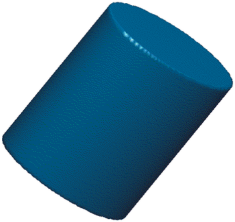
Computational Principles for High-dim Data Analysis

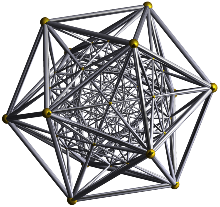
(Lecture Twelve)

Yi Ma

EECS Department, UC Berkeley

October 7, 2021





Decomposing Low-Rank and Sparse Matrices
(Principal Component Pursuit: Extensions)

"The whole is greater than the sum of the parts."
– Aristotle, *Metaphysics*

PCP and its Variants

Given $\boldsymbol{Y} = \boldsymbol{L}_o + \boldsymbol{S}_o$ with $\boldsymbol{L}_o$ low-rank and $\boldsymbol{S}_o$ sparse, PCP solves:

$$\text{minimize} \quad \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{S}\|_1 \quad \text{subject to} \quad \boldsymbol{L} + \boldsymbol{S} = \boldsymbol{Y}. \tag{1}$$

- λ can be adaptive to the density ρ_s of $\boldsymbol{S}_o$, for the range $0 \leq \rho_s < 1$.
- Signs of $\boldsymbol{S}_o$ can be deterministic, with guaranteed success up to density $\frac{1}{2}\rho_s$.
- If $\boldsymbol{Y} = \boldsymbol{L}_o + \boldsymbol{O}_o$ with $\boldsymbol{O}_o$ column sparse, we solve instead:

$$\min_{\boldsymbol{L},\boldsymbol{S}} \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{O}\|_{2,1} \quad \text{subject to} \quad \boldsymbol{L} + \boldsymbol{O} = \boldsymbol{Y}. \tag{2}$$

with $\|\boldsymbol{O}\|_{2,1} = \sum_i^{n_2} \|\boldsymbol{O}_i\|_2$. This is known as *sparse outlier pursuit*.[1]

[1] Robust PCA via outlier pursuit, Xu, Caramanis, and Sanghavi, *IEEE Transactions on Information Theory*, 2012.

Low-rank Matrix Recovery with Noise

Consider the measurement model with additive noise:

$$\boldsymbol{Y} = \boldsymbol{L}_o + \boldsymbol{S}_o + \boldsymbol{Z}_o, \tag{3}$$

where $\boldsymbol{Z}_o$ is a small error term $\|\boldsymbol{Z}_o\|_F \leq \epsilon$ for some $\epsilon > 0$.

Naturally, we solve a relaxed version to PCP (1):

$$\min_{\boldsymbol{L},\boldsymbol{S}} \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{S}\|_1 \quad \text{subject to} \quad \|\boldsymbol{Y} - \boldsymbol{L} - \boldsymbol{S}\|_F \leq \epsilon. \tag{4}$$

where we choose $\lambda = 1/\sqrt{n}$.

This combines classic PCA and robust PCA.

Stability of PCP

Theorem (Stability of PCP to Bounded Noise)

Under the same assumptions of PCP, that is, $\boldsymbol{L}_o$ *obeys the incoherence conditions and the support of* $\boldsymbol{S}_o$ *is uniformly distributed of size* m. *Then if* $\boldsymbol{L}_o$ *and* $\boldsymbol{S}_o$ *satisfy*

$$\operatorname{rank}(\boldsymbol{L}_o) \le \frac{\rho_r n}{\nu \log^2 n} \quad and \quad m \le \rho_s n^2, \tag{5}$$

with $\rho_r, \rho_s > 0$ *being sufficiently small numerical constants, with high probability in the support of* $\boldsymbol{S}_o$, *for any* $\boldsymbol{Z}_o$ *with* $\|\boldsymbol{Z}_o\|_F \le \epsilon$, *the solution* $(\hat{\boldsymbol{L}}, \hat{\boldsymbol{S}})$ *to the convex program* (4) *satisfies*

$$\|\hat{\boldsymbol{L}} - \boldsymbol{L}_o\|_F^2 + \|\hat{\boldsymbol{S}} - \boldsymbol{S}_o\|_F^2 \le C\epsilon^2, \tag{6}$$

where the constant $C = \left(16\sqrt{5}n + \sqrt{2}\right)^2$ *(which is not tight).*

