CSE 8803RS: Recommendation Systems

Lecture 13: Robust Principal Componet Analysis?

Liangda Li

School of Computational Science & Engineering
College of Computing
Georgia Institute of Technology

The General Question to Address

- For a data matrix which is the superposition of a low-rank component and a sparse component, can we recover each component individually?

Model Definition

- Suppose we are given a large data matrix M, and know that it may be decomposed as

$$M = L_0 + S_0$$

  where L_0 has low-rank and S_0 is sparse.

- **Principal Component Analysis**
  seek the best rank-k estimate of L_0 by solving

$$\begin{aligned} &\text{minimize} \|M - L\| \\ &\text{subject to } rank(L) \leq k \end{aligned}$$

  Throughout the paper, $\|M\|$ denotes the 2-norm; that is, the largest singular value of M.

A Surprising Message

- **Principal Component Pursuit(PCP)**
  Let $\|M\|_* = \sum_i \sigma_i(M)$ denote the nuclear norm of the matrix M, and $\|M\|_1 = \sum_{ij} |M_{ij}|$ denote the l_1 norm of M.

$$\begin{aligned} &\text{minimize} \|L\|_* + \lambda \|S\|_1 \\ &\text{subject to } L + S = M \end{aligned}$$

Assumptions

- The subset of observed entries Ω is uniformly random.
- With the singular value decomposition of L_0 as

$$L_0 = U\Sigma V^* = \sum_{i=1}^{r} \sigma_i u_i v_i^*$$

The factors U, V satisfy incoherence condition

$$max_i \|U^* e_i\|^2 \leq \frac{\mu r}{n_1}, max_i \|V^* e_i\|^2 \leq \frac{\mu r}{n_2}$$

and

$$\|UV^*\|_\infty \leq \sqrt{\frac{\mu r}{n_1 n_2}}$$

Here and below, $\|M\|_\infty = max_{i,j}|M_{ij}|$

- Ensure that low-rank matrix and sparse matrix can be distinguished.

Main Result

- Define $n_{(1)} = max(n_1, n_2)$, and $n_{(2)} = min(n_1, n_2)$.
- Suppose L_0 is $n \times n$, and the support set of S_0 is uniformly distributed among all sets of cardinality m. Then there is a numerical constant c such that with probability at least $1 - cn^{-10}$, `Principal Component Pursuit` with $\lambda = 1/\sqrt{n}$ is exact, i.e. $\hat{L} = L_0$ and $\hat{S} = S_0$, provided that
$$rank(L_0) \leq \rho_r n \mu^{-1} (\log n)^{-2} \quad and \quad m \leq \rho_s n^2$$
Above, ρ_r and ρ_s are positive numerical constants.
- In the general rectangular case where L_0 is $n_1 \times n_2$, PCP with $\lambda = 1/sqrt n_{(1)}$ succeeds with probability at least $1 - cn_{(1)}^{-10}$, provided that $rank(L_0) \leq \rho_r n_{(2)} \mu^{-1} (\log n_{(1)})^{-2}$ and $m \leq \rho_s n_1 n_2$.

Implications for Matrix Completion from Grossly Corrupted Data

- Let $\mathcal{P}_\Omega$ be the orthogonal projection onto the linear space of matrices supported on $\Omega \subset [n1] \times [n2]$,

$$\mathcal{P}_\Omega X = \begin{cases} X_{ij} & (i,j) \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

- Imagine we only have available a few entries of $L_0 + S_0$, which we conveniently write as

$$Y = \mathcal{P}_{\Omega_{obs}}(L_0 + S_0) = \mathcal{P}_{\Omega_{obs}} L_0 + S_0^{'}$$

  we propose recovering L_0 by solving the following problem:
  Principal Component Pursuit

$$\begin{aligned} &\text{minimize} \|L\|_* + \lambda \|S\|_1 \\ &\text{subject to } \mathcal{P}_{\Omega_{obs}}(L + S) = Y \end{aligned}$$

