Nonconvex Relaxation Approaches to Robust Matrix Recovery

Shusen Wang — Zhejiang University, China
Joint work with Dehua Liu and Zhihua Zhang

IJCAI 2013
Outline

1. Background
   - Problem Definition
   - Previous Work

2. Methodology
   - Better Sparse-Inducing Penalty Functions
   - Better Low-Rank-Inducing Penalty Functions
   - Model Formulation
   - Algorithm

3. Experiments
Outline

1 Background
   ■ Problem Definition
   ■ Previous Work

2 Methodology
   ■ Better Sparse-Inducing Penalty Functions
   ■ Better Low-Rank-Inducing Penalty Functions
   ■ Model Formulation
   ■ Algorithm

3 Experiments
Low-Rank Matrix Recovery Problem

- Problem: recover a low-rank matrix from contaminated observation.
- How? Factorize the observation into low-rank matrix + noise.
Low-Rank Matrix Recovery Problem

- Problem: recover a low-rank matrix from contaminated observation.
- How? Factorize the observation into low-rank matrix + noise.
Why Low-Rank Matrix?

- Many kinds of matrices can be well approximated by low-rank ones.
Why Low-Rank Matrix?

- Many kinds of matrices can be well approximated by low-rank ones.
- Example: video surveillance sequence
Why Low-Rank Matrix?

- Many kinds of matrices can be well approximated by low-rank ones.

- Example: natural image
Penalized Rank Minimization

- **Notation**
  - $L_0$: low-rank matrix
  - $S_0$: noise matrix
  - $D = L_0 + S_0$: observation

- Recover the low-rank matrix $L_0$ and noise $S_0$ by solving
  - $\min_{L,S} \text{rank}(L) + \lambda \|S\|_F^2$ \hspace{1cm} $\text{s.t. } L + S = D$;
  - or $\min_{L,S} \text{rank}(L) + \lambda \|S\|_0$ \hspace{1cm} $\text{s.t. } L + S = D$;

- However, the two optimization problems are intractable.
Penalized Rank Minimization

- **Notation**
  - $L_0$: low-rank matrix
  - $S_0$: noise matrix
  - $D = L_0 + S_0$: observation

- **Recover the low-rank matrix $L_0$ and noise $S_0$ by solving**
  - $\min_{L,S} \text{rank}(L) + \lambda \|S\|_F^2$ \quad s.t. $L + S = D$;
  - or $\min_{L,S} \text{rank}(L) + \lambda \|S\|_0$ \quad s.t. $L + S = D$;

- However, the two optimization problems are intractable.
Penalized Rank Minimization

Notation
- \( L_0 \): low-rank matrix
- \( S_0 \): noise matrix
- \( D = L_0 + S_0 \): observation

Recover the low-rank matrix \( L_0 \) and noise \( S_0 \) by solving
- \( \min_{L,S} \text{rank}(L) + \lambda \|S\|_F^2 \quad \text{s.t.} \quad L + S = D; \)
- or \( \min_{L,S} \text{rank}(L) + \lambda \|S\|_0 \quad \text{s.t.} \quad L + S = D; \)

However, the two optimization problems are intractable.
Penalized Rank Minimization

- **Notation**
  - \( L_0 \): low-rank matrix
  - \( S_0 \): noise matrix
  - \( D = L_0 + S_0 \): observation

- Recover the low-rank matrix \( L_0 \) and noise \( S_0 \) by solving
  - \( \min_{L,S} \text{rank}(L) + \lambda \|S\|_F^2 \quad \text{s.t. } L + S = D; \)
  - or \( \min_{L,S} \text{rank}(L) + \lambda \|S\|_0 \quad \text{s.t. } L + S = D; \)

- However, the two optimization problems are intractable.
Penalized Rank Minimization
Convex relaxations

- Replace matrix rank $\text{rank}(L)$ by the nuclear norm $\|L\|_*$
  - $\|L\|_* = \sum_i \sigma_i(L)$ (sum of singular values)
  - $\|L\|_*$ is the tightest convex approximation to $\text{rank}(L)$

- Convex Relaxations: $\text{rank}(L) \Rightarrow \|L\|_*$ and $\|S\|_0 \Rightarrow \|S\|_1$

$$\min_{L,S} \text{rank}(L) + \lambda \|S\|_0 \quad \text{s.t.} \quad L + S = D;$$

$$\downarrow$$

$$\min_{L,S} \|L\|_* + \lambda \|S\|_1 \quad \text{s.t.} \quad L + S = D;$$

Known as Robust PCA [Candès et al., 2011]
Penalized Rank Minimization
Convex relaxations

