ECE 18-898G: Special Topics in Signal Processing: Sparsity, Structure, and Inference

Robust PCA

Yuejie Chi

Department of Electrical and Computer Engineering

Carnegie Mellon University

Spring 2018

Demixing sparse and low-rank matrices

Suppose we are given a matrix

$$\boldsymbol{M} = \underbrace{\boldsymbol{L}}_{\text{low-rank}} + \underbrace{\boldsymbol{S}}_{\text{sparse}} \in \mathbb{R}^{n \times n}$$

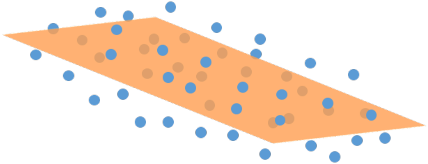
Question: Can we hope to recover both $\boldsymbol{L}$ and $\boldsymbol{S}$ from $\boldsymbol{M}$?

Principal component analysis (PCA)

- N samples $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N] \in \mathbb{R}^{n \times N}$ that are centered
- PCA: seeks r directions that explain most variance of data

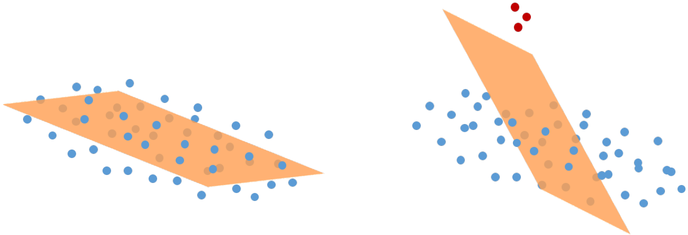
$$\text{minimize}_{\boldsymbol{L}:\text{rank}(\boldsymbol{L})=r} \quad \|\boldsymbol{X} - \boldsymbol{L}\|_{\mathrm{F}}$$

  - best rank-r approximation of $\boldsymbol{X}$



Sensitivity to corruptions / outliers

What if some samples are corrupted (e.g. due to sensor errors / attacks)?



Classical PCA fails even with a few outliers

Video surveillance

Separation of background (low-rank) and foreground (sparse)

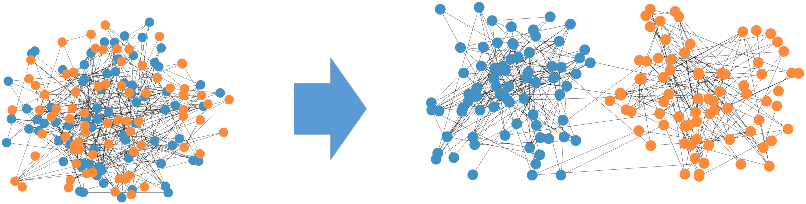
Candes, Li, Ma, Wright '11

Graph clustering / community recovery

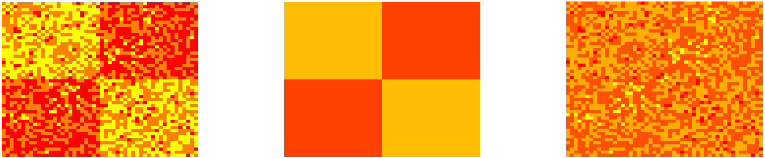
- n nodes, 2 (or more) clusters
- A friendship graph $\mathcal{G}$: for any pair (i, j),

$$M_{i,j} = \begin{cases} 1, & \text{if } (i,j) \in \mathcal{G} \\ 0, & \text{else} \end{cases}$$

- Edge density within clusters $>$ edge density across clusters
- **Goal:** recover cluster structure



Graph clustering / community recovery



$$\boldsymbol{M} = \underbrace{\boldsymbol{L}}_{\text{low-rank}} + \underbrace{\boldsymbol{M}-\boldsymbol{L}}_{\text{sparse}}$$

- An equivalent goal: recover ground truth matrix

$$L_{i,j} = \begin{cases} 1, & \text{if } i \text{ and } j \text{ are in same community} \\ 0, & \text{else} \end{cases}$$

- Clustering $\Longleftrightarrow$ robust PCA

When is decomposition possible?

