ECE 8201: Low-dimensional Signal Models for High-dimensional Data Analysis

Lecture 8: Robust PCA

Yuejie Chi
The Ohio State University

THE OHIO STATE UNIVERSITY

Main Reference

- E. J. Candès, X. Li, Y. Ma, and J. Wright. "Robust Principal Component Analysis?" Journal of ACM 58(1), 1-37.

- V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. "Rank-sparsity incoherence for matrix decomposition." SIAM Journal on Optimization 21, no. 2 (2011): 572-596.

Outline

- Motivating applications
- Mathematical formulation

Sparse+ Low-rank Matrix Decomposition

Suppose we are given a matrix of data observations:

$$\boldsymbol{M} = \boldsymbol{L} + \boldsymbol{S},$$

where $\boldsymbol{L}$ is low-rank and $\boldsymbol{S}$ is sparse. We do not know the rank of $\boldsymbol{L}$ nor the sparsity level of $\boldsymbol{S}$.

Question: Can we recover both $\boldsymbol{L}$ and $\boldsymbol{S}$ from $\boldsymbol{M}$? What if we only partially observe $\boldsymbol{M}$?

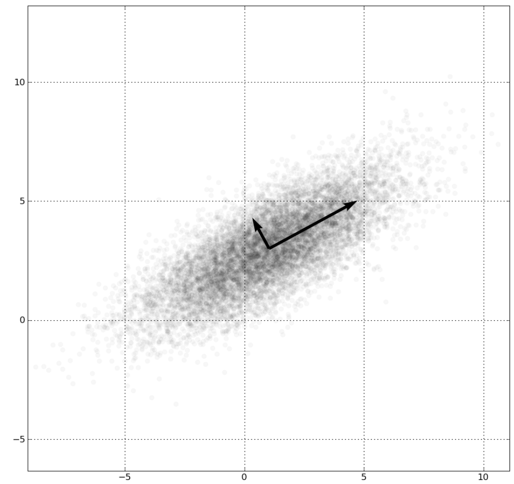
This problem has many applications in data-intensive problems.

Principal Component Analysis

Consider p data samples $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p]$ that are centered, $\boldsymbol{x}_i \in \mathbb{R}^n$. PCA seeks the direction that explains most of the variance of the data. Mathematically, we seek the direction $\boldsymbol{a} \in \mathbb{R}^n$ (principal component) that maximizes

$$\boldsymbol{a} = \underset{\|\boldsymbol{a}\|_2=1}{\operatorname{argmax}} \, \boldsymbol{a}^T \boldsymbol{X} \boldsymbol{X}^T \boldsymbol{a} = \underset{\|\boldsymbol{a}\|_2=1}{\operatorname{argmin}} \min_{\boldsymbol{b}} \|\boldsymbol{X} - \boldsymbol{a}\boldsymbol{b}^T\|_F^2$$

corresponding to seek the rank-one matrix approximation of $\boldsymbol{X}$.



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 $\boldsymbol{X}$, i.e.

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

where r is the number of PCs.

Many applications of PCA:

- feature extraction;
- dimensionality reduction;

PCA justifies the approximate low-rank assumption on $\boldsymbol{X}$.

Corruptions

What if the data samples $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p]$ are corrupted?

- Outliers/Gross errors due to sensor errors/attacks/etc: each entry in $\boldsymbol{x}_i$ corresponds to a sensor,

$$\boldsymbol{y}_i = \boldsymbol{x}_i + \boldsymbol{s}_i$$

  where $\boldsymbol{s}_i$ is a sparse vector with the nonzero entries corresponds to outliers. The corrupted data can be written as