- Replace matrix rank \( \text{rank}(L) \) by the nuclear norm \( \|L\|_* \)
  - \( \|L\|_* = \sum_i \sigma_i(L) \) (sum of singular values)
  - \( \|L\|_* \) is the tightest convex approximation to \( \text{rank}(L) \)

- Convex Relaxations: \( \text{rank}(L) \Rightarrow \|L\|_* \) and \( \|S\|_0 \Rightarrow \|S\|_1 \)
  \[
  \min_{L,S} \text{rank}(L) + \lambda \|S\|_0 \quad \text{s.t. } L + S = D;
  \]
  \[
  \Downarrow
  \]
  \[
  \min_{L,S} \|L\|_* + \lambda \|S\|_1 \quad \text{s.t. } L + S = D;
  \]

Known as Robust PCA [Candès et al., 2011]
Penalized Rank Minimization
Convex relaxations

Question:
Is the nuclear norm a good relaxation of matrix rank?
Outline

1 Background
   - Problem Definition
   - Previous Work

2 Methodology
   - Better Sparse-Inducing Penalty Functions
   - Better Low-Rank-Inducing Penalty Functions
   - Model Formulation
   - Algorithm

3 Experiments
Revisiting $\ell_1$-Norm Penalized Regression

- The $\ell_1$-norm penalized regression problem
  \[
  \min_{\mathbf{w} \in \mathcal{W}} \| \mathbf{y} - \mathbf{Xw} \|_2^2 + \lambda \| \mathbf{w} \|_1.
  \]
  The $\ell_1$-norm penalty contributes to the shrinkage and selection of the entries of $\mathbf{w}$, and the solutions are sparse.

- LASSO is an effective alternative of the intractable $\ell_0$-norm penalized problem
  \[
  \min_{\mathbf{w} \in \mathcal{W}} \| \mathbf{y} - \mathbf{Xw} \|_2^2 + \lambda \| \mathbf{w} \|_0.
  \]

- Is the $\ell_1$-norm a good approximation to the $\ell_0$-norm?
Revisiting $\ell_1$-Norm Penalized Regression

- The $\ell_1$-norm penalized regression problem
  \[
  \min_{w \in \mathcal{W}} \| y - Xw \|_2^2 + \lambda \| w \|_1.
  \]
  The $\ell_1$-norm penalty contributes to the shrinkage and selection of the entries of $w$, and the solutions are sparse.

- LASSO is an effective alternative of the intractable $\ell_0$-norm penalized problem
  \[
  \min_{w \in \mathcal{W}} \| y - Xw \|_2^2 + \lambda \| w \|_0.
  \]
  Is the $\ell_1$-norm a good approximation to the $\ell_0$-norm?
Revisiting $\ell_1$-Norm Penalized Regression

- The $\ell_1$-norm penalized regression problem
  \[
  \min_{w \in \mathcal{W}} \| y - Xw \|^2_2 + \lambda \| w \|_1.
  \]
  The $\ell_1$-norm penalty contributes to the shrinkage and selection of the entries of $w$, and the solutions are sparse.

- LASSO is an effective alternative of the intractable $\ell_0$-norm penalized problem
  \[
  \min_{w \in \mathcal{W}} \| y - Xw \|^2_2 + \lambda \| w \|_0.
  \]

- Is the $\ell_1$-norm a good approximation to the $\ell_0$-norm?
Revisiting $\ell_1$-Norm Penalized Regression

- $w_0 = [10, 2, 0, 0, 0, 0]$
- $w^* = [9.9, 1.8, 0, 0, 0, 0]$
Revisiting $\ell_1$-Norm Penalized Regression

- $w_0 = [10, 2, 0, 0, 0]$
- $w^* = [9.9, 1.8, 0, 0, 0]$
Revisiting $\ell_1$-Norm Penalized Regression

- Is the $\ell_1$-norm a good approximation to $\ell_0$-norm?
- NO. According to [Fan & Li, 2001]
  - The $\ell_1$-norm over-penalizes large variable
  - The resulting solution is a biased estimation
Revisiting $\ell_1$-Norm Penalized Regression

- Is the $\ell_1$-norm a good approximation to $\ell_0$-norm?
- NO. According to [Fan & Li, 2001]
  - The $\ell_1$-norm over-penalizes large variable
  - The resulting solution is a biased estimation

![Graph comparing $\ell_0$-norm and $\ell_1$-norm](image)
Several nonconvex functions better approximate the $\ell_0$-norm than the $\ell_1$-norm does, and their resulting solutions are nearly unbiased.

