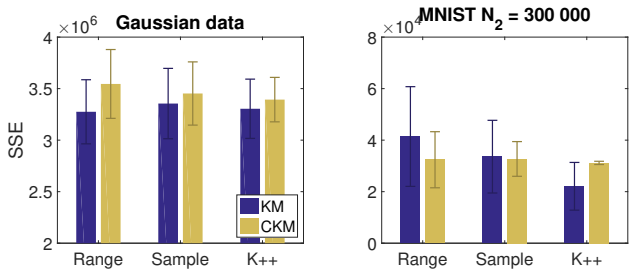

Fig. 1. Comparison of initialization strategies. Mean and variance of SSE cost over 100 experiments.

Unless indicated otherwise $m = 1000$ frequencies are used. Each result is averaged over 100 experiments.

4.2. Initialization strategies

Several strategies to initialize the gradient descent $\texttt{maximize}_{\boldsymbol{c}}$ in **step 1** of CKM are tested, along with their equivalents in the usual `kmeans` algorithm. Note that, for `kmeans` it corresponds to selecting K initial centroids then running the algorithm, while for CKM each iteration is initialized with a single new centroid.

- **Range:** for CKM, pick $\boldsymbol{c}$ where each component c_i is drawn uniformly at random with $l_i \leq c_i \leq u_i$; for `kmeans`, select K such points.
- **Sample:** for CKM, select a point $\boldsymbol{c} = \boldsymbol{x}_i$ from the data at random; for `kmeans`, select K such points.
- **K++,** a strategy analog to the K-means++ algorithm [9]: for CKM, select $\boldsymbol{c} = \boldsymbol{x}_i$ from the data with a probability inversely proportional to its distance to the *current* set of centroids C; for `kmeans`, run exactly the K++ algorithm [9].

NB: the last two strategies do not exactly fit in the "compressive" framework, where data are sketched then discarded, since they still require access to the data. They are implemented for testing purpose.

In Fig. 1 the different initialization strategies are compared, by displaying the mean and variance of the SSE over 100 experiments. On Gaussian data, both algorithms yield approximately the same SSE for all strategies. On MNIST data, for CKM all strategies approximately yield the same result, but the Sample and K++ initializations give a lower variance. The `kmeans` algorithm is more sensitive to initialization, and only outperforms the CKM algorithm for the K++ strategy. In all further experiments the "Range" strategy is used for both algorithms.

4.3. Number of frequencies

It is interesting to empirically evaluate how many frequencies m are required for CKM to be effective. In Fig 2, we show the SSE obtained with CKM divided by that of `kmeans` with respect to the relative number of frequencies $m/(Kn)$, and draw lines where the relative SSE becomes lower than 2. It is seen that these lines are almost constant at $m/(Kn) \approx 5$, except for a deviation at low n.

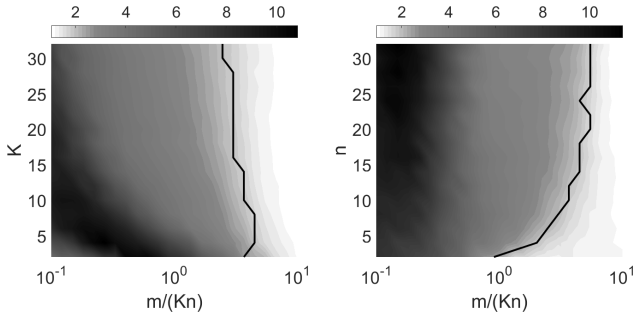

Fig. 2. Relative SSE (*i.e.* SSE obtained with CKM divided by that obtained with `kmeans`) on Gaussian data, with $n = 10$ (left) and $K = 10$ (right). Lines are drawn where the relative SSE becomes lower than 2.

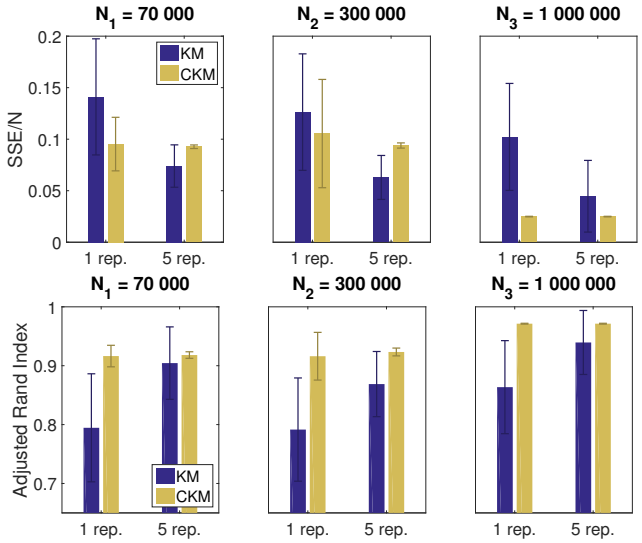

Fig. 3. Mean and variance over 100 experiments of SSE divided by N (top, lower is better) and Adjusted Rand Index [32] for comparing clustering results (bottom, higher is better) on MNIST, for 1 or 5 replicates.