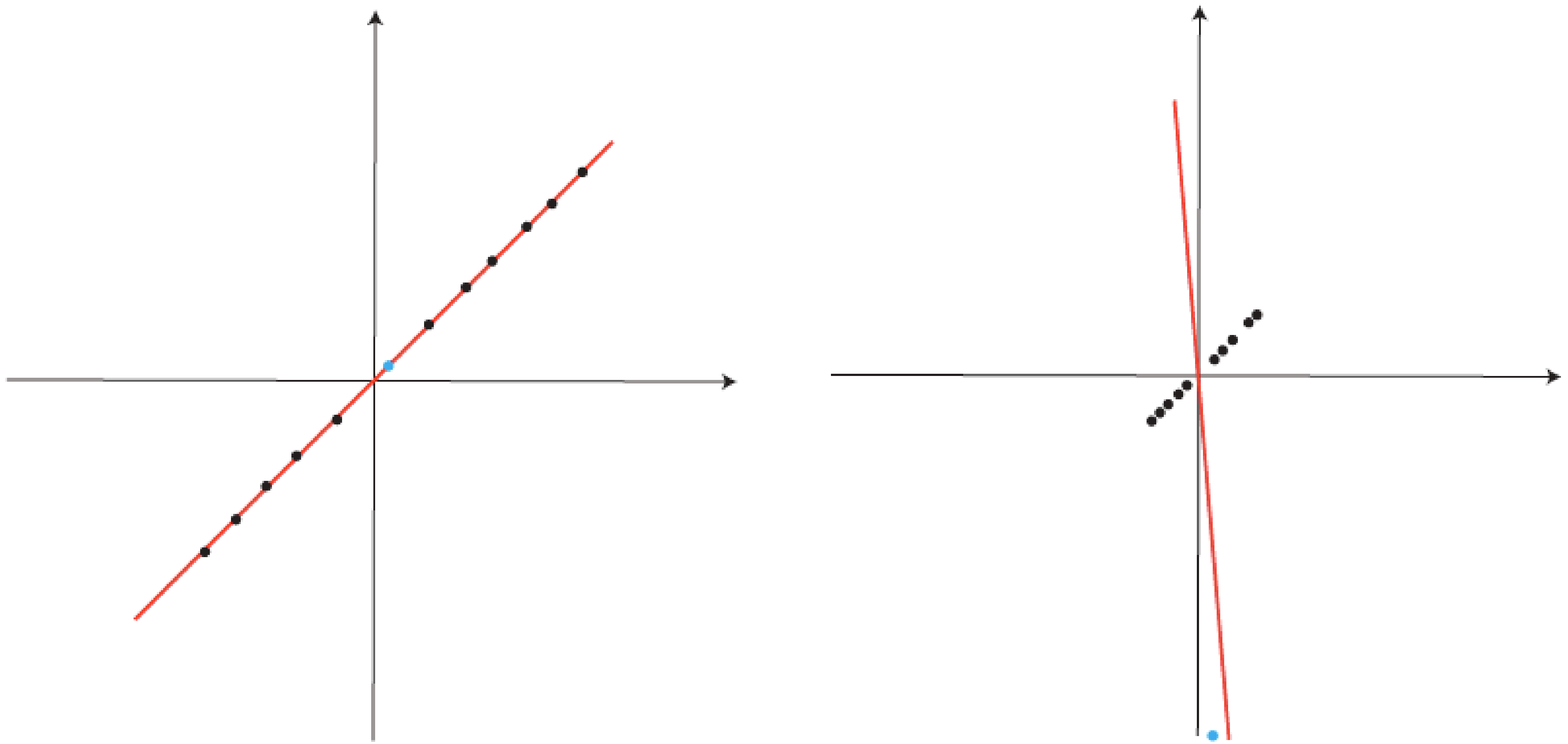
$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{S}$$

- The nominal PCA fails even with a few outliers:

