Average Case Column Subset Selection for Entrywise ℓ_1-Norm Loss

Zhao Song*, David P. Woodruff†, Peilin Zhong‡

*University of Washington, †Carnegie Mellon University, ‡Columbia University

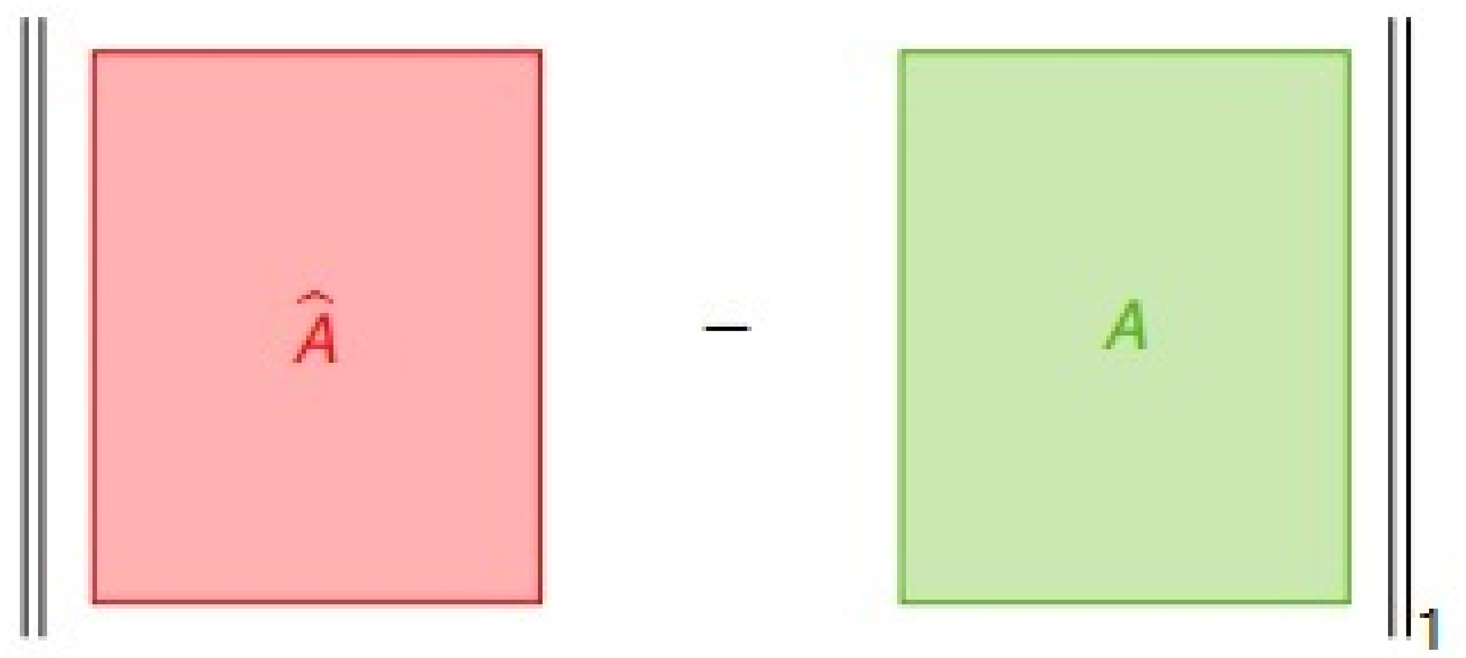
Problem Formulation

Given : $A \in \mathbb{R}^{n\times n}$, $k \geq 1$, $\alpha \geq 1$
Output : a rank-k matrix $\hat{A} \in \mathbb{R}^{n\times n}$ such that

$$\|\hat{A} - A\|_1 \leq \alpha \cdot \min_{\text{rank}-k\ A'} \|A' - A\|_1,$$

where

$$\|A\|_1 = \sum_{i,j} |A_{i,j}|.$$



- Unfortunately, a prior work shows in the worst case that a $2^{k^{\Omega(1)}}$ running time is necessary for any constant approximation given a standard conjecture in complexity theory.

