Dimension Reduction for Projective Clustering[*]

Rafael Zuolo Coppini Lima[1]

[1]Instituto de Matemática e Estatística – Universidade de São Paulo (USP)
São Paulo – SP – Brazil

rafaelzcl@ime.usp.br

Abstract. *The high dimensionality of data may be a barrier to algorithmic efficiency, mainly because of the well known "curse of dimensionality" that imposes exponential time and/or memory complexity for algorithms. It is natural then to search for ways to break this curse by relaxing the problem with approximate versions and by finding good ways to reduce the dimension of data. Our aim is to integrate and state slightly stronger results of approximation via dimension reduction for* clustering under the ℓ_2^2 metric. *The dimension reduction is achieved by combining randomized techniques (the Johnson-Lindenstrauss Lemma) and deterministic techniques (the singular value decomposition).*

1. Introduction

Let $\mathcal{C} \neq \emptyset$ be a family of non-empty sets of $\mathbb{R}^d$ and let $A \subset \mathbb{R}^d$ be a multiset with $|A| = n$. The $(\ell_2^2, \mathcal{C})$*-clustering* problem in $\mathbb{R}^d$ is to find a set $C \in \mathcal{C}$ that minimizes the *cost* function $\operatorname{dist}^2(A, C) := \sum_{a \in A} (\operatorname{dist}(a, C))^2$, where $\operatorname{dist}(a, C) := \inf \{\|a - c\| : c \in C\}$ and $\|\cdot\|$ is the usual Euclidean norm. We say that A is an *instance* and that the sets C are *solutions*. The sets C^* that minimize $\operatorname{dist}^2(A, C^*)$ are called *optimal solutions* of A.

This problem is a general formulation for clustering problems under the ℓ_2^2 metric. Let k and j be non-negative integers. Some examples of known clustering problems that fit this formulation are the following: k-means clustering, where the objective is to find a set C of k points of $\mathbb{R}^d$ that minimizes $\operatorname{dist}^2(A, C)$; best-fit j-subspace problem, where the objective is to find a subspace C of dimension j such that it minimizes $\operatorname{dist}^2(A, C)$; linear (affine) projective clustering problem, also known as linear (affine) j-subspace k-clustering problem, where the objective is to find a set C that is the union of k linear (affine) subspaces of dimension j that minimizes $\operatorname{dist}^2(A, C)$. It is known that k-means clustering is NP-hard for $k = 2$ [Aloise et al. 2009]. The best-fit j-subspace problem can be solved in time $O(\min\{nd^2, n^2d\})$ via singular value decomposition. Affine 2-subspace k-clustering is NP-hard to approximate for any multiplicative approximation factor [Megiddo and Tamir 1982].

Some of these problems have algorithms with time dependence considered inefficient on the dimension, and hence a natural idea for efficiently finding approximations for them is to "reduce the dimension" of the instance A in the following way: consider another instance $\widetilde{A}$ (called *sketch* of A) such that it lies in a low dimensional subspace of $\mathbb{R}^d$, and such that any optimal solution $\widetilde{C}^*$ for $\widetilde{A}$ is a good approximation for A. Solving the problem for $\widetilde{A}$ should be more efficient, and will give us approximations for A. In [Sarlós 2006] this idea is used to obtain a randomized $(1 + \varepsilon)$-approximation for best-fit j-subspace in time $o(\min\{nd^2, n^2d\})$, and in [Pratap and Sen 2018] it is

[*]The author was partially supported by CNPq (423833/2018-9) and by CAPES (Finance Code 001).

used to obtain a randomized algorithm that finds a sketch of any instance of projective clustering in linear time in n and d. In this work we present results of [Sarlós 2006] and [Pratap and Sen 2018] in a unified way, and we make a small improvement on the last result by generalizing it to a broader class of problems. We hope that our presentation will help further development and applications.

2. Definitions and preliminaries

2.1. Matrix notation

It will be useful to represent any set or multiset $A \subset \mathbb{R}^d$ with $|A| = n$ as a $d \times n$ matrix, where each point is a column vector of this matrix in some fixed orthonormal basis. Similarly we can see any $d \times n$ matrix as a multiset of $\mathbb{R}^d$ with cardinality n. We say that a matrix $A \in \mathbb{R}^{d\times n}$ has *orthonormal columns* if $A^T A$ is the identity matrix. Let $B \in \mathbb{R}^{d\times j}$. We denote by $\pi_B(A)$ the $d \times n$ matrix we obtain by projecting orthogonally each column of A onto the subspace spanned by the columns of B. The *Frobenius norm* of A is $\|A\|_F := \sqrt{\sum_{i=1}^{d} \sum_{j=1}^{n} A_{ij}^2}$, where A_{ij} is the entry of row i and column j.