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

Illustration

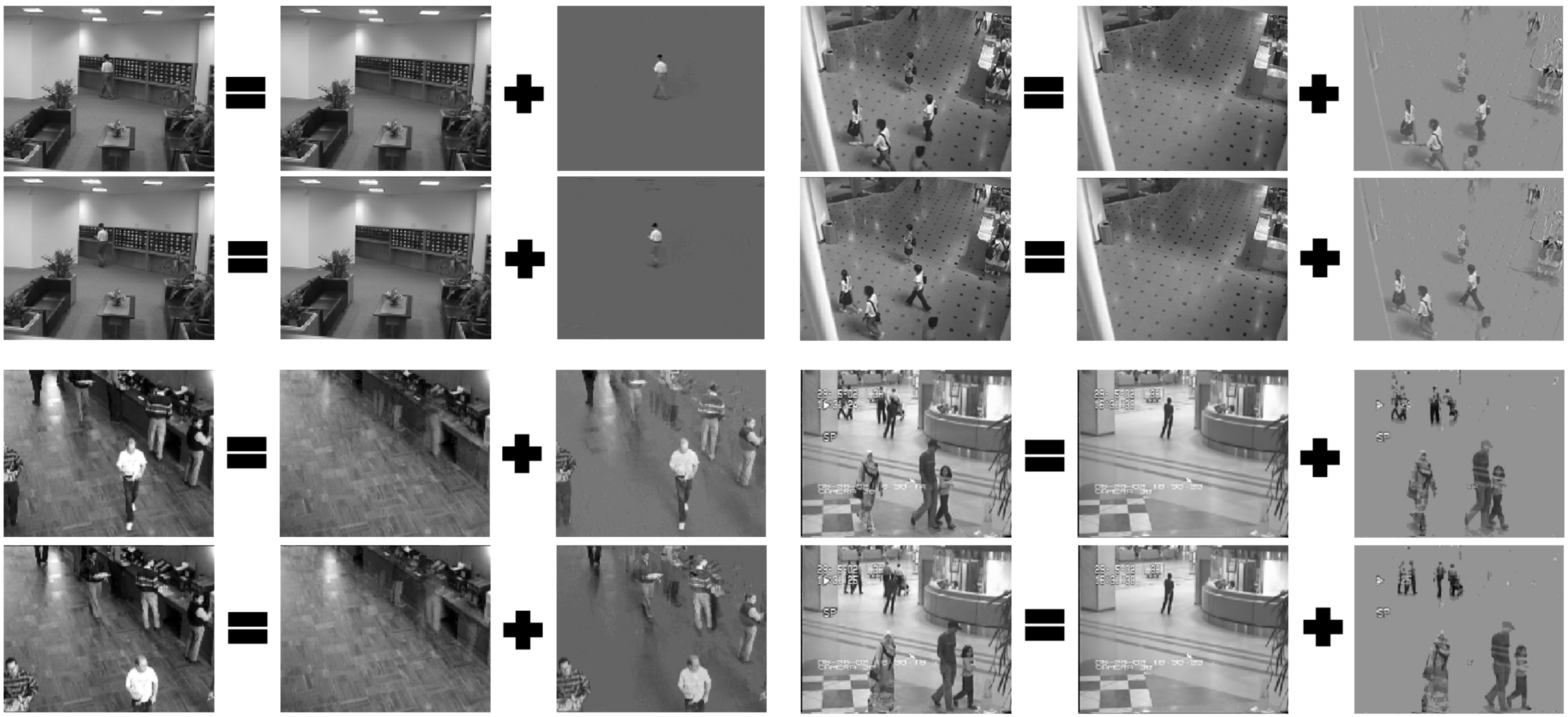
The nominal PCA could fail even with one outlier:



Video surveillance

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

$$\boldsymbol{M} = \boldsymbol{L} + \boldsymbol{S}$$



Graphical modeling

Consider a collection of random variables that are jointly Gaussian $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$:

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

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

- The nonzero entries of $\boldsymbol{P}$ describes the conditional independence between the variables, which can be depicted in a *graphical model*.
- **Graphical model learning:** Given i.i.d. samples of $\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, we want to learn the support of $\boldsymbol{P}$.
- An interesting case is when $\boldsymbol{P}$ 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 $\boldsymbol{x}_o$ as the observed variables;
- denote $\boldsymbol{x}_h$ as the hidden variables (latent factors);

