Literature

Vasily Ilin

July 8, 2020

1 Rademacher, 2005

[6] introduces fast approximate SVD via iterative sampling. They obtain bound

$$||A - \tilde{A}_k||^2 \leq \frac{1}{1-\epsilon}||A - A_k||^2 + \epsilon^t ||A||^2$$

with at least 3/4 probability via iterative sampling, where $\tilde{A}_k$ is the sampled matrix. The first round of sampling samples row i of A with probability proportional to lengths squared. The second round of sampling samples row i with probability proportional to the square of length of the residue of row i, after its projection onto the span of sampled rows. There are t such rounds of sampling, with k/ϵ rows sampled each round. Thus, there are tk/ϵ rows in $\tilde{A}_k$. The algorithm **does not** compute the actual SVD of A, as opposed to the algorithm from [4]. Intuitively, if an important direction is not accounted for in round i, it will have a higher chance of being sampled in round $i+1$. They also show existence of $4k^2/\epsilon$ rows that span matrix $\tilde{A}_k$ such that

$$||A - \tilde{A}_k||^2 \leq (1+\epsilon)||A - A_k||^2 = (1+\epsilon)\text{OPT}.$$

Even though the claim itself is weaker than that in [4] (mere existence vs an algorithm finding it w.h.p.), the technique used is different from [4], and can be of independent interest, In particular, this provides a certain guarantee of algorithm **??**.

2 Drineas, 2007

[4] does matrix decomposition using rows of the matrix as principal vectors. They first **compute the SVD**, and then sample rows, s.t. row i is sampled with probability proportional to the square of length of its projection onto the subspace spanned by principal vectors. They have two algorithms: one that samples $O(k \log k \log(1/\delta)/\epsilon^2)$ rows

in expectation, and another that samples exactly $O(k^2 \log(1/\delta)/\epsilon^2)$ rows They obtain multiplicative bound

$$||A - \tilde{A}_k||^2 \leq (1+\epsilon)||A - A_k||^2 = (1+\epsilon)\mathrm{OPT}$$

with probability at least $1 - \delta$.

3 Bernholt, 2004

[1] talks about minimum covariance determinant(MCD): they introduce an algorithm polynomial with running time $O(N^{\nu+1})$, where $\nu = d(d+3)/2$, and show that MCD is **NP**-hard if dimension d varies.

Definition 3.1 (MCD)**.** Let $\mathcal{X} \subset \mathbb{R}^d$ be a set of N points. Given natural number h, the *minimum covariance determinant problem* is to find an h-element subset $\mathcal{X}' \subset \mathcal{X}$ such that $\det(\mathcal{X}')$ is minimal, where $\det(S)$ is the determinant of the matrix SS^T, where S is both a set and a matrix, with columns from S. $(SS^T)_{d \times d}$ is the covariance matrix of S.

Definition 3.2 (Quadric)**.** A *quadric* Q in $\mathbb{R}^d$ is the set describable by a second-degree polynomial

$$a_0 + \sum_{i=1}^{d} a_i z_i + 2 \sum_{1 \leq i < j \leq d} a_{i,j} z_i z_j + \sum_{i=1}^{d} a_{i,i} z_i^2 = 0, \tag{1}$$

where $a_i, a_{i,j}$ for $1 \leq i \leq j \leq d$ are the coefficients defining the quadric, and $z = (z_1, ..., z_d) \in \mathbb{R}^d$ is a point in Q. Equivalently, it is the set $\left\{z \in \mathbb{R}^d : z^T A z + z^T b + a_0 = 0\right\}$, where $A_{i,j} = a_{i,j}$ is a symmetric matrix, and $b_i = a_i$ is a column vector. Q is determined by $\nu + 1$ parameters $a_0, ..., a_d$, $a_{i,j}$ for $1 \leq i \leq j \leq d$, where $\nu = d(d+3)/2$. It is a generalization of an ellipsoid, and is a $(d-1)$-dimensional (smooth) manifold.