Column Subset Selection

Given : $A \in \mathbb{R}^{n\times n}$, $k \geq 1$, $\alpha \geq 1$
Output : a subset of k' columns C of A such that

$$\min_X \|CX - A\|_1 \leq \alpha \cdot \min_{\text{rank}-k\ A'} \|A' - A\|_1.$$

Want k' as small as possible. Column subset selection gives a bicriteria solution for low rank approximation.

Distributional Assumption

- We propose an efficient bicriteria $(1+\epsilon)$-approximate column subset selection algorithm for the ℓ_1-norm.
- We bypass the running time lower bound mentioned above by making a mild assumption on the input data, and also show that our assumption is necessary in a certain sense.
- Suppose A can be decomposed into $A^* + \Delta$.
- $\text{rank}(A^*) = k$.
- $\Delta_{i,j}$ are i.i.d. symmetric random variables.
- $\mathbf{E}[|\Delta_{i,j}|] = 1$.
- $\mathbf{E}[|\Delta_{i,j}|^p] = O(1)$ for $p > 1$.

Algorithmic Result

Let $A = A^* + \Delta \in \mathbb{R}^{n\times n}$, where $\text{rank}(A^*) = k$, and where Δ is a matrix for which the $\Delta_{i,j}$ are i.i.d. symmetric random variables with $\mathbf{E}[|\Delta_{i,j}|] = 1$ and $\mathbf{E}[|\Delta_{i,j}|^p] = O(1)$ for $p > 1$. There is a linear-time algorithm which outputs a subset $S \subset [n]$ with $|S| \leq \text{poly}(k/\epsilon) + O(k \log n)$ with:

$$\min_{X \in \mathbb{R}^{|S|\times n}} \|A_S X - A\|_1 \leq (1+\epsilon)\|\Delta\|_1,$$

holds with probability at least $99/100$.

Necessity of $p > 1$

Let $A = \gamma \cdot \mathbf{1} \cdot \mathbf{1}^\top + \Delta \in \mathbb{R}^{n\times n}$ be a random matrix where $\gamma = n^{c_0}$ for a sufficiently large constant c_0, and $\forall i,j \in [n]$, $\Delta_{i,j} \sim C(0,1)$ are i.i.d. standard Cauchy random variables. Let $r = n^{o(1)}$. Then with probability at least $1 - O(1/\log\log n)$, $\forall S \subset [n]$ with $|S| = r$,

$$\min_{X \in \mathbb{R}^{r\times n}} \|A_S X - A\|_1 \geq 1.002 \cdot \|\Delta\|_1$$

Algorithm

1: $s \leftarrow \text{poly}(k/\epsilon)$
2: Sample a set I from $\binom{[n]}{s}$ uniformly at random
3: $\widehat{X} \leftarrow \min_{X \in \mathbb{R}^{|I|\times n}} \|A_I X - A\|_1$
4: Compute $T = \{i \in [n] \mid \|A_I \widehat{X}_i - A_i\|_1 > (1+\Theta(\epsilon))n\}$
5: $\widehat{A}_T \leftarrow \text{L1ApproxLowRank}(A_T, n, k)$
6: Return $\widehat{A}_T$ as the approximation of columns in T, and $A_I \widehat{X}_{[n]\setminus T}$ as the approximation of columns in $[n]\setminus T$