Identifiability issues: a matrix might be simultaneously low-rank and sparse!

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}}_{\text{sparse and low-rank}} \quad \text{vs.} \quad \underbrace{\begin{bmatrix} 1 & 0 & 1 & \cdots & 1 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix}}_{\text{sparse but not low-rank}}$$

Nonzero entries of sparse component need to be spread out
— assume locations of nonzero entries are random / restrict the number of nonzeros per row/column

When is decomposition possible?

Identifiability issues: a matrix might be simultaneously low-rank and sparse!

$$\underbrace{\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix}}_{\text{low-rank and dense}} \quad \text{vs.} \quad \underbrace{\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}}_{\text{low-rank but sparse}}$$

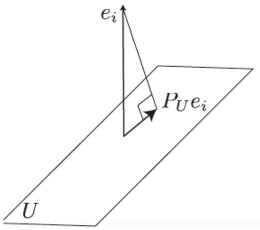
Low-rank component needs to be incoherent.

Low-rank component: coherence

Definition 8.1

Coherence parameter μ_1 of $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ is smallest quantity s.t.

$$\max_i \|\boldsymbol{U}^\top \boldsymbol{e}_i\|^2 \leq \frac{\mu_1 r}{n} \quad \text{and} \quad \max_i \|\boldsymbol{V}^\top \boldsymbol{e}_i\|^2 \leq \frac{\mu_1 r}{n}$$



Low-rank component: joint coherence

Definition 8.2 (Joint coherence)

Joint coherence parameter μ_2 of $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ is smallest quantity s.t.

$$\|\boldsymbol{U}\boldsymbol{V}^\top\|_\infty \leq \sqrt{\frac{\mu_2 r}{n^2}}$$

This prevents $\boldsymbol{U}\boldsymbol{V}^\top$ from being too peaky.

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

$$|(\boldsymbol{U}\boldsymbol{V}^\top)_{ij}| = |\boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{V}^\top \boldsymbol{e}_j| \leq \|\boldsymbol{e}_i^\top \boldsymbol{U}\| \cdot \|\boldsymbol{V}^\top \boldsymbol{e}_j\| \leq \frac{\mu_1 r}{n}$$

$$\|\boldsymbol{U}\boldsymbol{V}^\top\|_\infty^2 \geq \frac{\|\boldsymbol{U}\boldsymbol{V}^\top \boldsymbol{e}_j\|_{\mathrm{F}}^2}{n} = \frac{\|\boldsymbol{V}^\top \boldsymbol{e}_j\|^2}{n} = \frac{\mu_1 r}{n^2} \ \text{(suppose } \|\boldsymbol{V}^\top \boldsymbol{e}_j\|^2 = \frac{\mu_1 r}{n}\text{)}$$

Convex relaxation

$$\text{minimize}_{\boldsymbol{L},\boldsymbol{S}} \quad \text{rank}(\boldsymbol{L}) + \lambda\|\boldsymbol{S}\|_0, \quad \text{s.t.} \quad \boldsymbol{M} = \boldsymbol{L} + \boldsymbol{S} \tag{8.1}$$

$$\Downarrow$$

$$\text{minimize}_{\boldsymbol{L},\boldsymbol{S}} \quad \|\boldsymbol{L}\|_* + \lambda\|\boldsymbol{S}\|_1, \quad \text{s.t.} \quad \boldsymbol{M} = \boldsymbol{L} + \boldsymbol{S} \tag{8.2}$$

- $\|\cdot\|_*$ is nuclear norm; $\|\cdot\|_1$ is entry-wise ℓ_1 norm
- $\lambda > 0$: regularization parameter that balances two terms

Theoretical guarantee

Theorem 8.3 (Candes, Li, Ma, Wright '11)

- *rank*$(\boldsymbol{L}) \lesssim \frac{n}{\max\{\mu_1,\mu_2\}\log^2 n}$;
- *Nonzero entries of* $\boldsymbol{S}$ *are randomly located, and* $\|\boldsymbol{S}\|_0 \le \rho_s n^2$ *for some constant* $\rho_s > 0$ *(e.g.* $\rho_s = 0.2$*).*

Then (8.2) with $\lambda = 1/\sqrt{n}$ *is exact with high prob.*

- rank($\boldsymbol{L}$) can be quite high (up to $n/\text{polylog}(n)$)
- Parameter free: $\lambda = 1/\sqrt{n}$
- Ability to correct gross error: $\|\boldsymbol{S}\|_0 \asymp n^2$
- Sparse component $\boldsymbol{S}$ can have arbitrary magnitudes / signs!