An important fact they note in this paper is that selection by a quadric in d dimensions is equivalent to selection by a hyperplane in ν dimensions. Indeed, consider the map$\widehat{\ }$: $\mathbb{R}^d \to \mathbb{R}^\nu$ given by

$$\widehat{z} = (z_1, \dots, z_d, z_1 z_1, \dots, z_i z_j, \dots, z_d z_d), \ 1 \leq i \leq j \leq d.$$

Then the parameters $a_i, a_{i,j}$ define a hyperplane in $\mathbb{R}^\nu$ with the property that

$$z \text{ selected by quadric } Q \iff \widehat{z} \text{ is selected by the hyperplane} \iff \langle \widehat{z}, \widehat{a} \rangle + a_0 \leq 0,$$

where $\widehat{a} = (a_1, \dots, a_d, a_{1,1}, \dots, a_{i,j}, \dots, a_{d,d})$

Since by [7] we know that a set that solves MCD is selectable by an ellipsoid, it is enough to find an ellipsoid that selects the optimal set. They show that any set selectable by an ellipsoid is (almost) selectable by a quadric with the coefficients $a_i, a_{i,j} \subset \mathcal{X}$, and there are $O(N^{\nu+1})$ of those. The precise statement is

Theorem 3.3. Given a solid ellipsoid E, let $\mathcal{X}' = E \cap \mathcal{X}$, so $\mathcal{X}'$ is selected by E. Then there exists set $S \subset \mathcal{X}$ of $\nu := d(d+3)/2$ elements such that for the quadric defined by A, b, a_0 we have

$$x^T Ax + x^T b + a_0 \leq 0 \text{ for all } x \in \mathcal{X}'$$
$$x^T Ax + x^T b + a_0 \geq 0 \text{ for all } x \in \mathcal{X} \setminus \mathcal{X}'.$$

The proof comes from the fact that separation by a quadric in $\mathbb{R}^d$ is equivalent to separation by a hyperplane in $\mathbb{R}^\nu$ by mapping all $x \in \mathbb{R}^d$ to $\hat{x} = (x_1, ..., x_d, x_{1,1}, ..., x_{i,j}, ..., x_{d,d}) \in \mathbb{R}^\nu, 1 \leq i \leq j \leq d$.

So it is enough to check all subsets of $\mathcal{X}$ of size ν to uncover the quadric selecting the optimal set (gotta be careful about the boundary terms x s.t. $x^T Ax + x^T b + a_0 = 0$ but it works out).

4 Rousseeuw, 1999

[7] shows that any subset that minimizes the covariance determinant is selectable by an ellipsoid (corollary 1).

They also introduce a C-step, which, given a set, will find a new set with smaller or equal covariance determinant. It is done by selecting the h elemens with the smallest Mahalanobis distance.

Additionally, they present a fast algorithm for approximating MCD, which is iterative and finds a local optimum. They run a lot of randomly initialized parallel computations as an attempt to find the global optimum.

Their algorithm is to initialize a lot of subsets $\mathcal{X}'$ of the original set $\mathcal{X}$, and apply C-steps to each one until convergence. To initialize, for each subset $\mathcal{X}'$ they select k values u.i.r., compute the covariance matrix, and add $n - k$ values from $\mathcal{X}$ that have the lowest Mahalanobis distance.

5 chawla, 2013

[3] introduce k-means–, which works with outliers. They claim that for $k = 1, \ell > 1$, k-means– is a special case of MCD. This is true because the set $\mathcal{X}' = \mathcal{X} \setminus \{\text{outliers}\}$ is selectable by an ellipsoid (a sphere, in fact).

6 Hubert and Rousseeuw, 2005

[5] consider a problem quite similar to ours: given a set $\mathcal{X} \subset \mathbb{R}^d$ and integers k and $0.5|\mathcal{X}| \leq h$, find a subset of $\mathcal{X}$ of size h that has the best PCA approximation. Unfortunately, they never actually state the problem precisely, so it is unclear how they measure the goodness of PCA approximation of a set. Their algorithm is quite convoluted, but is quite different from all algorithms we have thought of for our problem. A big difference between their problem and ours is that their require $0.5|\mathcal{X}| \leq h$, and think of all elements that are discarded as outliers. We, on the other

hand, might choose h to be small, and think of the chosen points as a community, without labeling all other points as outliers.

In their algorithm they use a measure of "outlyingness“ for each point to select those are are not outliers. It is given by

$$\text{outl}_A(x_i) = \max_{v \in B} \frac{|x_i^T v - \text{med}(x_j^T v)|}{\text{mad}(x_j^T v)},$$

where B contains all non-zero vectors, $\text{med}(x_j^T v)$ is the median of $\left\{x_j^T v : x_j \in \mathcal{X}\right\}$, and $\text{mad}(x_j^T v) = \text{med}|x_j^T v - \text{med}(x_\ell^T v)|$.