Corresponding Result

- Suppose L_0 is $n \times n$, and Ω_{obs} is uniformly distributed among all sets of cardinality m obeying $m = 0.1n^2$. Suppose for simplicity, that each observed entry is corrupted with probability τ independently of the others. Then there is a numerical constant c such that with probability at least $1 - cn^{-10}$, `Principal Component Pursuit` with $\lambda = 1/\sqrt{0.1n}$ is exact, i.e. $\hat{L} = L_0$, provided that

$$rank(L_0) \leq \rho_r n \mu^{-1} (\log n)^{-2} \quad and \quad \tau \leq \tau_s$$

  Above, ρ_r and τ_s are positive numerical constants.

- In the general rectangular case where L_0 is $n_1 \times n_2$, PCP with $\lambda = 1/\sqrt{0.1 n_{(1)}}$ succeeds from $m = 0.1 n_1 n_2$ corrupted entries with probability at least $1 - c n_{(1)}^{-10}$, provided that $rank(L_0) \leq \rho_r n_{(2)} \mu^{-1} (\log n_{(1)})^{-2}$.

Exact Recovery from Varying Fractions of Error

Dimension n	$\text{rank}(L_0)$	$\|S_0\|_0$	$\text{rank}(\hat{L})$	$\|\hat{S}\|_0$	$\frac{\|\hat{L}-L_0\|_F}{\|L_0\|_F}$	# SVD	Time(s)
500	25	12,500	25	12,500	1.1×10^{-6}	16	2.9
1,000	50	50,000	50	50,000	1.2×10^{-6}	16	12.4
2,000	100	200,000	100	200,000	1.2×10^{-6}	16	61.8
3,000	250	450,000	250	450,000	2.3×10^{-6}	15	185.2

$\text{rank}(L_0) = 0.05 \times n$, $\|S_0\|_0 = 0.05 \times n^2$.

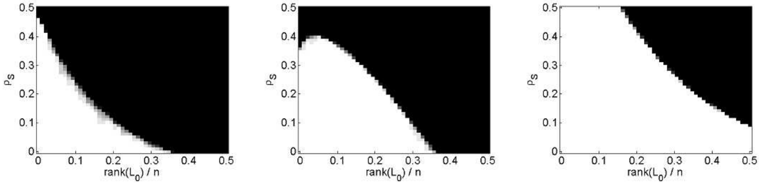
Dimension n	$\text{rank}(L_0)$	$\|S_0\|_0$	$\text{rank}(\hat{L})$	$\|\hat{S}\|_0$	$\frac{\|\hat{L}-L_0\|_F}{\|L_0\|_F}$	# SVD	Time(s)
500	25	25,000	25	25,000	1.2×10^{-6}	17	4.0
1,000	50	100,000	50	100,000	2.4×10^{-6}	16	13.7
2,000	100	400,000	100	400,000	2.4×10^{-6}	16	64.5
3,000	150	900,000	150	900,000	2.5×10^{-6}	16	191.0

$\text{rank}(L_0) = 0.05 \times n$, $\|S_0\|_0 = 0.10 \times n^2$.

Table 1: Correct recovery for random problems of varying size. Here, $L_0 = XY^* \in \mathbb{R}^{n \times n}$ with $X, Y \in \mathbb{R}^{n \times r}$; X, Y have entries i.i.d. $\mathcal{N}(0, 1/n)$. $S_0 \in \{-1, 0, 1\}^{n \times n}$ has support chosen uniformly at random and independent random signs; $\|S_0\|_0$ is the number of nonzero entries in S_0. Top: recovering matrices of rank $0.05 \times n$ from 5% gross errors. Bottom: recovering matrices of rank $0.05 \times n$ from 10% gross errors. In all cases, the rank of L_0 and ℓ_0-norm of S_0 are correctly estimated. Moreover, the number of partial singular value decompositions (# SVD) required to solve PCP is almost constant.

Notice that in all cases, solving the convex PCP gives a result (L, S) with the correct rank and sparsity. Moreover, the relative error $\|L - L_0\|_F / \|L0\|_F$ is small, less than 10^{-5} in all examples considered.

