Matrix Approximation and Projective Clustering via Volume Sampling

SODA 2006

Amit Deshpande, Luis Rademacher, Santosh Vempala, Grant Wang
Mathematics Department and CSAIL, MIT
Outline.

- *The matrix approximation and projective clustering problems and their motivations.*

- Our results. The additive error of matrix approximation drops exponentially as a function of the number of passes. Existence of a small sample of rows containing a relative approximation. A PTAS for projective clustering.
Matrix Approximation. Motivation.

- Given points in $\mathbb{R}^m$, find lower dimensional “representation”: a subspace such that the points are close to it ... 
- ... to “highlight” relevant features of data, obtain computational savings, and improve quality of retrieval.
- One formalization, minimum squares: see the points as rows of a matrix $A$ and find $\tilde{A}$ of rank $k$ that minimizes 

\[
\|A - \tilde{A}\|_F^2 = \sum_{ij} (A_{ij} - \tilde{A}_{ij})^2
\]
Singular Value Decomposition (SVD)

- Such minimization is solved by the SVD.
- SVD: any $m \times n$ real matrix $A$ can be written as

$$A = \sum_i \sigma_i u_i v_i^T$$

where $(u_i)_i$ orthonormal (left singular vectors), $(v_i)_i$ orthonormal (right singular vectors) and $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$

- Then the optimum for the approximation problem is

$$\tilde{A} = AY Y^T$$

where the columns of $Y$ are the top $k$ right singular vectors of $A$. 
SVD. Running time.

- SVD takes time $O(mn^2)$. Still too large for some applications; ...
- ... we could be satisfied with an *approximation* to the best, given in an implicit representation, obtained after only a few passes over the data.
A related problem, projective clustering: given \( n \) points in \( \mathbb{R}^d \), find \( j \)-dimensional subspaces that minimize the sum of squared distances of each point to its nearest subspace.

- \( j = 1 \) is matrix approximation,
- \( j \geq 2 \) is NP-hard (even for \( k = 1 \)).
Related Work.

- For matrix approximation:
  - [Drineas, Frieze, Kannan, Vempala.] Introduced matrix sampling for fast low-rank approximation.
  - [Achlioptas and McSherry.] Sparsification, uses only one pass.

- For projective clustering:
  - Multiple results for “$j$-means” (find $j$ points), and $k = 1$ (find $j$ lines)
  - [Har-Peled and Varadarajan.] A $1 + \epsilon$ approximation algorithm for the “maximum distance” objective function in time $d n^{O(j k^6 \log(1/\epsilon)/\epsilon^5)}$. 
Related Work.

- Two questions for matrix approximation:
  - Is there a small subset of rows in whose span lies a good low rank approximation?
  - Can such a subset be found efficiently?

- A result by Frieze, Kannan and Vempala gives an answer:
  **Theorem 1.** Let $S$ be a sample of $\frac{k}{\epsilon}$ rows where
  \[ \mathbb{P}(\text{row } i \text{ is picked}) = \frac{\|A^{(i)}\|^2}{\|A\|^2_F}. \]
  Then the span of $S$ contains* a matrix $\tilde{A}$ of rank $k$ for which
  \[ \mathbb{E}(\|A - \tilde{A}\|^2_F) \leq \|A - A_k\|^2_F + \epsilon \|A\|^2_F. \]

This can be turned into an efficient algorithm: 2 passes, complexity $O(nk^2/\epsilon^4)$. 

* Denotes a small constant factor.
Our Results.

- The additive error of matrix approximation drops exponentially in the number of passes and one can find a sample with the corresponding guarantee efficiently. The factor of the additive term is less than $\epsilon$ ...

<table>
<thead>
<tr>
<th>FKV</th>
<th>after 2 passes and $k\frac{1}{\epsilon}$ samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>our result</td>
<td>after $2 \log(1/\epsilon)$ passes and $k \log \frac{1}{\epsilon}$ samples.</td>
</tr>
</tbody>
</table>

- There exists a set of rows of size $O(k^2/\epsilon)$ in whose span lies a matrix that is no worse than $(1 + \epsilon)$ times the best.

- Projective Clustering: first PTAS for any fixed $j$ and $k$. Complexity: $d \left( \frac{n}{\epsilon} \right)^O(jk^3/\epsilon)$
The matrix approximation and projective clustering problems and their motivations.

Our results. The additive error of matrix approximation drops exponentially as a function of the number of passes. Existence of a small sample of rows containing a relative approximation. A PTAS for projective clustering.
Adaptive Sampling.

- Idea: Sample a few rows, then sample with weights proportional to the error that remains from the previous samples.
Adaptive Sampling.