Other Variants

If the magnitude of the low-rank component $\boldsymbol{L}_o$ is bounded, one could obtain better estimates by solving a *Lasso-type* program:

$$\min_{\boldsymbol{L},\boldsymbol{S}} \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{S}\|_1 + \tfrac{\mu}{2} \|\boldsymbol{L} + \boldsymbol{S} - \boldsymbol{Y}\|_F^2 \quad \text{subject to} \quad \|\boldsymbol{L}\|_\infty < \alpha. \tag{7}$$

The same analysis also applies to the stable version of the *outlier pursuit* program (2):

$$\min_{\boldsymbol{L},\boldsymbol{O}} \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{O}\|_{2,1} + \tfrac{\mu}{2} \|\boldsymbol{L} + \boldsymbol{O} - \boldsymbol{Y}\|_F^2 \quad \text{subject to} \quad \|\boldsymbol{L}\|_\infty < \alpha. \tag{8}$$

Both programs recover stable estimates for $\boldsymbol{L}$ and $\boldsymbol{S}$ with an error less than $C\epsilon^2$ where C does not depend on n.[2]

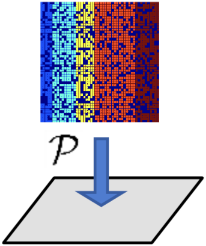
---

[2] Noisy matrix decomposition via convex relaxation: optimal rates in high dimensions, Agarwal, Negahban, and Wainwright. *The Annals of Statistics*, 2012.

Low-rank Matrix Recovery with Compressive Measurements

We are given only compressive linear measurements of a corrupted low-rank matrix:

$$\boldsymbol{Y} \doteq \mathcal{P}_{\mathsf{Q}}[\boldsymbol{L}_o + \boldsymbol{S}_o], \tag{9}$$



where $\mathcal{P}_{\mathsf{Q}}$ is a projection operator onto a subspace:

$$\mathsf{Q} \subseteq \mathbb{R}^{n_1 \times n_2}.$$

Consider the natural convex program

$$\min \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{S}\|_1 \quad \text{subject to} \quad \mathcal{P}_{\mathsf{Q}}[\boldsymbol{L} + \boldsymbol{S}] = \boldsymbol{Y}, \tag{10}$$

which is known as *compressive principal component pursuit* (CPCP).

Example: Transformed Low-rank Texture (Ch. 15)

An image of a low-rank texture from
an arbitrary view: $\boldsymbol{I} \circ \tau = \boldsymbol{L} + \boldsymbol{E}$.
To find out the correct deformation τ, solve:

$$\min_{\boldsymbol{L}, \boldsymbol{E}, \tau} \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{E}\|_1 \quad \text{subject to} \quad \boldsymbol{I} \circ \tau = \boldsymbol{L} + \boldsymbol{E}.$$

But this is nonlinear/nonconvex! Linearizing w.r.t. the deformation:

$$\boldsymbol{I} \circ \tau + \nabla \boldsymbol{I} \cdot d\tau \approx \boldsymbol{L} + \boldsymbol{E},$$

Let Q be the left kernel of the Jacobian $\nabla \boldsymbol{I}$: $\mathcal{P}_{\mathsf{Q}}[\nabla \boldsymbol{I}] = 0$, so we have:

$$\mathcal{P}_{\mathsf{Q}}[\boldsymbol{I} \circ \tau] = \mathcal{P}_{\mathsf{Q}}[\boldsymbol{L} + \boldsymbol{E}]. \tag{11}$$

Hence incrementally solve $d\tau$ via a convex program (CPCP):

$$\min_{\boldsymbol{L}, \boldsymbol{E}, d\tau} \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{E}\|_1 \quad \text{subject to} \quad \mathcal{P}_{\mathsf{Q}}[\boldsymbol{I} \circ \tau] = \mathcal{P}_{\mathsf{Q}}[\boldsymbol{L} + \boldsymbol{E}]. \tag{12}$$

Theoretical Guarantee for CPCP

Theorem (Compressive PCP)

Let $\boldsymbol{L}_o, \boldsymbol{S}_o \in \mathbb{R}^{n_1 \times n_2}$, *with* $n_1 \geq n_2$, *and suppose that* $\boldsymbol{L}_o \neq \boldsymbol{0}$ *is a rank-*r, ν*-incoherent matrix with* $r \leq \frac{c_r n_2}{\nu \log^2 n_1}$, *and* $\operatorname{sign}(\boldsymbol{S}_o)$ *is iid Bernoulli-Rademacher with nonzero probability* $\rho < c_\rho$. *Let* $\mathsf{Q} \subset \mathbb{R}^{n_1 \times n_2}$ *be a random subspace of dimension*