Geometry

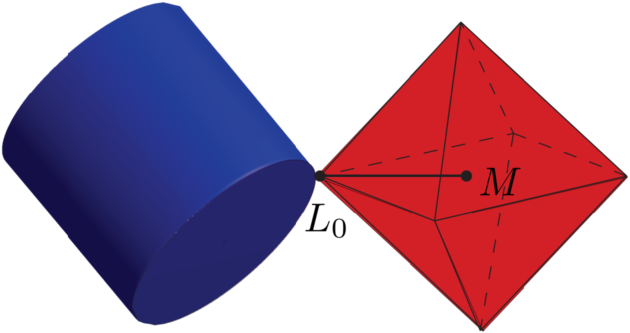
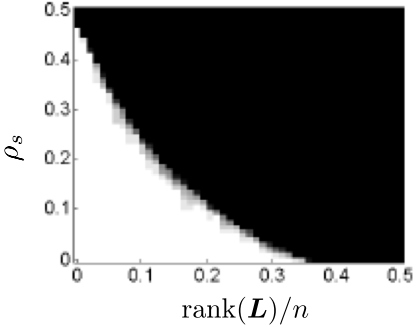


Fig. credit: Candes '14

Empirical success rate



$n = 400$

Fig. credit: Candes, Li, Ma, Wright '11

Dense error correction

Theorem 8.4 (Ganesh et al. '10, Chen et al. '13)

- *rank*$(\boldsymbol{L}) \lesssim \frac{n}{\max\{\mu_1,\mu_2\}\log^2 n}$;
- *Nonzero entries of* $\boldsymbol{S}$ *are randomly located, have random sign, and* $\|\boldsymbol{S}\|_0 = \rho_s n^2$.

Then (8.2) with $\lambda \asymp \sqrt{\frac{1-\rho_s}{\rho_s n}}$ *succeeds with high prob., provided that*

$$\underbrace{1-\rho_s}_{\text{non-corruption rate}} \gtrsim \sqrt{\frac{\max\{\mu_1,\mu_2\} r \,\mathrm{polylog}(n)}{n}}$$

- When additive corruptions have random signs, (8.2) works even when a dominant fraction of entries are corrupted

Is joint coherence needed?

- Matrix completion: does not need μ_2
- Robust PCA: so far we need μ_2

Question: can we remove μ_2? can we recover $\boldsymbol{L}$ with rank up to $\frac{n}{\mu_1 \text{polylog}(n)}$ (rather than $\frac{n}{\max\{\mu_1,\mu_2\}\text{polylog}(n)}$) with a *constant* fraction of outliers?

Answer: no (example: planted clique)

Planted clique problem

Setup: a graph $\mathcal{G}$ of n nodes generated as follows

1. connect each pair of nodes independently with prob. 0.5
2. pick n_0 nodes and make them a clique (fully connected)

Goal: find hidden clique from $\mathcal{G}$

Information theoretically, one can recover a clique if $n_0 > 2\log_2 n$

Conjecture on computational barrier

Conjecture: $\forall$ constant $\epsilon > 0$, if $n_0 \leq n^{0.5-\epsilon}$, then no tractable algorithm can find the clique from $\mathcal{G}$ with prob. $1 - o(1)$

— often used as hardness assumption

Lemma 8.5

If there is an algorithm that allows recovery of any $\boldsymbol{L}$ from $\boldsymbol{M}$ with rank$(\boldsymbol{L}) \leq \frac{n}{\mu_1 \text{polylog}(n)}$, then the above conjecture is violated

Proof of Lemma 8.5

Suppose $\boldsymbol{L}$ is true adjacency matrix,

$$L_{i,j} = \begin{cases} 1, & \text{if } i, j \text{ are both in the clique} \\ 0, & \text{else} \end{cases}$$