Definition 1. *Let $A \in \mathbb{R}^{d\times n}$ be an instance for $(\ell_2^2, \mathcal{C})$-clustering problem and let $\varepsilon \in (0,1)$ be fixed. We say that a matrix $\widetilde{A} \in \mathbb{R}^{d\times n}$ is an* ε-sketch *of A if there is a non-negative constant $\Delta = \Delta(A, \mathcal{C}, \varepsilon)$ such that for every $C \in \mathcal{C}$ we have*

$$(1-\varepsilon)\operatorname{dist}^2(A, C) \le \operatorname{dist}^2(\widetilde{A}, C) + \Delta \le (1+\varepsilon)\operatorname{dist}^2(A, C). \tag{1}$$

We say that a matrix $\widetilde{A} \in \mathbb{R}^{d\times n}$ is a weak ε-sketch *of A if for every optimal solution $\widetilde{C}^*$ for $\widetilde{A}$ and any optimal solution C^* for A we have*

$$\operatorname{dist}^2(A, \widetilde{C}^*) \le (1+\varepsilon)\operatorname{dist}^2(A, C^*). \tag{2}$$

Note that A is an ε-sketch of A for all $\varepsilon \in (0,1)$ and with $\Delta = 0$, and that if $\widetilde{A}$ is an ε-sketch of A, then it is also a weak 3ε-sketch if ε is small enough.

2.2. Singular value decomposition

Fix $A \in \mathbb{R}^{d\times n}$ and let $r = \operatorname{rank}(A)$. Any such A has a *singular value decomposition*: $A = U\Sigma V^T$, where $U \in \mathbb{R}^{d\times r}$ and $V \in \mathbb{R}^{n\times r}$ have orthonormal columns (called, respectively, the *left* and *right singular vectors* of A) and $\Sigma \in \mathbb{R}^{r\times r}$ is a positive diagonal matrix containing the *singular values* of A in non-increasing order: $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r$.

For any positive integer $j < r$, let Σ_j be Σ with its last $r - j$ diagonal elements zeroed out. Let U_j and V_j be, respectively, the matrices U and V with all but the first j columns zeroed out. Then $A_j := U\Sigma_j V^T = U_j \Sigma_j V_j^T$ is a closest rank j approximation of A in any unitarily invariant norm, in particular in the Frobenius norm:

$$\|A - A_j\|_F = \min\{\|A - B\|_F : B \in \mathbb{R}^{d\times n},\ \operatorname{rank}(B) = j\}.$$

Some important properties of the SVD is that the non-null columns of U_j form an orthonormal basis of an optimal solution of best-fit j-subspace problem for the instance A, and the matrix A_j is equal to $\pi_{U_j}(A)$.

2.3. Johnson-Lindenstrauss lemma

The Johnson-Lindenstrauss Lemma [Johnson and Lindenstrauss 1984] is the following result.

Theorem 2 (Johnson-Lindenstrauss Lemma)**.** *There exists a constant κ such that for any set A of n points in $\mathbb{R}^d$, any $\varepsilon \in (0,1)$ fixed and all integers $r \geq \kappa\varepsilon^{-2}\log n$ there exists a linear function $f\colon \mathbb{R}^d \to \mathbb{R}^r$ such that for every pair $x, y \in A$ we have*

$$(1-\varepsilon)\|x-y\|^2 \leq \|f(x)-f(y)\|^2 \leq (1+\varepsilon)\|x-y\|^2. \tag{3}$$

Linear functions that satisfy Theorem 2 can be found via random matrices: let $\beta > 0$ be fixed and let r be an integer such that $r \geq \frac{4+2\beta}{\varepsilon^2/2-\varepsilon^3/3}\log n$. If $S \in \mathbb{R}^{r\times d}$ is a random matrix where each entry is a independent random variable that assumes values uniformly in $\{1/\sqrt{r}, -1/\sqrt{r}\}$, then the function $v \mapsto Sv$ satisfies Theorem 2 with probability at least $1-n^{-\beta}$ [Achlioptas 2003].

3. Low rank weak sketch for the best-fit j-subspace problem

Suppose without loss of generality that $d \leq n$. Solving the best-fit j-subspace problem for any instance $A \in \mathbb{R}^{d\times n}$ with SVD requires just the first j left singular vectors, but finding them takes time $O(nd^2)$. We present now a result from [Sarlós 2006] on finding a $(1+\varepsilon)$-approximation of A_j in linear time on n and d when $j = O(1)$.

