Lecture 8: Robust PCA

Yuejie Chi
The Ohio State University
Main Reference


Outline

• Motivating applications

• Mathematical formulation
Sparse+ Low-rank Matrix Decomposition

Suppose we are given a matrix of data observations:

\[ M = L + S, \]

where \( L \) is low-rank and \( S \) is sparse. We do not know the rank of \( L \) nor the sparsity level of \( S \).

**Question:** Can we recover both \( L \) and \( S \) from \( M \)? What if we only partially observe \( M \)?

This problem has many applications in data-intensive problems.
Consider $p$ data samples $X = [x_1, x_2, \ldots, x_p]$ that are centered, $x_i \in \mathbb{R}^n$. PCA seeks the direction that explains most of the variance of the data. Mathematically, we seek the direction $a \in \mathbb{R}^n$ (principal component) that maximizes

$$a = \arg\max_{\|a\|_2 = 1} \|XX^Ta\|_2 = \arg\min_{\|a\|_2 = 1} \min_b \|X - ab^T\|_F^2$$

corresponding to seek the rank-one matrix approximation of $X$. 

![Image of data points with a principal component arrow]
Principal Component Analysis

In general, PCA is useful because the first few principal components (PCs) explains most of the variance of the data. This amounts to finding the low-rank approximation of $X$, i.e.

$$\min_{\text{rank}(L)=r} \|X - L\|_F^2$$

where $r$ is the number of PCs.

Many applications of PCA:

- feature extraction;
- dimensionality reduction;

PCA justifies the approximate low-rank assumption on $X$. 

Page 6
Corruptions

What if the data samples $X = [x_1, x_2, \ldots, x_p]$ are corrupted?

- Outliers/Gross errors due to sensor errors/attacks/etc: each entry in $x_i$ corresponds to a sensor,
  \[ y_i = x_i + s_i \]
  where $s_i$ is a sparse vector with the nonzero entries corresponds to outliers.

The corrupted data can be written as

\[ Y = X + S \]

- The nominal PCA fails even with a few outliers:
  \[ \min_{\text{rank}(L)=r} \| Y - L \|^2_F \]
The nominal PCA could fail even with one outlier:
Video surveillance

Separation of background (low-rank) and foreground (sparse) in video:

\[ M = L + S \]
Consider a collection of random variables that are jointly Gaussian $x \sim \mathcal{N}(0, \Sigma)$:

$$p(x) \propto \frac{1}{|\Sigma|} \exp \left\{ -x^T \Sigma^{-1} x \right\} := |P| \exp \left\{ -x^T P x \right\}$$

where $P = \Sigma^{-1}$ is the precision matrix.

- The nonzero entries of $P$ describes the conditional independence between the variables, which can be depicted in a graphical model.

- **Graphical model learning**: Given i.i.d. samples of $x_i \sim \mathcal{N}(0, \Sigma)$, we want to learn the support of $P$.

- An interesting case is when $P$ is sparse, corresponding to the case that most of the pairs of random variables are conditionally independent.
Graphical modeling with latent factors

What if we only observe a subset of the variables?

- denote $x_o$ as the observed variables;
- denote $x_h$ as the hidden variables (latent factors);

The precision matrix of all data can be written as

$$\Sigma^{-1} = \begin{bmatrix} P_o & P_{o,h} \\ P_{h,o} & P_h \end{bmatrix}$$

We only observe the marginal precision matrix on the observed variables $x_0$:

$$\Sigma_o^{-1} = P_o - P_{o,h} P_h^{-1} P_{h,o}$$

- $P_o$ is sparse due to conditional independence;
- $P_{o,h} P_h^{-1} P_{h,o}$ is low-rank if the number of hidden variables is small;
Structure from motion

In the pipeline of performing SFM, assume we’ve found a set of good feature points with their corresponding 2D locations in the images.