The precision matrix of all data can be written as

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{P}_o & \boldsymbol{P}_{o,h} \\ \boldsymbol{P}_{h,o} & \boldsymbol{P}_h \end{bmatrix}$$

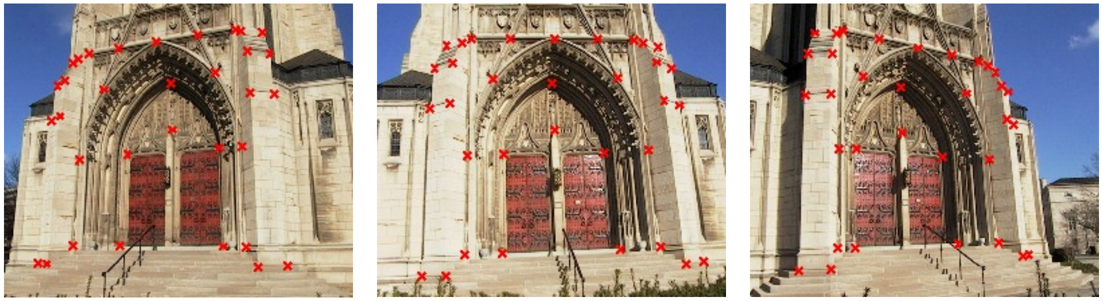
We only observe the *marginal* precision matrix on the observed variables $\boldsymbol{x}_0$:

$$\boldsymbol{\Sigma}_o^{-1} = \boldsymbol{P}_o - \boldsymbol{P}_{o,h}\boldsymbol{P}_h^{-1}\boldsymbol{P}_{h,o}$$

- $\boldsymbol{P}_o$ is sparse due to conditional independence;
- $\boldsymbol{P}_{o,h}\boldsymbol{P}_h^{-1}\boldsymbol{P}_{h,o}$ 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 $\boldsymbol{x}_{i,j}^T \in \mathbb{R}^2$ corresponding to the location of the ith point in the jth frame, define the matrix

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

- Occlusions: missing entries in $\boldsymbol{M}$;
- Wrong feature point/correspondence: sparse corruptions in $\boldsymbol{M}$;

Sparse+ Low-rank Decomposition: when is it possible?

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

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \quad \text{vs} \quad \begin{bmatrix} 1 & 0 & 1 & \cdots & 1 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & 1 & & \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Would the sparse component to be spread.
we assume its support is uniformly at random.

$$\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix} \quad \text{vs} \quad \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

Would the low-rank component to be incoherent.

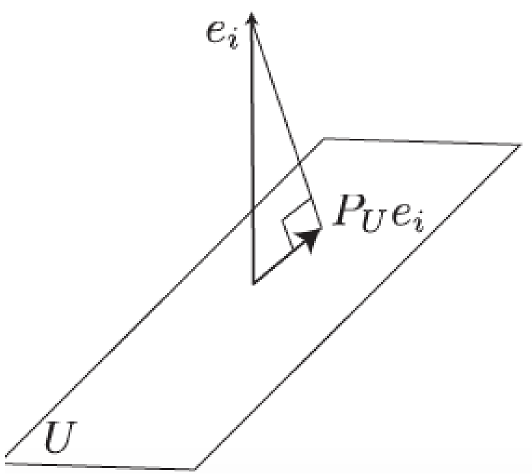
Low-rank component: Coherence

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

Definition 1. [Coherence] *Smallest scalar μ_1 obeying*

$$\max_{1 \le i \le n} \|\boldsymbol{U}^T \boldsymbol{e}_i\|_2^2 \le \mu_1 \frac{r}{n}, \quad \max_{1 \le i \le n} \|\boldsymbol{V}^T \boldsymbol{e}_i\|_2^2 \le \mu_1 \frac{r}{n},$$

where $\boldsymbol{e}_i$ is the ith standard basis vector.



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

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

Low-rank component: Joint Coherence

Definition 2. [Joint Coherence] *Smallest scalar μ_2 obeying*

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

This avoids $\boldsymbol{U}\boldsymbol{V}^T$ to be too peaky.