Phase Transition in Rank and Sparsity



(a) Robust PCA, Random Signs (b) Robust PCA, Coherent Signs (c) Matrix Completion

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, $\mathrm{sgn}(S_0)$ random. Middle: low-rank and sparse decomposition, $S_0 = \mathcal{P}_\Omega \mathrm{sgn}(L_0)$. Right: matrix completion. For matrix completion, ρ_s is the probability that an entry is omitted from the observation.

Notice that there is a large region in which the recovery is exact. This highlights an interesting aspect of our result: the recovery is correct even though in some cases $\|S_0\|_F \gg \|L_0\|_F$.

Background Modeling from Surveillance Video

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

Convex optimization (this work) Alternating minimization [47]

Figure 2: Background modeling from video. Three frames from a 200 frame video sequence taken in an airport [32]. (a) Frames of original video M. (b)-(c) Low-rank $\hat{L}$ and sparse components $\hat{S}$ obtained by PCP, (d)-(e) competing approach based on alternating minimization of an m-estimator [47]. PCP yields a much more appealing result despite using less prior knowledge.

Background Modeling from Surveillance Video-cont.

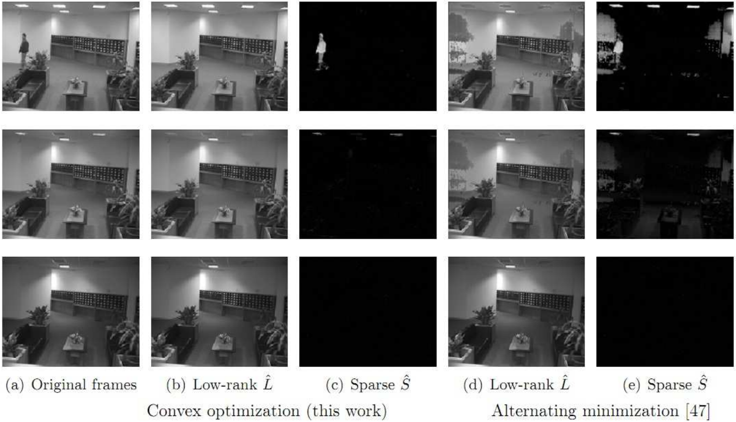


Figure 3: Background modeling from video. Three frames from a 250 frame sequence taken in a lobby, with varying illumination [32]. (a) Original video M. (b)-(c) Low-rank $\hat{L}$ and sparse $\hat{S}$ obtained by PCP. (d)-(e) Low-rank and sparse components obtained by a competing approach based on alternating minimization of an m-estimator [47]. Again, convex programming yields a more appealing result despite using less prior information.

Removing Shadows and Specularities from Face Images

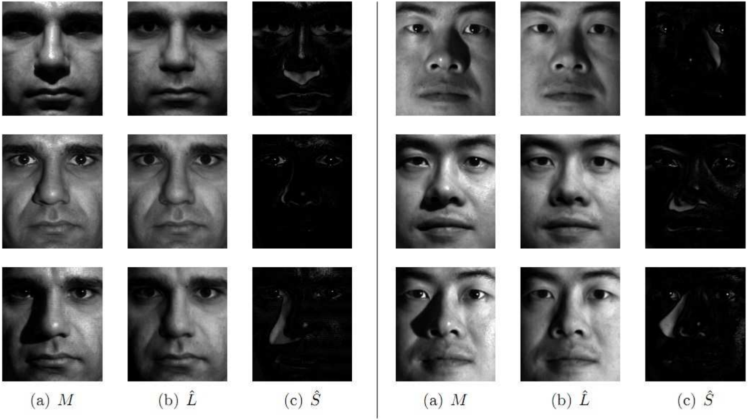


Figure 4: Removing shadows, specularities, and saturations from face images. (a) Cropped and aligned images of a person's face under different illuminations from the Extended Yale B database. The size of each image is 192×168 pixels, a total of 58 different illuminations were used for each person. (b) Low-rank approximation $\hat{L}$ recovered by convex programming. (c) Sparse error $\hat{S}$ corresponding to specularities in the eyes, shadows around the nose region, or brightness saturations on the face. Notice in the bottom left that the sparse term also compensates for errors in image acquisition.