Tomasi and Kanade's factorization: Given $n$ points $x_{i,j}^T \in \mathbb{R}^2$ corresponding to the location of the $i$th point in the $j$th frame, define the matrix

$$M = \begin{bmatrix} x_{1,1} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,m} \end{bmatrix} \in \mathbb{R}^{n \times 2m}, \quad \text{and} \quad \text{rank}(M) = 3$$

- Occlusions: missing entries in $M$;
- Wrong feature point/correspondence: sparse corruptions in $M$;
Identifiability issues: a matrix can be simultaneously low-rank and sparse!

\[
\begin{bmatrix}
1 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 0 \\
\end{bmatrix}
\]

\[\text{vs}\]

\[
\begin{bmatrix}
1 & 0 & 1 & \cdots & 1 \\
0 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 1 \\
\end{bmatrix}
\]

Would the sparse component to be spread.

we assume its support is uniformly at random.

\[
\begin{bmatrix}
1 & 1 & 1 & \cdots & 1 \\
1 & 1 & 1 & \cdots & 1 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 1 & 1 & \cdots & 1 \\
\end{bmatrix}
\]

\[\text{vs}\]

\[
\begin{bmatrix}
1 & 1 & 1 & \cdots & 1 \\
0 & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 0 \\
\end{bmatrix}
\]

Would the low-rank component to be incoherent.
Let $M$ be a rank-$r$ matrix with the SVD $M = U \Sigma V^T$, where $U, V \in \mathbb{R}^{n \times r}$.

**Definition 1. [Coherence]** Smallest scalar $\mu_1$ obeying

\[ \max_{1 \leq i \leq n} \|U^T e_i\|_2^2 \leq \mu_1 \frac{r}{n}, \quad \max_{1 \leq i \leq n} \|V^T e_i\|_2^2 \leq \mu_1 \frac{r}{n}, \]

where $e_i$ is the $i$th standard basis vector.

- Geometric condition: $U = \text{colspan}(M)$
- Since $\sum_{i=1}^{n} \|U^T e_i\|_2^2 = r$, $\mu_1 \geq 1$.
- If $e_i \in U$, $\mu_1 = n/r$;
- If $\frac{1}{\sqrt{n}} \mathbf{1} = U$, $\mu_1 = 1$.

We would like $\mu_1 = O(1)$. 

---

**Low-rank component: Coherence**

---
Definition 2. [Joint Coherence] Smallest scalar $\mu_2$ obeying

$$\|UV^T\|_{\infty} \leq \sqrt{\frac{\mu_2 r}{n^2}}$$

This avoids $UV^T$ to be too peaked.

- $\mu_1 \leq \mu_2 \leq \mu_1^2 r$, since

$$|(UV^T)_{ij}| = |u_i^T v_j| \leq \frac{\mu_1 r}{n}$$

$$\|UV^T\|_{\infty} \geq \frac{1}{n} \sum_i (UV^T)_{ij}^2 = \frac{1}{n} \left\| V^T e_j \right\|_2^2$$

- The incoherence parameter $\mu_1$ is sufficient and necessary for MC, while $\mu_2$ is necessary for Robust PCA (connection to the planted clique problem [c.f. Chen, 2015]).
Algorithm

Non-convex heuristic:

\[
(\hat{L}, \hat{S}) = \arg\min_{L,S} \text{rank}(L) + \lambda \|S\|_0, \quad \text{s.t.} \quad M = L + S.
\]

Convex relaxation: Principal Component Pursuit (PCP)

\[
(\hat{L}, \hat{S}) = \arg\min_{L,S} \|L\|_* + \lambda \|S\|_1, \quad \text{s.t.} \quad M = L + S
\]

where \(\| \cdot \|_*\) is the nuclear norm, and \(\| \cdot \|_1\) is the entry-wise \(\ell_1\) norm.

- \(\lambda > 0\) is some regularization parameter that balances the two terms.