Let $\boldsymbol{A}$ be adjacency matrix of $\mathcal{G}$, and generate $\boldsymbol{M}$ s.t.

$$M_{i,j} = \begin{cases} A_{i,j}, & \text{with prob. } 2/3 \\ 0, & \text{else} \end{cases}$$

Therefore, one can write

$$\boldsymbol{M} = \boldsymbol{L} + \underbrace{\boldsymbol{M} - \boldsymbol{L}}_{\text{each entry is nonzero w.p. } 1/3}$$

Proof of Lemma 8.5

Note that

$$\mu_1 = \frac{n}{n_0} \qquad \text{and} \qquad \mu_2 = \frac{n^2}{n_0^2}$$

If there is an algorithm that can recover any $\boldsymbol{L}$ of rank $\frac{n}{\mu_1 \mathsf{polylog}(n)}$ from $\boldsymbol{M}$, then

$$\mathsf{rank}(\boldsymbol{L}) = 1 \leq \frac{n}{\mu_1 \mathsf{polylog}(n)} \quad \Longleftrightarrow \quad n_0 \geq \mathsf{polylog}(n)$$

But this contradicts the conjecture (which claims computational infeasibility to recover $\boldsymbol{L}$ unless $n_0 \geq n^{0.5-o(1)}$)

Matrix completion with corruptions

What if we have missing data + corruptions?

- Observed entries

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

  for some observation set Ω, where $\boldsymbol{S} = (S_{ij})$ is sparse

- A natural extension of RPCA

$$\text{minimize}_{\boldsymbol{L},\boldsymbol{S}} \quad \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{S}\|_1 \quad \text{s.t. } \mathcal{P}_\Omega(\boldsymbol{M}) = \mathcal{P}_\Omega(\boldsymbol{L}+\boldsymbol{S})$$

- Theorems 8.3 - 8.4 easily extend to this setting

Efficient algorithm: proximal method

In the presence of noise, one needs to solve

$$\text{minimize}_{\boldsymbol{L},\boldsymbol{S}} \quad \|\boldsymbol{L}\|_* + \lambda\|\boldsymbol{S}\|_1 + \frac{\mu}{2}\|\boldsymbol{M}-\boldsymbol{L}-\boldsymbol{S}\|_{\mathrm{F}}^2$$

which can be solved efficiently via proximal method

Algorithm 8.1 Iterative soft-thresholding

for $t = 0, 1, \cdots$**:**

$$\begin{aligned} \boldsymbol{L}^{t+1} &= \mathcal{T}_{1/\mu}\left(\boldsymbol{M}-\boldsymbol{S}^t\right) \\ \boldsymbol{S}^{t+1} &= \psi_{\lambda/\mu}\left(\boldsymbol{M}-\boldsymbol{L}^{t+1}\right) \end{aligned}$$

where $\mathcal{T}$ is singular-value thresholding operator, and ψ is soft thresholding operator

Nonconvex approach

Alternatively, we can directly solve the nonconvex problem without relaxation with the assumptions

- $\mathrm{rank}(\boldsymbol{L}) \le r$; if we write the SVD of $\boldsymbol{L} = \boldsymbol{U\Sigma V}^\top$, set

$$\boldsymbol{X}^\star = \boldsymbol{U\Sigma}^{1/2}; \quad \boldsymbol{Y}^\star = \boldsymbol{V\Sigma}^{1/2}.$$

- the non-zero entries of $\boldsymbol{S}$ are "spread out" (no more than α fraction of non-zeros per row/column), but otherwise arbitrary.

$$\mathcal{S}_\alpha = \{\boldsymbol{S} \in \mathbb{R}^{n\times n} : \quad \|\boldsymbol{S}_{i,:}\|_0 \le \alpha n;\ \|\boldsymbol{S}_{:,j}\|_0 \le \alpha n\}$$

$$\text{minimize}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{S}\in\mathcal{S}_\alpha} \quad \underbrace{\|\boldsymbol{M} - \boldsymbol{XY}^\top - \boldsymbol{S}\|_{\mathrm{F}}^2}_{\text{quadratic loss}} + \underbrace{\frac{1}{4}\|\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{Y}^\top\boldsymbol{Y}\|_{\mathrm{F}}^2}_{\text{fix scaling ambiguity}}$$

where $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times r}$.