Architecture of the Proof

- Any subgradient of the l_1 norm at S_0 supported on Ω, is of the form

$$\operatorname{sgn}(S_0) + F$$

  where F vanishes on Ω, i.e. $\mathcal{P}_\Omega F = 0$, and obeys $\|F\|_\infty \leq 1$.
- Any subgradient of the nuclear norm at L_0 is of the form

$$UV^* + W$$

  where $U^*W = 0$, $WV = 0$ and $\|W\| \leq 1$. Denote by T the linear space of matrices

$$T = \{UX^* + YV^*, X, T \in \mathcal{R}^{n \times r}\}$$

  and by $T^\perp$ its orthogonal complement.

An Elimination Theorem

- We will say that $S^{'}$ is a trimmed version of S if $supp(S^{'}) \subset supp(S)$ and $S^{'}_{ij} = S_{ij}$ whenever $S^{'}_{ij} \neq 0$.
- Suppose the solution to PCP with input data $M_0 = L_0 + S_0$ is unique and exact, and consider $M^{'}_0 = L_0 + S^{'}_0$, where $S^{'}_0$ is a trimmed version of S_0. Then the solution to PCP with input $M^{'}_0$ is exact as well.
- **The Bernoulli model**
  $\Omega = \{(i;j) : \delta_{ij} = 1\}$, where the Ω_{ij}'s are i.i.d. variables Bernoulli taking value one with probability ρ and zero with probability $1-\rho$, so that the expected cardinality of Ω is ρn^2. From now on, we will write $\Omega \sim Ber(\rho)$ as a shorthand for Ω is sampled from the Bernoulli model with parameter ρ.

Derandomization

- Suppose L_0 obeys the conditions and that the locations of the nonzero entries of S_0 follow the Bernoulli model with parameter $2\rho_s$, and the signs of S_0 are i.i.d. ± 1 as above (and independent from the locations). Then if the PCP solution is exact with high probability, then it is also exact with at least the same probability for the model in which the signs are fixed and the locations are sampled from the Bernoulli model with parameter ρ_s.

Dual Certificates

- Assume that $\|\mathcal{P}_\Omega \mathcal{P}_T\| < 1$. With the standard notations, (L_0, S_0) is the unique solution of there is a pair (W, F) obeying

$$UV^* + W = \lambda(sgn(S_0) + F)$$

  with $\mathcal{P}_T W = 0$, $\|W\| < 1$, $\mathcal{P}_\Omega F = 0$ and $\|F\|_\infty < 1$.
- Assume that $\|\mathcal{P}_\Omega \mathcal{P}_T\| < 1/2$ and $\lambda < 1$. With the standard notations, (L_0, S_0) is the unique solution of there is a pair (W, F) obeying

$$UV^* + W = \lambda(sgn(S_0) + F + \mathcal{P}_\Omega D)$$

  with $\mathcal{P}_T W = 0$, $\|W\| < 1/2$, $\mathcal{P}_\Omega F = 0$ and $\|F\|_\infty < 1/2$, and $\|P_\Omega D\|_F \leq 1/4$.
- (a)$W \in T^\perp$; (b) $\|W\| < 1/2$; (c) $\|P_\Omega(UV^* - \lambda \mathrm{sgn}(S_0) + W\|_F \leq \lambda/4$; (d) $\|P_{\Omega^\perp}(UV^* + W)\|_\infty \leq \lambda/2$.