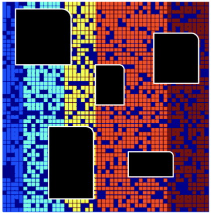
$$\mathit{dim}(\mathsf{Q}) \;\geq\; C_{\mathsf{Q}} \cdot (\rho n_1 n_2 + n_1 r) \cdot \log^2 n_1 \tag{13}$$

distributed according to the Haar measure, independent of $\operatorname{sign}(\boldsymbol{S}_o)$. *Then with probability at least* $1 - C n_1^{-9}$ *in* $(\operatorname{sign}(\boldsymbol{S}_o), \mathsf{Q})$, *the solution to*

$$\min \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{S}\|_1 \quad \text{s.t.} \quad \mathcal{P}_{\mathsf{Q}}[\boldsymbol{L} + \boldsymbol{S}] = \mathcal{P}_{\mathsf{Q}}[\boldsymbol{L}_o + \boldsymbol{S}_o] \tag{14}$$

with $\lambda = 1/\sqrt{n_1}$ *is unique, and equal to* $(\boldsymbol{L}_o, \boldsymbol{S}_o)$. *Above,* $c_r, c_\rho, C_{\mathsf{Q}}, C$ *are positive numerical constants.*

Incomplete and Corrupted Low-rank Matrix



Imagine we only observe a fraction entries of a corrupted matrix $\boldsymbol{Y} = \boldsymbol{L}_o + \boldsymbol{S}_o$ on a support $\mathsf{O} \sim Ber(\rho_o)$. Hence the measurement model is:

$$\mathcal{P}_{\mathsf{O}}[\boldsymbol{Y}] = \mathcal{P}_{\mathsf{O}}[\boldsymbol{L}_o + \boldsymbol{S}_o] = \mathcal{P}_{\mathsf{O}}[\boldsymbol{L}_o] + \boldsymbol{S}_o'.$$

A natural convex program to solve here is:

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{L}\|_* + \lambda\|\boldsymbol{S}\|_1 \\ \text{subject to} & \mathcal{P}_{\mathsf{O}}[\boldsymbol{L} + \boldsymbol{S}] = \mathcal{P}_{\mathsf{O}}[\boldsymbol{Y}]. \end{array} \tag{15}$$

This combines matrix completion and robust PCA.

Theoretical Guarantee

Theorem (Matrix Completion with Corruptions)

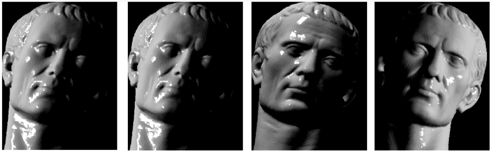
Suppose $\boldsymbol{L}_o$ is $n \times n$, obeys the incoherence conditions. Suppose $\rho_0 > C_0 \frac{\nu r \log^2 n}{n}$ and $\rho_s \leq C_s$, and let $\lambda = \frac{1}{\sqrt{\rho_0 n \log n}}$. Then the optimal solution to the convex program (15) is exactly $\boldsymbol{L}_o$ and $\boldsymbol{S}_o'$ with probability at least $1 - Cn^{-3}$ for some constant C, provided the constants C_0 is large enough and C_s is small enough.

- Robust PCA: If $\rho_0 = 1$, the above condition $1 > C_0 \frac{\nu r \log^2 n}{n}$ gives $r < C_0^{-1} n \nu^{-1} (\log n)^{-2}$ for small enough C_0^{-1}, the condition for robust PCA.
- Matrix Completion: if $\rho_s = 0$, the above theorem guarantees perfect recovery as long as $\rho_0 > C_0 \frac{\nu r \log^2 n}{n}$ for large enough C_0, the condition for matrix completion.

Example: Photometric Stereo (Ch. 14)

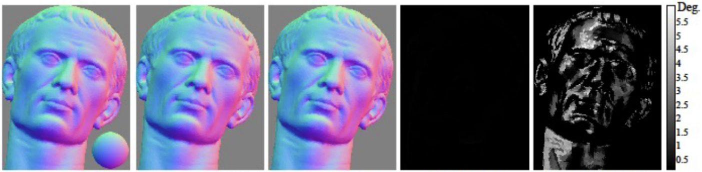
Recovering 3D shape of an object from images under different lightings.