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

$$|(\boldsymbol{U}\boldsymbol{V}^T)_{ij}| = |\boldsymbol{u}_i^T \boldsymbol{v}_j| \leq \frac{\mu_1 r}{n}$$

$$\|\boldsymbol{U}\boldsymbol{V}^T\|_\infty \geq \frac{1}{n}\sum_i (\boldsymbol{U}\boldsymbol{V}^T)_{ij}^2 = \frac{1}{n}\left\|\boldsymbol{V}^T\boldsymbol{e}_j\right\|_2^2$$

- The incoherence parameter μ_1 is sufficient and necessary for MC, while μ_2 is necessary for Robust PCA (connection to the planted clique problem [c.f. Chen, 2015]).

Algorithm

Non-convex heuristic:

$$(\hat{\boldsymbol{L}}, \hat{\boldsymbol{S}}) = \operatorname*{argmin}_{\boldsymbol{L},\boldsymbol{S}} \mathsf{rank}(\boldsymbol{L}) + \lambda \|\boldsymbol{S}\|_0, \quad \text{s.t.} \quad \boldsymbol{M} = \boldsymbol{L} + \boldsymbol{S}.$$

Convex relaxation: Principal Component Pursuit (PCP)

$$(\hat{\boldsymbol{L}}, \hat{\boldsymbol{S}}) = \operatorname*{argmin}_{\boldsymbol{L},\boldsymbol{S}} \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{S}\|_1, \quad \text{s.t.} \quad \boldsymbol{M} = \boldsymbol{L} + \boldsymbol{S}$$

where $\|\cdot\|_*$ is the nuclear norm, and $\|\cdot\|_1$ is the entry-wise ℓ_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:*

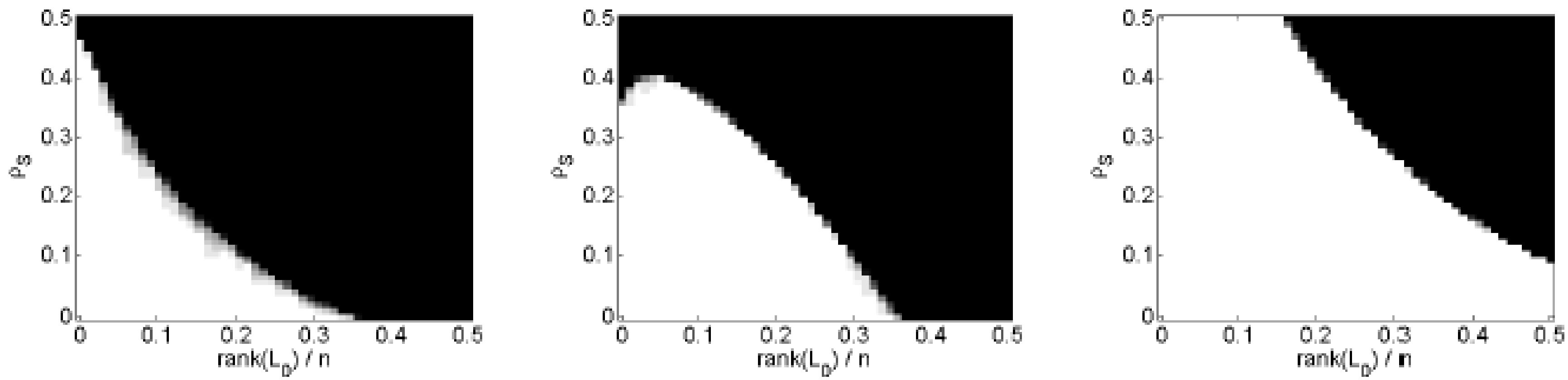
$$\hat{L} = L_0, \quad \hat{S} = S_0$$

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

Remark:

- No tuning parameters: $\lambda = 1/\sqrt{n}$ is prefixed by the theorem.
- Essentially optimal: rank$(\boldsymbol{L}) = O(n)$, $\|\boldsymbol{S}\|_0 = O(n^2)$
- Arbitrary magnitudes and sign patterns of $\boldsymbol{L}$ and $\boldsymbol{S}$!