Dual Certification via the Golfing Scheme

- We propose constructing a dual certificate

$$W = W^L + W^S$$

- Construction of W^L via the golfing scheme.
  For an integer $j_0 \geq 1$, and let Ω_j, $1 \leq j \leq j_0$, be defined so that $\Omega^c = \cup_{1 \leq j \leq j_0} \Omega_j$. Then starting with $Y_0 = 0$, inductively define

$$Y_j = Y_{j-1} + q^{-1} \mathcal{P}_{\Omega_j} \mathcal{P}_T (UV^* - Y_{j-1})$$

  and set

$$W^L = \mathcal{P}_{T^\perp} Y_{j_0}$$

- Construction of W^S via the method of least squares.
  Assume that $\|\mathcal{P}_\Omega \mathcal{P}_T\| < 1/2$. Then $\|\mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega\| < 1/4$, and thus the operator $\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega$ mapping Ω onto itself is invertible. Set

$$W^S = \lambda \mathcal{P}_{\Omega_\perp} (\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega)^{-1} \mathrm{sgn}(S_0)$$

- (a) $\|W^L + W^S\| < 1/2$; (b) $\|P_\Omega(UV^* + W^L\|_F \leq \lambda/4$; (c) $\|P_{\Omega^\perp}(UV^* + W^L + W^S)\|_\infty \leq \lambda/2$.

Key Lemmas

- Suppose Ω_0 is sampled from the Bernoulli model with parameter ρ_0. Then with high probability,

$$\|\mathcal{P}_T - \rho_0^{-1}\mathcal{P}_T\mathcal{P}_{\Omega_0}\mathcal{P}_T\| \leq \epsilon$$

  provided that $\rho_0 \geq C_0\epsilon^{-2}\frac{\mu r \log n}{n}$ for some numerical constant $C_0 > 0$. For rectangular matrices, we need $\rho_0 \geq C_0\epsilon^{-2}\frac{\mu r \log n_{(1)}}{n_{(2)}}$.
- Assume that $\Omega \sim Ber(\rho)$, then $\|\mathcal{P}_\Omega\mathcal{P}_T\|^2 \leq \rho + \epsilon$, provided that $1 - \rho \geq C_0\epsilon^{-2}\frac{\mu r \log n}{n}$.
- Assume that $\Omega \sim Ber(\rho)$ with parameter $\rho \leq \rho_s$ for some $\rho_s > 0$. Set $j_0 = 2\lceil \log n \rceil$. Then the matrix W^L obeys
  (a) $\|W^L\| < 1/4$,
  (b) $\|\mathcal{P}_\Omega(UV^* + W^L)\|_F < \lambda/4$,
  (c) $\|\mathcal{P}_{\Omega^\perp}(UV^* + W^L)\|_\infty < \lambda/4$.
- Assume that S_0 is supported on set Ω, and that the signs of S_0 are i.i.d. symmetric. Then the matrix W^S obeys
  (a) $\|W^S\| < 1/4$, (b) $\|\mathcal{P}_{\Omega^\perp}W^S\|_\infty < \lambda/4$.

Algorithm for Principal Component Pursuit

The PCP problem is solved using an augmented Lagrange multiplier (ALM) algorithm, which operates on the augmented Lagrangian

$$l(L, S, Y) = \|L\|_* + \lambda\|S\|_1 + < Y, M - L - S > + \frac{\mu}{2}\|M - L - S\|_F^2$$

Let $\mathcal{S}_\tau$ denote the shrinkage operator $\mathcal{S}_\tau[x] = sgn(x)max(|x| - \tau, 0)$, and $\mathcal{D}_\tau(X)$ denote the singular value thresholding operator given by $\mathcal{D}_\tau(X) = U\mathcal{S}_\tau(\Sigma)V^*$, where $X = U\Sigma V^*$ is any singular value decompositon, we propose algorithm as below:

Algorithm 1 Principal Component Pursuit by Alternating Directions

initialize: $S_0 = Y_0 = 0$, $\mu > 0$.
while not converged **do**
  compute $L_{k+1} = \mathcal{D}_{\mu^{-1}}(M - S_k + \mu^{-1}Y_k)$;
  compute $S_{k+1} = \mathcal{S}_{\lambda\mu^{-1}}(M - L_{k+1} + \mu^{-1}Y_k)$;
  compute $Y_{k+1} = Y_k + \mu(M - L_{k+1} - S_{k+1})$;
end while