Theorem 3 ([Sarlós 2006])**.** *Let $A \in \mathbb{R}^{d\times n}$ be a matrix, let $j < \min\{d, n\}$ be a positive integer and let $\varepsilon \in (0,1)$ be fixed. There is an integer $r = \Theta(\varepsilon^{-1}j + j\log j)$ such that if S is an $r \times n$ random matrix where each entry is an independent random variable that assumes values uniformly in $\{1/\sqrt{r}, -1/\sqrt{r}\}$, then with probability at least $1/2$ we have*

$$\|A - (\pi_{AS^T}(A))_j\|_F \leq (1+\varepsilon)\|A - A_j\|_F.$$

Computing $(\pi_{AS^T}(A))_j$ can be done with two readings of the matrix A and in time $O(ndr + (n+d)r^2)$.

Note that $\pi_{AS^T}(A)$ is a weak 3ε-sketch of A for the best-fit j-subspace problem (see (2)). The probability of success in Theorem 3 can be boosted to $1-\delta$, for any $\delta \in (0,1)$, since by Pythagoras' Theorem $\|A\|_F^2 = \|A - (\pi_{AS^T}(A))_j\|_F^2 + \|(\pi_{AS^T}(A))_j\|_F^2$. Therefore if we run $\log_2(1/\delta)$ independent instances of S and choose the one that maximizes $\|(\pi_{AS^T}(A))_j\|_F^2$, we will have a probability of success of $1-\delta$.

4. Low-rank sketch for the $(\ell_2^2, \mathcal{C})$-clustering problem

We now present a low-rank ε-sketch that can be found in time linear on n and d. We say that a family $\mathcal{C}$ of subsets of $\mathbb{R}^d$ is m dimensional if for every $C \in \mathcal{C}$ there exists a subspace $L = L(C)$ such that C is contained in L. For example, the family $\mathcal{C}$ implicit in the k-means clustering problem is k dimensional, while in the affine j-subspace k-clustering it is $(j+1)k$ dimensional.

Theorem 4 (based on [Pratap and Sen 2018]). *Let $A \in \mathbb{R}^{d\times n}$ be an instance for the $(\ell_2^2, \mathcal{C})$-clustering problem in $\mathbb{R}^d$, where $\mathcal{C}$ is an m dimensional family and $m < \min\{d, n\}$. Let $\varepsilon \in (0,1)$ be fixed. There exists an integer $s = \lceil \varepsilon^{-2}m/8 \rceil$ such that if a matrix $\widetilde{A}^T \in \mathbb{R}^{n\times d}$ is an orthogonal projection of A^T to some subspace of dimension s of $\mathbb{R}^n$ and satisfies*

$$\left\|A - \widetilde{A}\right\|_F^2 \leq \left(1 + \frac{\varepsilon^2}{8}\right) \|A - A_s\|_F^2 , \tag{4}$$

then $\widetilde{A}$ is an ε-sketch of A with constant $\Delta = \|A - A_s\|_F^2$.

Note that taking $\widetilde{A} = A_s$ trivially satisfies equation (4), and thus A_s is an ε-sketch of A. But this takes time $O(nd^2)$, and as we saw in the previous section this can be improved. Using Theorem 3 it follows that taking $\widetilde{A} = (\pi_{A^TS^T}(A^T))_s^T$ satisfies inequality (4) with probability at least $1/2$ and can be computed in time

$$O\left(nd(\varepsilon^{-4}m + \varepsilon^{-2}m\log(\varepsilon^{-2}m)) + (n+d)(\varepsilon^{-8}m^2 + \varepsilon^{-4}m^2\log^2(\varepsilon^{-2}m))\right).$$

5. Final remarks

The ε-sketch is a powerful tool that can be applied to any $(\ell_2^2, \mathcal{C})$-clustering problem to reduce the dimension of any instance. This is useful not only for approximation algorithms, but also for *succinct representation of data* with a technique known as *coresets*. A coreset construction for the projective clustering problem which uses the singular value decomposition as dimension reduction tool is developed in [Feldman et al. 2020]. This construction can be improved with Theorem 4 by accelerating the dimension reduction step. More information can be found in [Pratap and Sen 2018].

References

Achlioptas, D. (2003). Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687.

Aloise, D., Deshpande, A., Hansen, P., and Popat, P. (2009). NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248.

Feldman, D., Schmidt, M., and Sohler, C. (2020). Turning big data into tiny data: Constant-size coresets for k-means, PCA, and projective clustering. *SIAM J. Comput.*, 49(3):601–657.

Johnson, W. B. and Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, volume 26 of *Contemp. Math.*, pages 189–206. Amer. Math. Soc., Providence, RI.

Megiddo, N. and Tamir, A. (1982). On the complexity of locating linear facilities in the plane. *Operations Research Letters*, 1(5):194–197.

Pratap, R. and Sen, S. (2018). Faster coreset construction for projective clustering *via* low-rank approximation. In *International Workshop on Combinatorial Algorithms*, pages 336–348. Springer.

Sarlós, T. (2006). Improved approximation algorithms for large matrices via random projections. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 143–152. IEEE.