Properties of the Noise Matrix

- Lower bound on its norm:
  - Let $\Delta \in \mathbb{R}^{n\times n}$ be a matrix where $\Delta_{i,j}$ are i.i.d. samples from a symmetric distribution. Suppose $\mathbf{E}[|\Delta_{i,j}|] = 1$ and $\mathbf{E}[|\Delta_{i,j}|^p] = O(1)$ for $p > 1$. Then, $\forall \epsilon \in (0, 1/2)$,
  $$\Pr\left[\|\Delta\|_1 \geq (1-\epsilon)n^2\right] \geq 1 - e^{-\Theta(n)}.$$
- Averaging "reduces" noise:
  - Let $\Delta_1, \Delta_2, \cdots, \Delta_t \in \mathbb{R}^n$ be t random vectors. The $\Delta_{i,j}$ are i.i.d. symmetric random variables with $\mathbf{E}[|\Delta_{i,j}|] = 1$ and $\mathbf{E}[|\Delta_{i,j}|^p] = O(1)$ for $p > 1$. Let $\alpha_1, \alpha_2, \cdots, \alpha_t \in [-1, 1]$ be t real numbers.
  If $\forall i \in [n], j \in [t], |\Delta_{i,j}| \leq n^{1/2+1/(2p)}$,
  then
  $$\Pr\left[\|\sum_{i=1}^t \alpha_i \Delta_i\|_1 \leq O(t^{1/p} n)\right] \geq 1 - 2^{-n^{\Theta(1)}}.$$
- Only a small number of columns have large entries:
  - Let $\Delta \in \mathbb{R}^{n\times n}$ be a matrix where the $\Delta_{i,j}$ are i.i.d. symmetric random variables with $\mathbf{E}[|\Delta_{i,j}|] = 1$ and $\mathbf{E}[|\Delta_{i,j}|^p] = O(1)$ for $p > 1$. Let
  $$H = \{j \in [n] \mid \exists i \in [n], |\Delta_{i,j}| > n^{1/2+1/(2p)}\}.$$
  Then
  $$\Pr\left[|H| \leq O(n^{1-(p-1)/2})\right] \geq 0.999.$$
- Entrywise ℓ_1-cost of all columns containing large entries can be bounded:
  - Let $\Delta \in \mathbb{R}^{n\times n}$ be a matrix where $\Delta_{i,j}$ are i.i.d. symmetric random variables with $\mathbf{E}[|\Delta_{i,j}|] = 1$ and $\mathbf{E}[|\Delta_{i,j}|^p] = O(1)$ for $p > 1$. Let $r \geq (1/\epsilon)^{1+1/(p-1)}$. With probability $.999$,
  $$\forall S \subset [n] \text{ with } |S| \leq n/r, \sum_{j \in S} \|\Delta_j\|_1 = O(\epsilon n^2).$$
- Cost of good noise columns is small:
  - Let $\Delta \in \mathbb{R}^n$ be a vector where the Δ_i are i.i.d. symmetric random variables with $\mathbf{E}[|\Delta_i|] = 1$ and $\mathbf{E}[|\Delta_i|^p] = O(1)$ for $p > 1$. Let $\epsilon \in (0,1)$ satisfy $1/\epsilon = n^{o(1)}$. If $\forall i \in [n], |\Delta_i| \leq n^{1/2+1/(2p)}$, then
  $$\Pr[\|\Delta\|_1 \leq (1+\epsilon)n] \geq 1 - 2^{-n^{\Theta(1)}}.$$

Median Heuristic

The take-home message from our theoretical analysis is that although the noise distribution may be heavy-tailed, if the p-th $(p > 1)$ moment of the distribution exists, averaging the noise may reduce the noise.

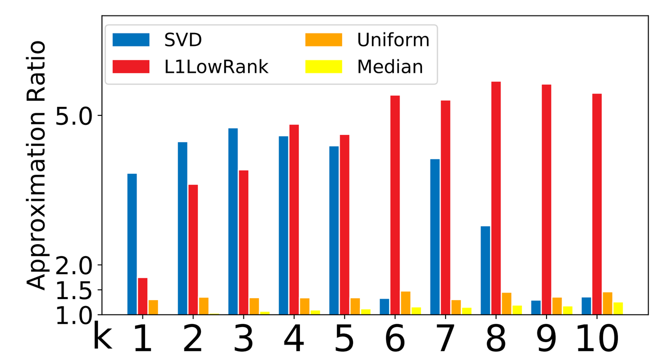
In the spirit of averaging, we found that taking a median works a bit better in practice. Inspired by our theoretical analysis, we propose a simple heuristic algorithm which can output a rank-k solution.