Phase transition



(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, 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, ρ_s is the probability that an entry is omitted from the observation.

Connections with Matrix Completion

Comparison with Matrix Completion:

MC: missing

RPCA: corrupted

- In MC we know where the entries are missing; while in RPCA we do not know the locations of corruptions.

MC with Corruptions

What if we have both missing data and corruptions?

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

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

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

- A natural extension of RPCA:

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

MC with Corruptions: Guarantee

Theorem

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

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

$$\hat{L} = L_0$$

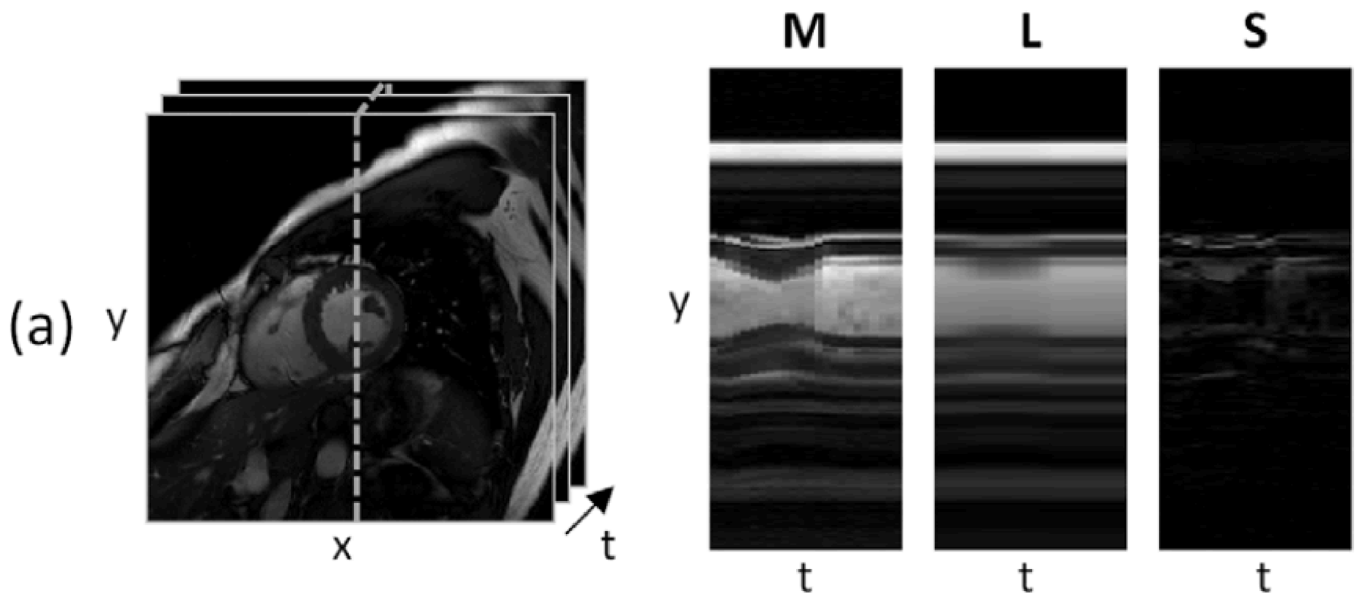
Same conclusion for rectangular matrices with $\lambda = 1/\sqrt{0.1 \text{max dim}}$

[a] missing fraction is arbitrary

- No tuning parameters: $\lambda = 1/\sqrt{n}$ is prefixed by the theorem.
- Essentially optimal: rank$(\boldsymbol{L}) = O(n)$, $\|\boldsymbol{S}\|_0 = O(m)$
- Arbitrary magnitudes and sign patterns of $\boldsymbol{L}$ and $\boldsymbol{S}$!