Recent, preliminary theoretical results on GMMs [5] hint that for a fixed error level the required number of frequencies grows proportional to the number of parameters $m \approx \mathcal{O}(Kn)$. We postulate that the same phenomenon might be valid for K-means clustering.

4.4. Scalability and performance of CKM

Often Lloyd-Max is repeated several times with random initializations, and the set of centroids yielding the lowest SSE is kept. In the CKM algorithm, we do not have access to the SSE in practice since the data are discarded after computation of the sketch. Hence, when several replicates of CKM are performed, *we select instead the set of centroids that minimizes the cost function* (4).

In Fig. 3, we evaluate the SSE and classification performance on the MNIST dataset, for 1 or 5 replicates of the algorithms. As expected, `kmeans` greatly benefits from being performed several times, while CKM is more stable between number of replicates. This allows CKM to be run with (much) fewer replicates than `kmeans` in practice. Moreover, for a large dataset ($N_3 = 10^6$), the performance of CKM has negligible variance and negligible difference between 1 and 5 replicates. Hence, *although the size m of the sketch is kept fixed for all N's, the method is actually more efficient when applied to large datasets*.

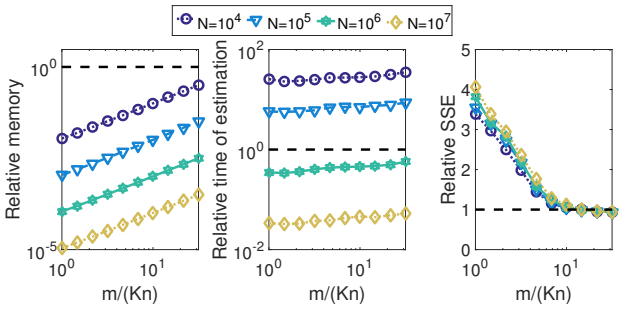

Fig. 4. Relative time, memory and SSE of CKM algorithm with respect to *one run* of `kmeans` (10^0 represent the `kmeans` result), on Gaussian data.

Interestingly, in all cases CKM outperforms `kmeans` in terms of classification (Fig. 3, bottom). This might mean that the proposed cost function is more adapted than the SSE on this particular task.

We finally examine the time and memory complexities of CKM, relatively to that of `kmeans`, in Fig. 4. Note that the computational complexity of computing the sketch is not outlined on this figure, since it can be done in an online and massively parallelized manner and is highly dependent on the user's available hardware. As expected, given the sketch CKM is far more efficient than `kmeans` on large datasets, even for a high number of frequencies. Overall, on a dataset with 10^7 elements, one run of CKM is up to 150 times faster than `kmeans` with 5 replicates.

5. CONCLUSION

We presented a method for performing K-means on large datasets, where the centroids are derived from a *sketch* of the data. The problem was linked to generalized Compressive Sensing, where a probability distribution is measured through a linear operator and reconstructed as a sparse mixture of K Diracs. A modified version of the CLOMPR algorithm [4, 5] is used for estimating this mixture.

Results showed that, although the proposed objective is not directly linked to the traditional SSE cost, our method compares favorably with usual algorithms for K-means. It is much more stable to initialization, and generally succeeds with only one replicate. Although the size of the sketch does not depend on N, compared to usual K-means the proposed algorithm is all the more effective when applied on large datasets, in terms of complexity, SSE and classification performance on the MNIST dataset for instance.

Outlooks As already mentioned, it is possible to combine the proposed approach with dimension reduction [8] and/or fast transforms [7] to speed-up the method even more.

The proposed method was observed to outperform usual K-means for classification on the MNIST dataset, even when it performs worse in terms of SSE. These encouraging results may lead to further engineering of objective functions and innovative clustering methods. As mentioned in the introduction, the cost function (4) can be connected to a finite embedding of probability distributions in Hilbert Spaces [15] with Random Features [14]. In this framework, theoretical results about the information preservation property of sketches have been derived on GMMs [5], and further work will examine such results in the context of mixtures of Diracs.

6. REFERENCES

[1] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "NP-hardness of Euclidean sum-of-squares clustering," *Machine Learning*, vol. 75, no. 2, pp. 245–248, 2009.

[2] S. Lloyd, "Least Squares Quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[3] H. Steinhaus, "Sur la division des corps materiels en parties," *Bull. Acad. Polon. Sci. IV (C1.III)*, vol. IV, no. 12, pp. 801–804, 1956.

[4] N. Keriven, A. Bourrier, R. Gribonval, and P. Pérèz, "Sketching for Large-Scale Learning of Mixture Models," in *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2015.

[5] N. Keriven, A. Bourrier, R. Gribonval, and P. Pérèz, "Sketching for Large-Scale Learning of Mixture Models," *arXiv preprint arXiv:1606.02838*, pp. 1–50, 2016.

[6] T. Do, L. Gan, N. Nguyen, and T. Tran, "Fast and Efficient Compressive Sensing using Structurally Random Matrices," *IEEE Transactions on Signal Processing*, vol. 30, 2011.

[7] L. Le Magoarou and R. Gribonval, "Flexible Multi-layer Sparse Approximations of Matrices and Applications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 688–700, 2016.

[8] C. Boutsidis, A. Zouzias, and P. Drineas, "Random Projections for k-means Clustering," in *Advances in Neural Information and Processing Systems (NIPS)*, 2010, pp. 298–306.

[9] D. Arthur and S. Vassilvitskii, "k-means++: The Advantages of Careful Seeding," in *ACM-SIAM symposium on Discrete algorithms*, 2007, pp. 1027–1035.

[10] J. Altschuler, A. Bhaskara, G. Fu, V. Mirrokni, A. Rostamizadeh, and M. Zadimoghaddam, "Greedy Column Subset Selection: New Bounds and Distributed Algorithms," in *Proceedings of The 33rd International Conference on Machine Learning*, 2016.

[11] C. Boutsidis, P. Drineas, and M. W. Mahoney, "Unsupervised feature selection for the k-means clustering problem," in *Advances in Neural Information and Processing Systems (NIPS)*, 2009, pp. 153–161.

[12] G. Frahling and C. Sohler, "A fast k -means implementation using coresets," *Proceedings of the twenty-second annual symposium on Computational geometry (SoCG)*, vol. 18, no. 6, pp. 605–625, 2005.

[13] D. Feldman and M. Langberg, "A unified framework for approximating and clustering data," *Proceedings of the forty-third annual ACM symposium on Theory of computing*, , no. 46109, pp. 569–578, 2011.

[14] A. Rahimi and B. Recht, "Random Features for Large Scale Kernel Machines," *Advances in Neural Information Processing Systems (NIPS)*, , no. 1, pp. 1–8, 2007.

[15] B. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. Lanckriet, "Hilbert space embeddings and metrics on probability measures," *The Journal of Machine Learning Research*, vol. 11, pp. 1517–1561, 2010.

[16] D. J. Sutherland, J. B. Oliva, P. Barnabas, and J. Schneider, "Linear-time Learning on Distributions with Approximate Kernel Embeddings," *arXiv:1509.07553*, pp. 1–10, 2015.

[17] E. Candes and C. Fernandez-Granda, "Super-Resolution from Noisy Data," , no. October 2012, pp. 1–22, 2012.

[18] Y. De Castro, F. Gamboa, D. Henrion, and J. Lasserre, "Exact solutions to Super Resolution on semi-algebraic domains in higher dimensions," *arXiv preprint arXiv:1502.02436*, pp. 1–22, 2015.

[19] E. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[20] E. Candes, J.K. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 480–509, 2006.

[21] P. Zhang and Y. Gao, "Matrix Multiplication on High-Density Multi-GPU Architectures: Theoretical and Experimental Investigations Peng," in *ISC High Performance*, vol. 1, pp. 17–30. Springer International Publishing, 2015.

[22] G. Cormode, M. Garofalakis, P. J. Haas, and C. Jermaine, "Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches," *Foundations and Trends in Databases*, vol. 4, no. xx, pp. 1–294, 2011.

[23] N. Keriven, N. Tremblay, and R. Gribonval, "SketchMLbox : a Matlab toolbox for large-scale learning of mixture models," `http://sketchml.gforge.inria.fr`, 2016.

[24] S. Uw, A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in Neural Information Processing Systems 14*, pp. 849–856, 2001.

[25] Y. Lecun, C. Cortes, and C. Burges, "The MNIST database of handwritten digits," 1998.

[26] G. Loosli, S. Canu, and L. Bottou, "Training Invariant Support Vector Machines using Selective Sampling," *Large Scale Kernel Machines*, pp. 301–320, 2007.

[27] A. Vedaldi and B. Fulkerson, "VLFeat - An open and portable library of computer vision algorithms," Tech. Rep., 2010.

[28] M. Muja and D. Lowe, "Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration," *International Conference on Computer Vision Theory and Applications (VISAPP '09)*, pp. 1–10, 2009.

[29] N. Tremblay, G. Puy, P. Borgnat, R. Gribonval, and P. Vandergheynst, "Accelerated Spectral Clustering Using Graph Filtering Of Random Signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4094–4098.

[30] N. Tremblay, G. Puy, R. Gribonval, and P. Vandergheynst, "Compressive Spectral Clustering," *Proceedings of The 33rd International Conference on Machine Learning*, pp. 1–15, 2016.

[31] R. Chitta, R. Jin, and A. K. Jain, "Efficient Kernel Clustering using Random Fourier Features," *2012 IEEE 12th International Conference on Data Mining*, pp. 161–170, 2012.

[32] William M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," *Journal of American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.