- The algorithm is convex.
Performance Guarantee

Theorem

- $L_0$ is $n \times n$ of rank($L_0$) $\leq \rho_r n \mu^{-1} (\log n)^{-2}$
- $S_0$ is $n \times n$, random sparsity pattern of cardinality $m \leq \rho_s n^2$

Then with probability $1 - O(n^{-10})$, PCP with $\lambda = 1/\sqrt{n}$ is exact:

$$\hat{L} = L_0, \quad \hat{S} = S_0$$

Same conclusion for rectangular matrices with $\lambda = 1/\sqrt{\max \text{dim}}$

Remark:

- No tuning parameters: $\lambda = 1/\sqrt{n}$ is prefixed by the theorem.
- Essentially optimal: rank($L$) = $O(n)$, $\|S\|_0 = O(n^2)$
- Arbitrary magnitudes and sign patterns of $L$ and $S$!!
Figure 1: Correct recovery for varying rank and sparsity. Fraction of correct recoveries across 10 trials, as a function of rank($L_0$) (x-axis) and sparsity of $S_0$ (y-axis). Here, $n_1 = n_2 = 400$. In all cases, $L_0 = XY^*$ is a product of independent $n \times r$ i.i.d. $\mathcal{N}(0, 1/n)$ matrices. Trials are considered successful if $\|\hat{L} - L_0\|_F / \|L_0\|_F < 10^{-3}$. Left: low-rank and sparse decomposition, $\text{sgn}(S_0)$ random. Middle: low-rank and sparse decomposition, $S_0 = \mathcal{P}_\Omega \text{sgn}(L_0)$. Right: matrix completion. For matrix completion, $\rho_s$ is the probability that an entry is omitted from the observation.
Connections with Matrix Completion

Comparison with Matrix Completion:

- In MC we know where the entries are missing; while in RPCA we do not know the locations of corruptions.
What if we have both missing data and corruptions?

- Consider we only have partial observations of a low-rank matrix $L$ on the index set $\Omega$, and the observed matrix $M$ satisfies

  $$M_{ij} = L_{ij} + S_{ij}, \quad (i, j) \in \Omega$$

  where $S = (S_{ij})$ is a sparse matrix supported on $\Omega$.

- A natural extension of RPCA:

  $$(\hat{L}, \hat{S}) = \arg \min_{L, S} \|L\|_* + \lambda \|S\|_1, \quad \text{s.t.} \quad M = \mathcal{P}_\Omega (L + S)$$
MC with Corruptions: Guarantee

**Theorem**

- $L_0$ is $n \times n$ as before, $\text{rank}(L_0) \leq \rho_r n \mu^{-1} (\log n)^{-2}$
- $\Omega_{\text{obs}}$ random set of size $m = 0.1n^2$
- each observed entry is corrupted with probability $\tau \leq \tau_s$

Then with probability $1 - O(n^{-10})$, PCP with $\lambda = 1/\sqrt{0.1n}$ is exact:

$$\hat{L} = L_0$$

Same conclusion for rectangular matrices with $\lambda = 1/\sqrt{0.1 \text{max dim}}$

---

\(^a\)missing fraction is arbitrary

- No tuning parameters: $\lambda = 1/\sqrt{n}$ is prefixed by the theorem.

- Essentially optimal: $\text{rank}(L) = O(n)$, $\|S\|_0 = O(m)$

- Arbitrary magnitudes and sign patterns of $L$ and $S$!
[Otazo et.al. 2014]: “The combination of compressed sensing and low-rank matrix completion represents an attractive proposition for further increases in imaging speed...”

L+S decomposition of fully-sampled 2D cardiac cine data corresponding to the central x location. The low-rank component captures the correlated background among temporal frames and the sparse component S the remaining dynamic information (heart motion).
Application in Accelerated MRI

L+S decomposition improves the performance of CS in accelerated MRI significantly with lower residual aliasing artifacts.

Constructing better priors on the signals helps performance!