Application in Accelerated MRI

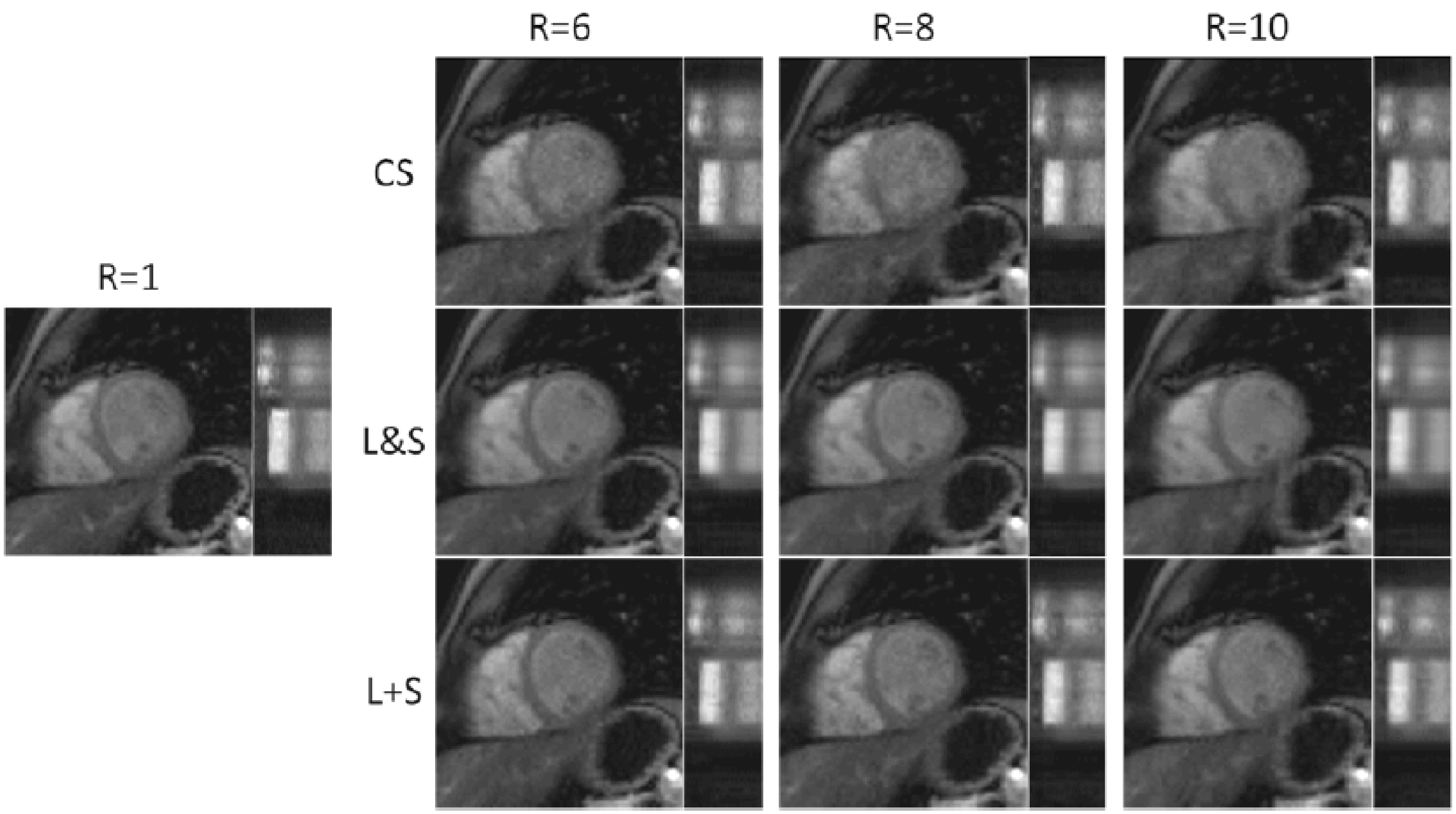
[Otazo et.al. 2014]: "The combination of compressed sensing and low-rank matrix completion represents an attractive proposition for further increases in imaging speed..."



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 betters priors on the signals helps performance!