**Theorem 2.** Let $S = S_1 \cup \cdots \cup S_t$ be a random sample of rows of an $m \times n$ matrix $A$ where for $j = 1, \ldots, t$, each set $S_j$ is a sample of $s$ rows of $A$ chosen independently from the following distribution: row $i$ is picked with probability

$$P_i^{(j)} = \frac{\|E_{j}^{(i)}\|^2}{\|E_{j}\|^2_F}$$

where $E_1 = A$, $E_j = A - \pi_{S_1 \cup \cdots \cup S_{j-1}}(A)$. Then for $s \geq k/\epsilon$, the span of $S$ contains a matrix $\tilde{A}_k$ of rank $k$ such that

$$\mathbb{E}_S(\|A - \tilde{A}_k\|^2_F) \leq \frac{1}{1 - \epsilon} \|A - A_k\|^2_F + \epsilon t \|A\|^2_F.$$

**Complexity:** $O \left( \frac{Mkt}{\epsilon} + (m + n)k^2t^2/\epsilon^2 \right)$ ($M =$ number of non-zeros).
Proof Idea. Inductive Step.

Proof Idea: Induction and use the following theorem for the inductive step:

**Theorem 3.** Let $A \in \mathbb{R}^{m \times n}$. Let $V \subseteq \mathbb{R}^n$ be a vector subspace. Let $E = A - \pi_V(A)$. Let $S$ be a random sample of $s$ rows of $A$ from a distribution such that row $i$ is chosen with probability

$$P_i = \frac{\|E^{(i)}\|^2}{\|E\|_F^2}. \quad (1)$$

Then, for any nonnegative integer $k$,

$$E_S(\|A - \pi_V + \text{span}(S),k(A)\|_F^2) \leq \|A - A_k\|_F^2 + \frac{k}{s}\|E\|_F^2.$$

The proof of the inductive step is very similar to the proof of FKV.
Volume Sampling, Arbitrary $k$.

- “In any matrix there are $k$ rows such that the projection of the matrix to those rows is a $k + 1$ approximation to $A_k$, the best of rank $k$”. More precisely (probabilistic method),

**Theorem 4.** Let $S$ be a random subset of $k$ rows of $A$ so that

$$\mathbb{P}(S \text{ is picked}) = \frac{\text{vol}(\Delta(S))}{\sum_{T:|T|=k} \text{vol}(\Delta(T))^2}.$$

Then $\tilde{A}_k$, the projection of $A$ to the span of $S$, satisfies

$$\mathbb{E}(\|A - \tilde{A}_k\|_F^2) \leq (k + 1)\|A - A_k\|_F^2.$$

- Tight: factor $k + 1$ is best possible.
The previous result combined with the “inductive step” gives

**Theorem 5.** For any $A$, there exists a subset of $O(k^2/\epsilon)$ rows in whose span lies a rank-$k$ matrix $\tilde{A}_k$ such that

$$||A - \tilde{A}_k||_F^2 \leq (1 + \epsilon)||A - A_k||_F^2.$$
The $k + 1$ approximation theorem gives a simple $k + 1$ approximation for projective clustering:
If the best partition of the points into $j$ subsets is $P_1, \ldots, P_j$, then each $P_i$ contains a subset $S_i$ of $k$ points whose span is a $k + 1$ approximation.
We can find the $S_i$’s by enumerating all subsets of $k$ points, considering $j$ of these subsets at a time, and taking the best of these.

- Complexity: $O(dn^{jk})$. 

\[
\begin{array}{c}
\text{Projective Clustering.}
\end{array}
\]
The $1 + \epsilon$ approximation theorem gives that there exists a subset $\hat{P}_i \subseteq P_i$ of size $O(k^2/\epsilon)$ in whose span lies an approximately optimal $k$-dimensional subspace.

We enumerate over all combinations of $j$ subsets, each of size $O(k^2/\epsilon)$ to find the $\hat{P}_i$.

We cannot enumerate then all the $k$-dimensional subspaces of the span of $\hat{P}_i$, but we can put an appropriate $\epsilon$-net and enumerate over subspaces induced by this net.

Complexity: $d\left(\frac{n}{\epsilon}\right)^{O(jk^3/\epsilon)}$. 
Conclusion: Summary and Open Problems.

Summary:
- The additive error of matrix approximation drops exponentially with the number of passes.
- Existence of $O(k^2/\epsilon)$ rows containing a relative approximation.
- A PTAS for projective clustering.

Open problems:
- Lower bound for multiplicative error, $k^2/\epsilon$?
- Is there an efficient implementation of volume sampling, or another efficient algorithmic way of getting a multiplicative approximation?
- Fix the mismatch of the exponents of the projective clustering approximations for $\epsilon = k$ and for arbitrary $\epsilon$. 