7 Candes, 2011, Robust PCA

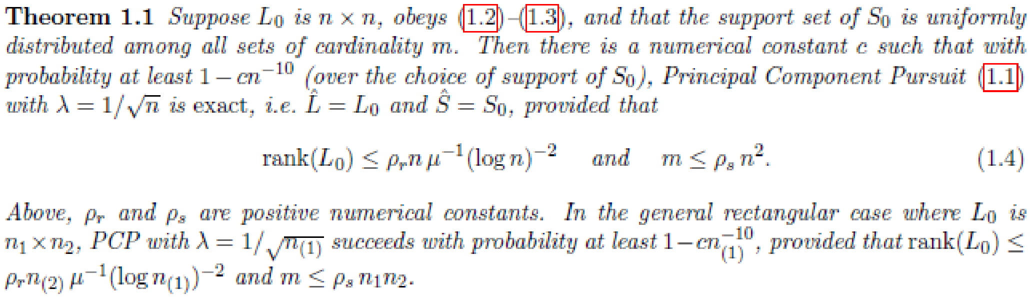
[2] describes an algorithm to find low-rank L_0 and sparse N_0 given $M_{n_1 \times n_2} = L_0 + N_0$. They assume the incoherence condition with parameter μ, which is that for $L_0 = U \Sigma V^* = \sum_{i=1}^{r} \sigma_i u_i v_i^*$, the following hold:

$$\max_i ||U^* e_i||^2 \leq \frac{\mu r}{n_1}, \quad \max_i ||V^* e_i||^2 \leq \frac{\mu r}{n_2}, \quad ||UV^*||_\infty \leq \sqrt{\frac{\mu r}{n_1 n_2}}, \tag{2}$$

where $||.||_\infty$ is the ℓ_∞ norm, i.e. the largest absolute value, and $||.||$ is the Frobenius norm. The incoherence condition 2 ensures that singular vectors are not sparse, i.e. their entries are more or less uniform in absolute value. They solve the optimization problem

$$\text{minimize } ||L||_* + \lambda ||S||_1$$
$$\text{subject to } L + S = M,$$

which under specific conditions provides an *exact* solution.



Theorem 1.1 *Suppose L_0 is $n \times n$, obeys (1.2)–(1.3), and that the support set of S_0 is uniformly distributed among all sets of cardinality m. Then there is a numerical constant c such that with probability at least $1 - cn^{-10}$ (over the choice of support of S_0), Principal Component Pursuit (1.1) with $\lambda = 1/\sqrt{n}$ is* exact, *i.e. $\hat{L} = L_0$ and $\hat{S} = S_0$, provided that*

$$\operatorname{rank}(L_0) \leq \rho_r n \, \mu^{-1} (\log n)^{-2} \quad \text{and} \quad m \leq \rho_s \, n^2. \tag{1.4}$$

Above, ρ_r and ρ_s are positive numerical constants. In the general rectangular case where L_0 is $n_1 \times n_2$, PCP with $\lambda = 1/\sqrt{n_{(1)}}$ succeeds with probability at least $1 - cn_{(1)}^{-10}$, provided that $\operatorname{rank}(L_0) \leq \rho_r n_{(2)} \, \mu^{-1} (\log n_{(1)})^{-2}$ *and $m \leq \rho_s \, n_1 n_2$.*

This is the main theorem:

References

[1] Thorsten Bernholt and Paul Fischer. “The complexity of computing the MCD-estimator”. In: *Theoretical Computer Science* 326.1-3 (2004), pp. 383–398.

[2] Emmanuel J Candès et al. “Robust principal component analysis?” In: *Journal of the ACM (JACM)* 58.3 (2011), pp. 1–37.

[3] Sanjay Chawla and Aristides Gionis. “k-means–: A unified approach to clustering and outlier detection”. In: *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM. 2013, pp. 189–197.

[4] Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. *Relative-Error CUR Matrix Decompositions*. 2007. arXiv: `0708.3696 [cs.DS]`.

[5] Mia Hubert, Peter J Rousseeuw, and Karlien Vanden Branden. “ROBPCA: a new approach to robust principal component analysis”. In: *Technometrics* 47.1 (2005), pp. 64–79.

[6] Luis Rademacher, Santosh Vempala, and Grant Wang. “Matrix Approximation and Projective Clustering via Iterative Sampling”. In: (Dec. 2005).

[7] Peter J Rousseeuw and Katrien Van Driessen. “A fast algorithm for the minimum covariance determinant estimator”. In: *Technometrics* 41.3 (1999), pp. 212–223.