Gradient descent and hard thresholding

$$\text{minimize}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{S}\in\mathcal{S}_\alpha} \quad F(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{S})$$

where $F(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{S}) := \|\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{S}\|_{\mathrm{F}}^2 + \frac{1}{4}\|\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{Y}^\top\boldsymbol{Y}\|_{\mathrm{F}}^2$.

Algorithm 8.2 Gradient descent + Hard thresholding for RPCA

Input: $\boldsymbol{M}$, r, α, γ, η.

Spectral initialization: Set $\boldsymbol{S}^0 = \mathcal{H}_{\gamma\alpha}(\boldsymbol{M})$. Let $\boldsymbol{U}^0\boldsymbol{\Sigma}^0\boldsymbol{V}^{0\top}$ be the rank-r SVD of $\boldsymbol{M}^0 := \mathcal{P}_\Omega(\boldsymbol{M} - \boldsymbol{S})$; set $\boldsymbol{X}^0 = \boldsymbol{U}^0\left(\boldsymbol{\Sigma}^0\right)^{1/2}$ and $\boldsymbol{Y}^0 = \boldsymbol{V}^0\left(\boldsymbol{\Sigma}^0\right)^{1/2}$.

for $t = 0, 1, 2, \ldots, T-1$ **do**

1. **Hard thresholding**: $\boldsymbol{S}^{t+1} = \mathcal{H}_{\gamma\alpha}(\boldsymbol{M} - \boldsymbol{X}^t\boldsymbol{Y}^{t\top})$.
2. **Gradient updates**:

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta\nabla_{\boldsymbol{X}}F\left(\boldsymbol{X}^t, \boldsymbol{Y}^t, \boldsymbol{S}^{t+1}\right),$$

$$\boldsymbol{Y}^{t+1} = \boldsymbol{Y}^t - \eta\nabla_{\boldsymbol{Y}}F\left(\boldsymbol{X}^t, \boldsymbol{Y}^t, \boldsymbol{S}^{t+1}\right).$$

Efficient nonconvex recovery

Theorem 8.6 (Yi et al. '16)

Set $\gamma = 2$ *and* $\eta = 1/(36\sigma_{\max})$*. Suppose that*

$$\alpha \lesssim \min\left\{ \frac{1}{\mu_1\sqrt{\kappa r^3}}, \frac{1}{\mu_1 \kappa^2 r} \right\}.$$

The nonconvex approach (GD+HT) satisfies

$$\left\| \boldsymbol{X}^t \boldsymbol{Y}^{t\top} - \boldsymbol{L} \right\|_{\mathrm{F}}^2 \lesssim \left(1 - \frac{1}{288\kappa}\right)^t \mu_1^2 \kappa r^3 \alpha^2 \sigma_{\max}$$

- $O(\kappa \log 1/\epsilon)$ iterations to reach ϵ-accuracy.
- For adversarial outliers, the optimal fraction of $\alpha = O(1/\mu_1 r)$; the bound is worse by a factor of $\sqrt{r}$.
- extendable to partial observation case.

Reference

[1] "*Robust principal component analysis?*," E. Candes, X. Li, Y. Ma, and J. Wright, *Journal of ACM*, 2011.

[2] "*Rank-sparsity incoherence for matrix decomposition*," V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky, *SIAM Journal on Optimization*, 2011.

[3] "*Incoherence-optimal matrix completion*," Y. Chen, *IEEE Transactions on Information Theory*, 2015.

[4] "*Dense error correction for low-rank matrices via principal component pursuit*," A. Ganesh, J. Wright, X. Li, E. Candes, Y. Ma, *ISIT*, 2010.

[5] "*Low-rank matrix recovery from errors and erasures*," Y. Chen, A. Jalali, S. Sanghavi, C. Caramanis, *IEEE Transactions on Information Theory*, 2013.

[6] "*Fast Algorithms for Robust PCA via Gradient Descent*," X. Yi, D. Park, Y. Chen, and C. Caramanis, *NIPS*, 2016.