For example, the minimax concave penalty (MCP) [Zhang, 2010] is defined by

$$M_\gamma(w) = \sum_{i=1}^{p} \psi_\gamma(w_i),$$

where

$$\gamma > 0 \quad \text{and} \quad \psi_\gamma(t) = \begin{cases} \frac{\gamma}{2} & \text{if } |t| \geq \gamma, \\ |t| - \frac{t^2}{2\gamma} & \text{otherwise}. \end{cases}$$
Better Sparse-Inducing Penalty
The minimax concave penalty (MCP)

- Plot of the minimax concave penalty (MCP) [Zhang, 2010]:

- When $\gamma \to \infty$, MCP function $M_\gamma(w) \to \|w\|_1$;
- When $\gamma \to 0^+$, MCP function behaves like the $\ell_0$-norm.
Low-Rank-Inducing Penalty & Sparse-Inducing Penalty

- Let $\mathbf{L}$ be a matrix and $\sigma(\mathbf{L})$ be a vector containing the singular values of $\mathbf{L}$, then

$$\text{rank}(\mathbf{L}) = \|\sigma(\mathbf{L})\|_0$$
$$\|\mathbf{L}\|_* = \|\sigma(\mathbf{L})\|_1.$$  

- Optimizing over $\|\sigma(\mathbf{L})\|_1$ shrinks small singular values to zero.

- Bad News: the $\ell_1$-norm over-penalizes large entries

  $\implies \|\mathbf{L}\|_* = \|\sigma(\mathbf{L})\|_1$ over-penalizes large singular values!

- Good News: apply MCP function to $\sigma(\mathbf{L})$ for low-rank-inducing.
Low-Rank-Inducing Penalty & Sparse-Inducing Penalty

- Let $L$ be a matrix and $\sigma(L)$ be a vector containing the singular values of $L$, then
  \[
  \text{rank}(L) = \| \sigma(L) \|_0 \\
  \| L \|_* = \| \sigma(L) \|_1.
  \]

- Optimizing over $\| \sigma(L) \|_1$ shrinks small singular values to zero.

- Bad News: the $\ell_1$-norm over-penalizes large entries
  \[
  \implies \| L \|_* = \| \sigma(L) \|_1
  \]
  over-penalizes large singular values!

- Good News: apply MCP function to $\sigma(L)$ for low-rank-inducing.
Let $\mathbf{L}$ be a matrix and $\sigma(\mathbf{L})$ be a vector containing the singular values of $\mathbf{L}$, then

$$\text{rank}(\mathbf{L}) = \|\sigma(\mathbf{L})\|_0,$$

$$\|\mathbf{L}\|_* = \|\sigma(\mathbf{L})\|_1.$$

Optimizing over $\|\sigma(\mathbf{L})\|_1$ shrinks small singular values to zero.

Bad News: the $\ell_1$-norm over-penalizes large entries

$$\implies \|\mathbf{L}\|_* = \|\sigma(\mathbf{L})\|_1$$

over-penalizes large singular values!

Good News: apply MCP function to $\sigma(\mathbf{L})$ for low-rank-inducing.
Better Low-Rank-Inducing Penalty
MCP on singular values

- $M_\gamma(\sigma(L))$ is a tighter approximation to $\text{rank}(L)$ than $\|L\|_*$ is, and it alleviates the over-penalization on large singular values.
  - $M_\gamma(\sigma(L))$ bridges the matrix rank and the nuclear norm;
  - When $\gamma \to \infty$, $M_\gamma(\sigma(L)) \to \|L\|_*$;
  - When $\gamma \to 0_+$, $M_\gamma(\sigma(L))$ corresponds to $\text{rank}(L)$. 
Better Low-Rank-Inducing Penalty

- The MCP function $M_\gamma(w)$ better approximates $\|w\|_0$ than $\|w\|_1$ does.

- The MCP function on singular values $M_\gamma(\sigma(L))$ better approximates $\text{rank}(L)$ than $\|L\|_*$ does.
Nonconvex Optimization Model for Matrix Recovery

- **Matrix recovery by optimization:**
  \[
  \min_{L,S} \text{rank}(L) + \lambda \|S\|_0; \quad \text{s.t. } L + S = D. \tag{1}
  \]
  \[
  \min_{L,S} \|L\|_* + \lambda \|S\|_1; \quad \text{s.t. } L + S = D. \tag{2}
  \]

- **Nonconvex optimization for matrix recovery:**
  \[
  \min_{L,S} M_{\gamma_1}(\sigma(L)) + \lambda M_{\gamma_2}(S); \quad \text{s.t. } L + S = D.
  \]
Nonconvex Optimization Model for Matrix Recovery

Matrix recovery by optimization:

\[
\min_{L,S} \|\sigma(L)\|_0 + \lambda\|S\|_0; \quad \text{s.t. } L + S = D. \tag{1}
\]

\[
\min_{L,S} \|\sigma(L)\|_1 + \lambda\|S\|_1; \quad \text{s.t. } L + S = D. \tag{2}
\]