Input images



$$\min \ \|A\|_* + \lambda \|E\|_1 \quad \text{subj} \quad D = \mathcal{P}_\Omega(A + E).$$

$\Omega^c \sim \text{shadow}(20.7\%)$
$E \sim \text{specularities}(13.6\%)$



(a) Ground truth (b) Our method (c) Least Squares (d) Error map (our method) (e) Error map (LS)

Mean error	**0.014°**	0.96°
Max error	**0.20°**	8.0°

Summary: Sparse & Low-Rank

Sparse v.s. Low-rank	Sparse Vector	Low-rank Matrix
Low-dimensionality of	individual signal $\boldsymbol{x}$	a set of signals $\boldsymbol{X}$
Low-dim measure	ℓ^0 norm $\|\boldsymbol{x}\|_0$	$\mathrm{rank}(\boldsymbol{X})$
Convex surrogate	ℓ^1 norm $\|\boldsymbol{x}\|_1$	nuclear norm $\|\boldsymbol{X}\|_*$
Compressive sensing	$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$	$\boldsymbol{Y} = \mathcal{A}(\boldsymbol{X})$
Stable recovery	$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{z}$	$\boldsymbol{Y} = \mathcal{A}(\boldsymbol{X}) + \boldsymbol{Z}$
Error correction	$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$	$\boldsymbol{Y} = \mathcal{A}(\boldsymbol{X}) + \boldsymbol{E}$
Recovery of mixed structures	$\mathcal{P}_{\mathsf{Q}}[\boldsymbol{Y}] = \mathcal{P}_{\mathsf{Q}}[\boldsymbol{L}_o + \boldsymbol{S}_o] + \boldsymbol{Z}$	

"An idea which can be used once is a trick. If one can use it more than once it becomes a method."

– *George Pólya and Gábor Szegö*

General Low-Dim Structures (Ch. 6)

Definition (Atomic Gauge)

The *atomic gauge* associated with a dictionary $\mathcal{D}$ is the function

$$\|\boldsymbol{x}\|_{\mathcal{D}} \doteq \inf \left\{ \sum_{i=1}^{k} \alpha_i \;\middle|\; \alpha_1, \ldots, \alpha_k \geq 0,\ \boldsymbol{d}_1, \ldots, \boldsymbol{d}_k \in \mathcal{D} \text{ s.t. } \sum_i \alpha_i \boldsymbol{d}_i = \boldsymbol{x} \right\}.$$

To recover $\boldsymbol{x}_o$ from $\boldsymbol{y} = \mathcal{A}(\boldsymbol{x}_o)$, solve the convex minimization problem:

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_{\mathcal{D}} \quad \text{subject to} \quad \mathcal{A}[\boldsymbol{x}] = \boldsymbol{y}. \tag{16}$$

Let D denote the *descent cone* of the atomic norm $\|\cdot\|_{\mathcal{D}}$ at $\boldsymbol{x}_o$. Then

- $\mathbb{P}[\text{(16) recovers } \boldsymbol{x}_o] \;\leq\; C \exp\left(-c\frac{(\delta(\mathsf{D})-m)^2}{n}\right), \qquad m \leq \delta(\mathsf{D});$
- $\mathbb{P}[\text{(16) recovers } \boldsymbol{x}_o] \;\geq\; 1 - C \exp\left(-c\frac{(m-\delta(\mathsf{D}))^2}{n}\right), \; m \geq \delta(\mathsf{D}).$

Limitations of Convex Programs

- Limitations of Convexification (Ch. 7):
  for example, $\boldsymbol{X}$ is simultaneously sparse and low-rank, the convex relaxation $\lambda_1\|\boldsymbol{X}\|_1 + \lambda_2\|\boldsymbol{X}\|_*$ is not optimal.
- Nonlinearity due to Domain Transformation (Ch. 15):
  for example, $\boldsymbol{I} \circ \tau = \boldsymbol{L} + \boldsymbol{S}$ for a low-rank $\boldsymbol{L}$ and sparse $\boldsymbol{S}$.
- Nonlinearity due to Nonlinear Observation (Ch. 16):
  $\boldsymbol{Y} = g(\boldsymbol{X})$ for some nonlinear function $g(\cdot)$ and low-dim $\boldsymbol{X}$.

We will deal with nonconvex and nonlinearity in later lectures.

Assignments

- Reading: Section 5.4 - 5.6 of Chapter 5.
- Programming Homework #3.