Algorithm 1 Median Heuristic

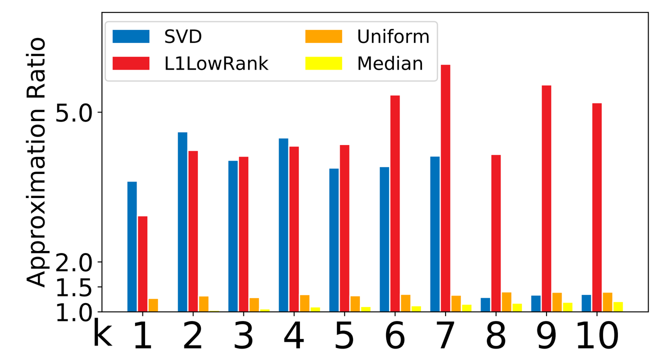
1: Sample a set $I = \{i_1, i_2, \cdots, i_{sk}\}$ from $\binom{[n]}{sk}$ uniformly at random.
2: Compute $B \in \mathbb{R}^{n\times k}$ s.t., for $t \in [n], q \in [k]$, $B_{t,q} = \text{median}(A_{t,i_{s(q-1)+1}}, \cdots, A_{t,i_{sq}})$.
3: Solve $\min_{X \in \mathbb{R}^{k\times d}} \|BX - A\|_1$ and let the solution be X^*. Output BX^*.

Experiments

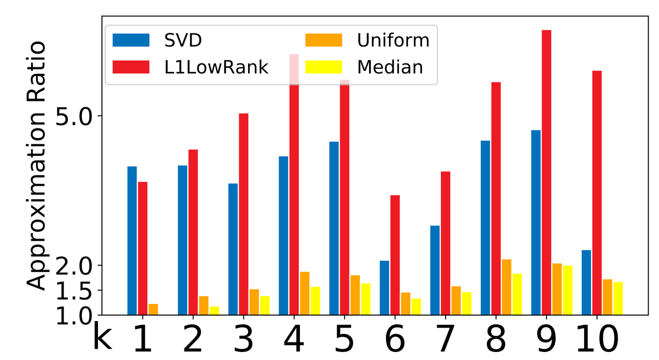
- Synthetic data $+$ 1.1-stable distribution:



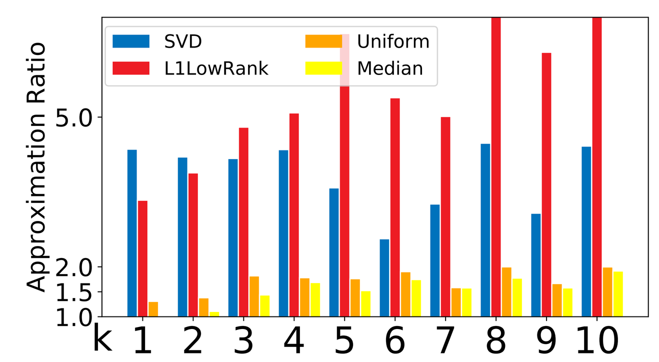
- Synthetic data $+$ 1.1-th root of a Cauchy distribution:



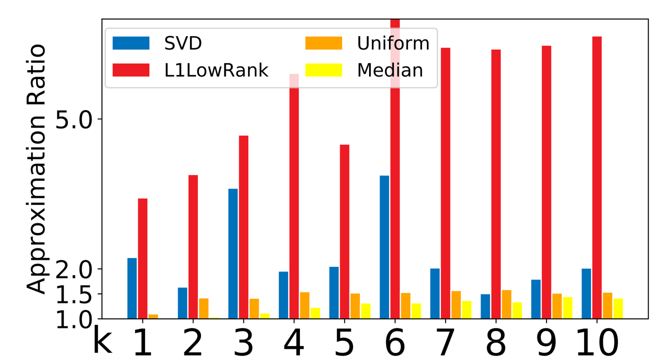
- isolet $+$ 1.1-stable distribution:



- isolet $+$ 1.1-th root of a Cauchy distribution:



- mfeat $+$ 1.1-stable distribution:



- mfeat $+$ 1.1-th root of a Cauchy distribution: