Robust Principal Component Analysis

Yuxin Chen

Wharton Statistics & Data Science, Spring 2022

Disentangling sparse and low-rank matrices

Suppose we are given a matrix

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

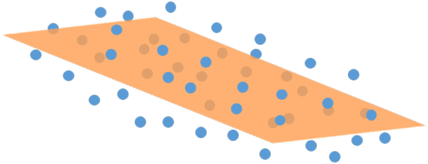
Question: can we hope to recover both $\boldsymbol{L}$ and $\boldsymbol{S}$ from $\boldsymbol{M}$?

Principal component analysis (PCA)

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

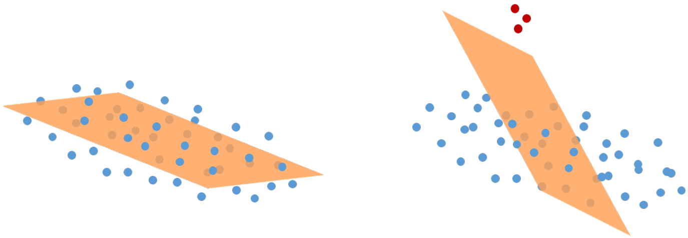
$$\text{minimize}_{\boldsymbol{L}:\text{rank}(\boldsymbol{L})=r} \quad \|\boldsymbol{X} - \boldsymbol{L}\|_{\mathrm{F}}$$

  - 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)

(a) Original frames (b) Low-rank $\hat{L}$ (c) Sparse $\hat{S}$

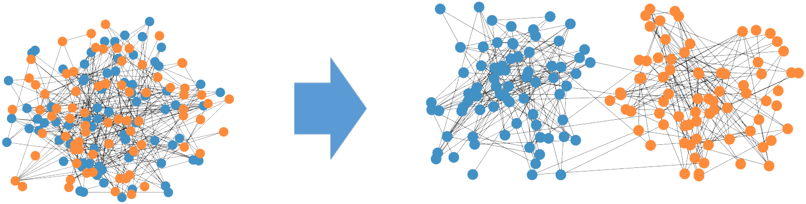
Candès, Li, Ma, Wright '11

Graph clustering / community recovery

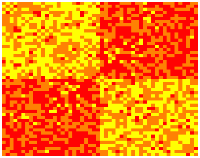
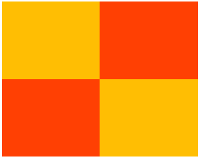
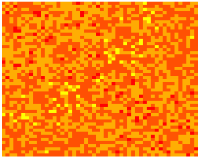
- n nodes, 2 (or more) clusters
- A friendship graph $\mathcal{G}$: for any pair (i,j),

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

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



Graph clustering / community recovery

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

- An equivalent goal: recover the ground truth matrix

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

- Clustering $\quad \Longleftrightarrow \quad$ robust PCA

Gaussian graphical models

Fact 14.1

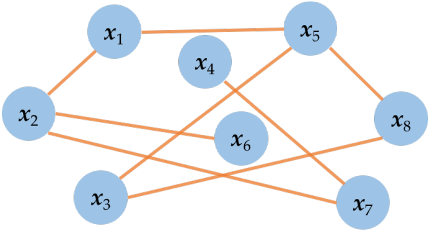
Consider a Gaussian vector $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$*. For any* u *and* v*,*

$$x_u \perp\!\!\!\perp x_v \;\Big|\; \boldsymbol{x}_{\mathcal{V}\backslash\{u,v\}}$$

iff $\Theta_{u,v} = 0$*, where* $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ *is the inverse covariance matrix*

conditional independence $\quad\Longleftrightarrow\quad$ sparsity

Gaussian graphical models



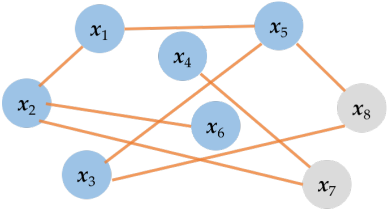
$$\underbrace{\begin{bmatrix} * & * & 0 & 0 & * & 0 & 0 & 0 \\ * & * & 0 & 0 & 0 & * & * & 0 \\ 0 & 0 & * & 0 & * & 0 & 0 & * \\ 0 & 0 & 0 & * & 0 & 0 & * & 0 \\ * & 0 & * & 0 & * & 0 & 0 & * \\ 0 & * & 0 & 0 & 0 & * & 0 & 0 \\ 0 & * & 0 & * & 0 & 0 & * & 0 \\ 0 & 0 & * & 0 & * & 0 & 0 & * \end{bmatrix}}_{\boldsymbol{\Theta}}$$

The inverse covariance matrix $\boldsymbol{\Theta}$ is often sparse

Graphical models with latent factors

What if one only observes a subset of variables?

$$\begin{bmatrix} \boldsymbol{x}_{\mathrm{o}} \\ \boldsymbol{x}_{\mathrm{h}} \end{bmatrix} \quad \begin{array}{c} \text{(observed variables)} \\ \text{(hidden variables)} \end{array}$$



$$\boldsymbol{x}_{\mathrm{o}} = [x_1, \cdots, x_6]^\top, \boldsymbol{x}_{\mathrm{h}} = [x_7, x_8]^\top$$

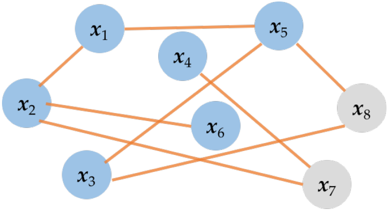
The covariance and precision matrices can be partitioned as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \overbrace{\boldsymbol{\Sigma}_{\mathrm{o}}}^{\text{observed part}} & \boldsymbol{\Sigma}_{\mathrm{o,h}} \\ \boldsymbol{\Sigma}_{\mathrm{o,h}}^\top & \boldsymbol{\Sigma}_{\mathrm{h}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Theta}_{\mathrm{o}} & \boldsymbol{\Theta}_{\mathrm{o,h}} \\ \boldsymbol{\Theta}_{\mathrm{o,h}}^\top & \boldsymbol{\Theta}_{\mathrm{h}} \end{bmatrix}^{-1}$$

Graphical models with latent factors

What if one only observes a subset of variables?

$$\begin{bmatrix} \boldsymbol{x}_{\rm o} \\ \boldsymbol{x}_{\rm h} \end{bmatrix} \quad \begin{matrix} \text{(observed variables)} \\ \text{(hidden variables)} \end{matrix}$$



$\boldsymbol{x}_{\rm o} = [x_1, \cdots, x_6]^\top, \boldsymbol{x}_{\rm h} = [x_7, x_8]^\top$

$$\underbrace{\boldsymbol{\Sigma}_{\rm o}^{-1}}_{\text{observed}} = \underbrace{\boldsymbol{\Theta}_{\rm o}}_{\text{sparse}} - \underbrace{\boldsymbol{\Theta}_{\rm o,h}\boldsymbol{\Theta}_{\rm h}^{-1}\boldsymbol{\Theta}_{\rm h,o}}_{\text{low-rank if \# latent vars is small}}$$

sparse + low-rank decomposition

When is decomposition possible?

Identifiability issue: 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

— This lecture: assume locations of the nonzero entries are random

When is decomposition possible?

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

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

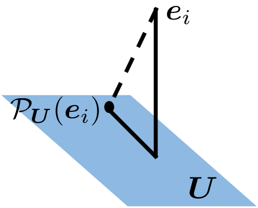
The low-rank component needs to be incoherent

Low-rank component: coherence

Definition 14.2

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

$$\max_i \|\boldsymbol{U}^\top \boldsymbol{e}_i\|_2^2 \leq \frac{\mu_1 r}{n} \quad \text{and} \quad \max_i \|\boldsymbol{V}^\top \boldsymbol{e}_i\|_2^2 \leq \frac{\mu_1 r}{n}$$



Low-rank component: joint coherence

Definition 14.3 (Joint coherence)

Joint coherence parameter μ_2 of $\boldsymbol{M} = \boldsymbol{U\Sigma V}^\top$ is the 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

$$|(\boldsymbol{U}\boldsymbol{V}^\top)_{ij}| = |\boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{V}^\top \boldsymbol{e}_j| \leq \|\boldsymbol{e}_i^\top \boldsymbol{U}\|_2 \cdot \|\boldsymbol{V}^\top \boldsymbol{e}_j\|_2 \leq \frac{\mu_1 r}{n}$$

$$\|\boldsymbol{U}\boldsymbol{V}^\top\|_\infty^2 \geq \frac{\|\boldsymbol{U}\boldsymbol{V}^\top \boldsymbol{e}_j\|_{\mathrm{F}}^2}{n} = \frac{\|\boldsymbol{V}^\top \boldsymbol{e}_j\|_2^2}{n} = \frac{\mu_1 r}{n^2} \quad (\text{suppose } \|\boldsymbol{V}^\top \boldsymbol{e}_j\|_2^2 = \frac{\mu_1 r}{n})$$

