Scalable and Robust PCA Approach with Random Column/Row Sampling

Mostafa Rahmani and George Atia

University of Central Florida
Outline

- Robust PCA problem & data corruption models
- Randomized approaches & existing results
- Proposed approach
- New result
- Numerical experiments
Applications

- Background Subtraction
- Removing Shadow
- Network Data Analysis
- Phased array systems signal processing
Robust PCA Problem

Data Model

- The problem is defined to
  - Learn the column-space/row-space of \( L \)
  - Decomposing matrix \( D \)

\[ D = L + C \]

- Low Rank Matrix
- Data Corruption
Data Corruption Models

Matrix C

- **Element-wise corruption model**
  - Matrix $C$ is a sparse matrix with arbitrary support.
  - All the columns/rows might be affected.
  - Known as low rank plus sparse matrix decomposition problem.

- **Column-wise corruption model**
  - A subset of the columns of $C$ are non-zero columns.
  - These non-zero columns do not lie in the Column space of $L$.
  - Known as subspace recovery or outlier detection problem.
Algorithms

- **Element-wise model**
  - Principal Component Pursuit [Chandrasekaran et al. 2011]
  - Alternating minimization [Ke et al. 2005]

- **Column-wise model**
  - Algorithms based on column-sparsity [Xu et al. 2010, Ding et al. 2006]
  - Algorithms based on outliers linear independence [Soltanolkotabi et al, 2012]
  - Algorithm based on low coherency of outliers [Rahmani et. Al, 2016]
Complexity of Robust PCA

- Computation complexity
  \[ \geq O(r N_1 N_2 T) \]

- Memory requirement
  \[ O(N_1 N_2) \]

Can we solve the problem with few random linear measurements?
Randomized approach

Element wise model
(Matrix Decomposition)
[Mackey et al. 2011, Rahmani et al. 2015]

Column-space learning
\[ D = L + C \]
\[ D_s \in \mathbb{R}^{N_1 \times m} \]
\[ D_r \in \mathbb{R}^{n \times N_2} \]

Row-space learning
\[ D = L + C \]

Column wise model
(Subspace Recovery)
[Li et al., 2014]

Matrix Embedding
\[ D_s^\phi \in \mathbb{R}^{n \times m} \]

Outlier Removal

\[ m: \text{Number of randomly sampled columns} \]
\[ n: \text{Number of randomly sampled rows} \]

Basis for the column space

Low Rank Matrix Recovery
Existing Results

- Elements-wise model (matrix decomposition)
  - Sample complexity $O(r\mu \max(N_1, N_2))$ [Rahmani et al., 2015]
  - Computation complexity $O(r^2 \mu \max(N_1, N_2)T)$

- Column-wise model (Subspace recovery)
  - Sample complexity $O(rN_2)$ [Li et al., 2015], $O(r^2 \mu)$ [Rahmani et al., 2015]
  - Computation complexity $O(N_1r^2\mu + r^3\mu)$
Motivation

In the column-wise outlier model, can we make the computation complexity of subspace recovery independent from the size of data?
Proposed Randomized Design

Column-wise model
(Subspace Recovery)

Row Sampling

Column Sampling

D = L + C

$D_s \in \mathbb{R}_{N_1 \times m}$

Outlier Removal

D = L + C

$D_s^\phi \in \mathbb{R}^{n \times m}$

Complexity can be independent from size data

Basis for the column space
Outlier Removal

- Outlier column sparsity

\[
\min_{\hat{L}, \hat{C}} \quad \|\hat{L}\|_* + \lambda \|\hat{C}\|_{1,2} \\
\text{subject to} \quad \hat{L} + \hat{C} = D
\]

- Outlier linear independence

Checking if a column is linearly dependent on other columns or has sparse representation w.r.t. them
**Data Model**: The given data matrix $\mathbf{D} \in \mathbb{R}^{N_1 \times N_2}$ satisfies the following conditions

1. $\mathbf{D} = \mathbf{L} + \mathbf{C}$ and the columns of $\mathbf{D}$ are normalized.
2. $\text{Rank}(\mathbf{L}) = r$.
3. Matrix $\mathbf{C}$ has $K$ non-zero columns. The non-zero columns of $\mathbf{C}$ are i.i.d. random vectors uniformly distributed on the unit sphere. \[ \frac{K}{N_2} = \frac{\# \text{ outliers}}{\# \text{ Columns}} \]
Performance Guarantee

Sufficient Conditions, Outlier detection: Column sparsity

**Theorem 1:** If the given data follows the data model, the columns/rows are sampled randomly, and

\[
\frac{K}{N_2} \leq \frac{N_2'/2N_2'}{1 + 6r_\mu_v(121/9)}
\]

\[
m \geq \max \left( 12 \frac{K}{N_2} (1 + 6r_\mu_v(121/9))^{2\log \frac{2}{\delta}}, 10r_\mu_v \log \frac{2r}{\delta} \right)
\]

\[
n \geq \max \left[ r_\mu_u \max \left( c_1 \log r, c_2 \log \left( \frac{3}{\delta} \right) \right), r + 1 + 2\log \frac{2K}{\delta} + \sqrt{8 \log \frac{2K}{\delta}} \right]
\]

then the proposed method recovers the exact subspace with probability at least \(1 - 4\delta\).
Theorem 2: If data follows the data model, the columns/rows are sampled randomly, and

\[
m_1 \geq C \mu_r r \log \frac{4r}{\delta}
\]

\[
m_2 \geq \max \left[ r \mu_u \max \left( c_1 \log r, c_2 \log \left( \frac{3}{\delta} \right) \right), r + q + 2 \log \frac{2}{\delta} + \sqrt{8 q \log \frac{K}{\delta}} \right]
\]

then the proposed method recovers the exact subspace with probability at least \(1 - 6\delta\).
Proposed Randomized Design

Column-wise model
(Subspace Recovery)

Row Sampling

Outlier Removal

Basis for the column space
New Result

- The computation and sample complexity for exact subspace recovery is almost independent from the size of data.
  - Sample complexity
    - Column sparsity: $O(r^2\mu_u \max(\mu_v, r\mu_v^2K/N_2))$
    - Linear independence: $O(r^2\mu_v \max(\mu_u, r\mu_vK/N_2))$
  - Computation complexity:
    - Column sparse: $O(rmnT)$
    - Linear independence: $O(rm^2n)$

Both $m$ and $n$ were shown to be independent from data size.
Numerical Experiment-Phase transition

with different values for the rank of $L$

$D \in \mathbb{R}^{2000 \times 4000}$

$r = 20, \rho = 0.2$

$r = 45, \rho = 0.2$

$r = 70, \rho = 0.2$
Numerical Experiment-Phase transition

with different data dimensions

\[ N_1 = \frac{N_2}{2} = 2000 \]

\[ N_1 = \frac{N_2}{2} = 20000 \]

\[ N_1 = \frac{N_2}{2} = 50000 \]
Numerical Experiment-Phase transition

with different $\rho = \frac{K}{N_2}$

$D \in \mathbb{R}^{2000 \times 4000}$

$r = 5, \; \rho = 0.2$

$r = 5, \; \rho = 0.5$

$r = 5, \; \rho = 0.7$
Phase Transition with Real Data

\[
D \in \mathbb{R}^{62 \times 512} \\
\approx 3
\]
Row Sampling vs Random Embedding

Preserving the low rank component

The component of outliers which does not lie in the span of inliers
Thank you.

Questions?!
New Result

- The computation and sample complexity for exact subspace recovery is almost independent from the size of data.

- Sample complexity
  - Column sparse: $O(r^2 \mu_u \max(\mu_v, r\mu_v^2 K/N_2))$
  - Linear independence: $O(r^2 \mu_v \max(\mu_u, r\mu_v K/N_2))$

- Computation complexity:
  - Column sparse: $O(rmnT)$
  - Linear independence: $O(rm^2n)$

Both $m$ and $n$ were shown to be independent from data size.