Learning to Sketch for Compressive Clustering

Antoine Chatalic[1], Rémi Gribonval[2].

[1]Univ. Rennes, Inria, CNRS, IRISA; [2]Univ Lyon, Inria, CNRS, ENS de Lyon, UCB Lyon 1, LIP

Abstract— **Clustering of large-scale collections can be performed efficiently based on low-dimensional sketches obtained by averaging random Fourier features of the items in the training collection. Some prior knowledge about the data distribution is however required to design the sketching mechanism and optimize its performance. We show empirically the importance of estimating the inter-cluster separation, and give a proof of concept of how to learn it.**

1 Introduction

With the availability of ever larger datasets, learning accurate models becomes easier but efficient algorithms are needed to process this information. Standard machine learning approaches, which typically require a few passes over the data, give way to approximate and randomized methods with reduced computational costs and memory usage. In the compressive learning framework, the dataset $X = [\mathbf{x}_1, ..., \mathbf{x}_n]$ is compressed into a single vector of generalized random moments $\mathbf{s}$ (the sketch, or mean map embedding [1]), from which the learning task is then performed. For example K-means clustering [2], which aims at finding k cluster centers $C = [\mathbf{c}_1, ..., \mathbf{c}_k]$ minimizing the error $\mathrm{SSE}(X, C) = \sum_{1 \le i \le n} \min_{1 \le j \le k} \|\mathbf{x}_i - \mathbf{c}_j\|_2^2$, can be performed compressively [3] using random Fourier features, i.e. computing the sketch as

$$\mathbf{s} \triangleq \frac{1}{n} \sum_{i=1}^{n} \Phi(\mathbf{x}_i) \quad \text{with} \quad \Phi(\mathbf{x}) = \begin{bmatrix} \exp(i\boldsymbol{\omega}_j^T \mathbf{x}) \\ \vdots \\ \exp(i\boldsymbol{\omega}_m^T \mathbf{x}) \end{bmatrix},$$

where the sketch size m is a parameter and $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_m$ are frequency vectors typically drawn i.i.d. according to a multivariate normal distribution $\boldsymbol{\omega}_i \sim \mathcal{N}(0, \frac{1}{\sigma_\kappa^2}\mathbf{I}_d)$. The k centers are then recovered so that their mean sketch matches $\mathbf{s}$. Although we focus only on clustering in this paper, note that this problem bears close similarities with super-resolution (where k spikes have to be estimated precisely), and Gaussian modeling which can be performed using the same sketch.

Goal and contribution. We provide empirical insights on the choice of the parameter σ_κ^2 driving the draw of the random frequencies $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_m$. We show that it connects to the inter-cluster separation $\varepsilon \triangleq \min_{i \ne j} \|\mathbf{c}_i^* - \mathbf{c}_j^*\|_2$, where $(\mathbf{c}_1^*, ..., \mathbf{c}_k^*)$ denotes the optimal solution, and discuss some possible approaches to learn ε from the training dataset.

2 Related work

From a theoretical perspective, statistical learning guarantees have been studied [4] for compressive clustering using a sketch computed via (weighted) random Fourier features, and a separation assumption $\varepsilon > 0$ has been shown to be necessary for these guarantees to hold. Vice-versa, and more quantitatively, learning guarantees have been established when $\sigma_\kappa^2 \lesssim \varepsilon^2 / \log k$, using an interpretation of σ_κ^2 as the variance of a spatial smoothing kernel. Yet, when σ_κ^2 is too small, reconstruction algorithms can get stuck in local minima and theoretical error bounds become vacuous. When performing compressive clustering in practice, some empirical works suggested to tune σ_κ^2 using an estimate of the intra-cluster variance [5] rather than the inter-cluster separation ε^2. In certain scenarios it was also observed that the second moment of the data [6] can yield lower empirical error.

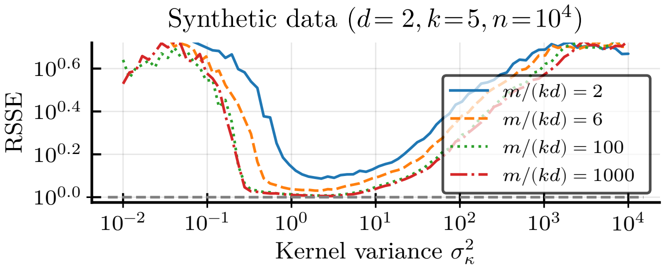

Figure 1: Impact of the sketch size m on clustering error.

3 Simulations with synthetic data

To investigate more systematically the role of the kernel variance σ_κ^2 in the clustering performance, we generate data according to the Gaussian mixture $\pi_0 \triangleq \frac{1}{k}\sum_{1 \le i \le k} \mathcal{N}(\boldsymbol{\mu}_i, \sigma_{\text{intra}}^2 \mathbf{I}_d)$, where $\boldsymbol{\mu}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_{\text{inter}}^2 \mathbf{I}_d)$. We run compressive k-means with the CLOMPR algorithm [3], and measure all the errors with the relative SSE (RSSE), i.e. we normalize the SSE by the error of the best solution obtained with the k-means algorithm (which performs multiple passes over the data).

Figure 1 shows how the clustering error varies with σ_κ^2 for different sketch sizes m. As the goal is to estimate k points in dimension d, there are kd parameters to fit in total and previous empirical work suggest indeed that $m = \Omega(kd)$ observations are required for successful recovery, so we choose the sketch sizes accordingly. One can see that the order of magnitude of the optimal variance, that we denote ${\sigma_\kappa^*}^2$, only mildly depends on m and that the range of kernel variances σ_κ^2 for which near optimal performance is achieved gets larger as the sketch size grows. Overall there is a tradeoff between the sketch size m and the precision at which the variance needs to be tuned. To achieve high compression ratios, i.e. small $m/(kd)$ with good performance it seems important to have an accurate estimate of the optimal kernel variance ${\sigma_\kappa^*}^2$.

We now fix the sketch size m and analyse the impact of other parameters on ${\sigma_\kappa^*}^2$ in Figure 2. On the left subfigure, we vary the ratio $\rho^2 \triangleq \sigma_{\text{inter}}^2 / \sigma_{\text{intra}}^2$, which drives how separated are the different clusters. We observe that at fixed k, d and up to a multiplicative constant ($\approx 1/2$), the optimal variance ${\sigma_\kappa^*}^2$ scales linearly with ρ^2. As shown by the dashed curves, the minimum inter-cluster distance and the second moment σ_X^2 of the dataset both have the same scaling. However as shown on the other subfigures, the variation of ${\sigma_\kappa^*}^2$ with the dimension d and the number of clusters k when ρ^2 is fixed seem to better fit the observed

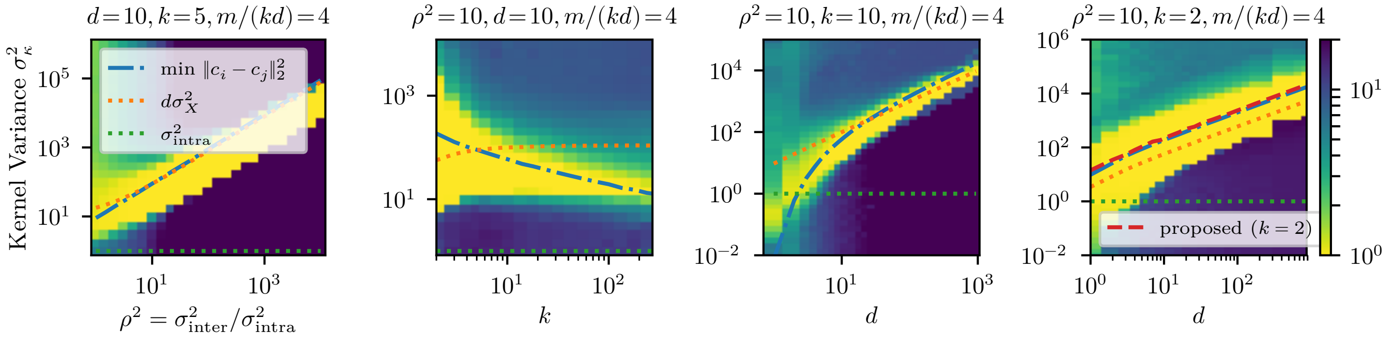

Figure 2: RSSE (lower = better = yellow) as a function of the kernel variance σ_κ^2 and the variance ratio ρ^2 (left) the number of clusters k (middle left), and the dimension d (right). Medians over 100 trials obtained with CLOMPR.

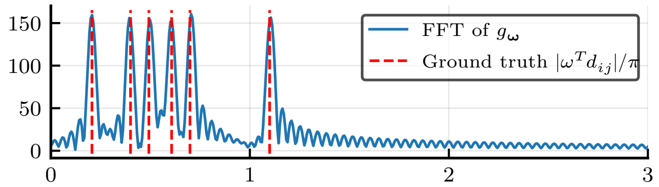

Figure 3: FFT of $g_{\boldsymbol{\omega}}$ for one random $\boldsymbol{\omega}$. $k = 4$ (6 pairs).

evolution of the minimum inter-cluster distance ε^2 with these parameters. This suggests that ε is the quantity to tune, and that the previously used heuristics of using the intra-cluster variance $\sigma_{\mathrm{intra}}^2$ or the global dataset variance (whose expectation is of the order of $d(\sigma_{\mathrm{intra}}^2 + \sigma_{\mathrm{inter}}^2)$, and thus close to the mean inter-cluster distance in our setting) are sub optimal, although they might have been observed to work well for some specific settings [5, 6].

4 Learning the separation?

We now discuss how the separation ε could be learned from the dataset. A simple method would be to run the k-means algorithm on a smaller randomly chosen subset, but this would not scale in high dimension and might perform poorly in the presence of unbalanced clusters. We focus on a proof of concept using sketches.

Imagine the data distribution is a pure mixture of diracs $P_C = \frac{1}{k}\sum_{1\leq i\leq k}\delta_{c_i}$. In the following, we denote $\varphi(\boldsymbol{\omega})$ the characteristic function of P_C at $\boldsymbol{\omega}$, and define $\boldsymbol{\mu}_{ij} \triangleq \frac{1}{2}(\mathbf{c}_i + \mathbf{c}_j), \mathbf{d}_{ij} \triangleq \frac{1}{2}(\mathbf{c}_i - \mathbf{c}_j)$ for any i, j.

When $\mathbf{k} = \mathbf{2}$, we have

$$\varphi(\boldsymbol{\omega}) \triangleq \tfrac{1}{2}(e^{i\boldsymbol{\omega}^\top \mathbf{c}_1} + e^{i\boldsymbol{\omega}^\top \mathbf{c}_2}) = e^{i\boldsymbol{\omega}^\top \boldsymbol{\mu}_{12}} \cos(\boldsymbol{\omega}^\top \mathbf{d}_{12})$$

$$\text{and } f(\boldsymbol{\omega}) \triangleq 1 - |\varphi(\boldsymbol{\omega})|^2 = \sin^2(\boldsymbol{\omega}^\top \mathbf{d}_{12})$$

Thus, if we draw $\boldsymbol{\omega} \sim \mathcal{N}(0, \sigma_\omega^2 \mathbf{I}_d)$ with $\sigma_\omega \ll 1/\|\mathbf{d}_{12}\|$, then we have with high probability $f(\boldsymbol{\omega}) \approx |\boldsymbol{\omega}^\top \mathbf{d}_{12}|^2$. Bounds on the data itself can be used to bound $\|\mathbf{d}_{12}\|$ from above and choose an appropriate variance σ_ω^2. Multiple frequencies can be used to improve the concentration, i.e. we use $2\sigma_\omega^{-1}(\sum_{1\leq i\leq m} f(\omega_i)/m)^{1/2}$ as an estimator of ε. Using in practice the true data sketch instead of the sketch of P_C is not problematic given that we are sampling only low frequencies. Estimations of ${\sigma_\kappa^*}^2$ obtained with this method on the data sketch are shown in Figure 2 (right).

When $\mathbf{k} > \mathbf{2}$, robust estimation is not straightforward but defining $f_{ij} \triangleq \frac{1}{\pi}|\boldsymbol{\omega}^T \mathbf{d}_{ij}|$ we have in a similar manner

$$g_{\boldsymbol{\omega}}(t) \triangleq \tfrac{1}{2}(k^2|\varphi(t\boldsymbol{\omega})|^2 - k) = \sum_{i<j} \cos(2\pi f_{ij} t).$$

Hence the (f_{ij}) can be recovered from the spectrum of $g_{\boldsymbol{\omega}}$ as shown in Figure 3. Recovering ε from the f_{ij} is at least straightforward in dimension $d = 1$, and a challenge is to understand how to estimate ε for $d > 1$ by combining estimations obtained in multiple directions $\boldsymbol{\omega}$, possibly leveraging the sparse FFT [7] to use as few samples of the characteric function as possible.

If all the clusters are normally distributed and have similar scales, we can model the data distribution P as the convolution of a mixture of diracs P_C (located at the cluster centers) and a Gaussian multivariate distribution $P_{\mathrm{intra}} = \mathcal{N}(0, \sigma_{\mathrm{intra}}^2 \mathbf{I}_d)$. The intra-cluster variance $\sigma_{\mathrm{intra}}^2$ can itself be estimated using a small sketch [5], hence any empirical sketch $\mathbf{s}$ measured w.r.t. P can be deconvolved by dividing it pointwise by the (analytical) sketch of P_{intra} in order to estimate the sketch of the mixture of diracs P_C.

5 Perspectives

We showed with empirical simulations that a good estimation of the minimum inter-cluster separation is essential for compressive clustering, and that information relative to this quantity is contained in the characteristic function. Building a robust algorithm to estimate the separation ε for any values of k, d is left for future work.

References

[1] K. Muandet et al. "Kernel mean embedding of distributions: A review and beyond". In: *Foundations and Trends in Machine Learning* 10.1 (2017).

[2] A. K. Jain. "Data clustering: 50 years beyond K-meansq". In: *Pattern Recognition Letters* 31 (2010).

[3] N. Keriven et al. "Compressive K-means". In: ICASSP. Mar. 5, 2017.

[4] R. Gribonval et al. "Compressive statistical learning with random feature moments". In: *arXiv:1706.07180* (2017).

[5] N. Keriven et al. "Sketching for large-scale learning of mixture models". In: *Information and Inference: A Journal of the IMA* 7.3 (2017).

[6] E. Byrne et al. "Sketched Clustering via Hybrid Approximate Message Passing". In: *IEEE Transactions on Signal Processing* 67.17 (Sept. 2019).

[7] A. C. Gilbert et al. "Recent Developments in the Sparse Fourier Transform - A compressed Fourier transform for big data." In: *IEEE Signal Process. Mag.* 31.5 (2014).