Convex relaxation

$$\text{minimize}_{\boldsymbol{L},\boldsymbol{S}} \quad \text{rank}(\boldsymbol{L}) + \lambda\|\boldsymbol{S}\|_0 \quad \text{s.t.} \quad \boldsymbol{M} = \boldsymbol{L} + \boldsymbol{S} \tag{14.1}$$

$$\Downarrow$$

$$\text{minimize}_{\boldsymbol{L},\boldsymbol{S}} \quad \|\boldsymbol{L}\|_* + \lambda\|\boldsymbol{S}\|_1 \quad \text{s.t.} \quad \boldsymbol{M} = \boldsymbol{L} + \boldsymbol{S} \tag{14.2}$$

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

Theoretical guarantee

Theorem 14.4 (Candès, 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 (14.2) 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

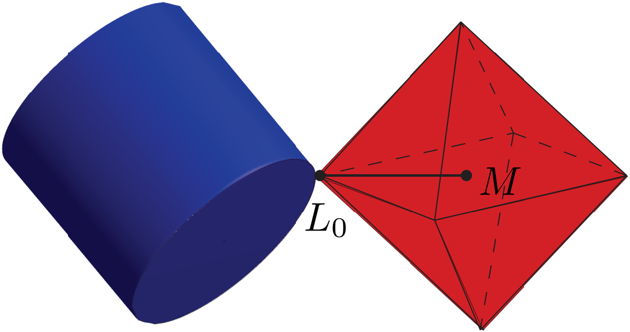
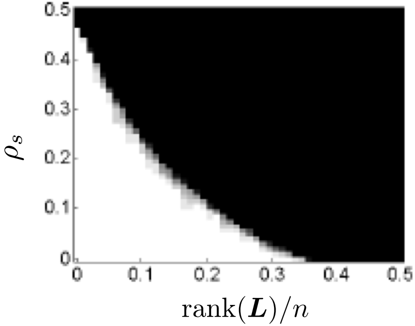


Fig. credit: Candès '14

Empirical success rate



$n = 400$

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

Dense error correction

Theorem 14.5 (Ganesh, Wright, Li, Candès, Ma '10, Chen, Jalali, Sanghavi, Caramanis '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 (14.2) with $\lambda \asymp \sqrt{\frac{1-\rho_s}{\rho_s n}}$ succeeds with high prob., provided that

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

- When additive corruptions have random signs, (14.2) works even when a dominant fraction of the entries are corrupted

Is joint coherence needed?

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

Question: is μ_2 needed? can we recover $\boldsymbol{L}$ with rank up to $\frac{n}{\mu_1 \mathsf{polylog}(n)}$ (rather than $\frac{n}{\max\{\mu_1,\mu_2\}\mathsf{polylog}(n)}$)?

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 the hidden clique from $\mathcal{G}$

Information theoretically, one can recover the 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 a hardness assumption

Lemma 14.6

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

Proof of Lemma 14.6

Suppose $\boldsymbol{L}$ is the 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 the 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 14.6

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 14.4 - 14.5 easily extend to this setting

Efficient algorithm

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

Algorithm 14.1 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 $\mathcal{T}$: singular-value thresholding operator; ψ: soft thresholding operator

Reference

- "*Lecture notes, Advanced topics in signal processing (ECE 8201)*," Y. Chi, 2015.
- "*Robust principal component analysis?*," E. Candes, X. Li, Y. Ma, and J. Wright, *Journal of ACM*, 2011.
- "*Rank-sparsity incoherence for matrix decomposition*," V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky, *SIAM Journal on Optimization*, 2011.
- "*Latent variable graphical model selection via convex optimization*," V. Chandrasekaran, P. Parrilo, and A. Willsky, *Annals of Statistics*, 2012.
- "*Incoherence-optimal matrix completion*," Y. Chen, *IEEE Transactions on Information Theory*, 2015.
- "*Dense error correction for low-rank matrices via principal component pursuit*," A. Ganesh, J. Wright, X. Li, E. Candes, Y. Ma, *ISIT*, 2010.

Reference

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