Nonconvex optimization for matrix recovery:

\[
\min_{L,S} M_{\gamma_1}(\sigma(L)) + \lambda M_{\gamma_2}(S); \quad \text{s.t. } L + S = D.
\]
Matrix recovery by optimization:

\[
\begin{align*}
\min_{L,S} & \quad \| \sigma(L) \|_0 + \lambda \| S \|_0; \\
& \quad \text{s.t. } L + S = D. \quad (1)
\end{align*}
\]

\[
\begin{align*}
\min_{L,S} & \quad \| \sigma(L) \|_1 + \lambda \| S \|_1; \\
& \quad \text{s.t. } L + S = D. \quad (2)
\end{align*}
\]

Nonconvex optimization for matrix recovery:

\[
\min_{L,S} M_{\gamma_1}(\sigma(L)) + \lambda M_{\gamma_2}(S); \quad \text{s.t. } L + S = D.
\]
Algorithm for the Nonconvex Optimization Problem

How to solve the nonconvex optimization problem?

\[
\min_{L,S} M_{\gamma_1}(\sigma(L)) + \lambda M_{\gamma_2}(S); \quad \text{s.t. } L + S = D,
\]

The Majorization-Minimization (MM) Algorithm [Hunter & Li, 2005; Zou & Li 2008] alternating between

1. Majorization: find a local convex approximation for the nonconvex objective function.
2. Minimization: solve the convex approximation.
Algorithm for the Nonconvex Optimization Problem

How to solve the nonconvex optimization problem?

\[ \min_{L, S} M_{\gamma_1}(\sigma(L)) + \lambda M_{\gamma_2}(S); \quad \text{s.t. } L + S = D, \]

The Majorization-Minimization (MM) Algorithm [Hunter & Li, 2005; Zou & Li 2008] alternating between

1. Majorization: find a local convex approximation for the nonconvex objective function.
2. Minimization: solve the convex approximation.
Algorithm for the Nonconvex Optimization Problem

The Majorization-Minimization (MM) Algorithm

1. Majorization: find a local convex approximation to the nonconvex objective function

\[
M_{\gamma_1}(\sigma(L)) + \lambda M_{\gamma_2}(S) \implies Q_{\gamma_1}(\sigma(L) \mid \sigma(L^{\text{old}})) + \lambda Q_{\gamma_2}(S \mid S^{\text{old}})
\]
The Majorization-Minimization (MM) Algorithm

1. Majorization
2. Minimization: solve the convex approximation

\[
\min_{L,S} Q_{\gamma_1}(\sigma(L) \sigma(L^{\text{old}})) + \lambda Q_{\gamma_2}(S \mid S^{\text{old}}); \quad \text{s.t. } L + S = D,
\]

Then replace \((L^{\text{old}}, S^{\text{old}})\) by the solution \((L^*, S^*)\)
1 Background
   - Problem Definition
   - Previous Work

2 Methodology
   - Better Sparse-Inducing Penalty Functions
   - Better Low-Rank-Inducing Penalty Functions
   - Model Formulation
   - Algorithm

3 Experiments
Experiments

- Image denoising: with 20% entries added with salt-and-pepper noise.

(a) original  
(b) noisy
Experiments

- Results: with all parameters tuned best.

(a) RPCA  
(b) NRMR
Experiments

- Results: with parameters set such that $\text{rank}(L^*) = 100$.

(a) RPCA  
(b) NRMR
Experiments

Relative error on 50 images, where

$$\text{Relative Error} = \frac{\|L^* - L_0\|_F}{\|L_0\|_F}.$$
Summary

- The $\ell_1$-norm over-penalizes large variable in regression problems, and the solution is biased. [Fan & Li, 2001].

- The nonconvex MCP function is a better relaxation of the $\ell_0$-norm, and it alleviates over-penalization. [Zhang, 2010].

- The nuclear norm over-penalizes large singular values because $\|L\|_* = \|\sigma(L)\|_1$.

- We propose to use the MCP function on singular values in low-rank matrix recovery.

- Our nonconvex approach achieves much higher performance than its convex counterpart—Robust PCA.
Summary

- The $\ell_1$-norm over-penalizes large variable in regression problems, and the solution is biased. [Fan & Li, 2001].

- The nonconvex MCP function is a better relaxation of the $\ell_0$-norm, and it alleviates over-penalization. [Zhang, 2010].

- The nuclear norm over-penalizes large singular values because $\|L\|_* = \|\sigma(L)\|_1$.

- We propose to use the MCP function on singular values in low-rank matrix recovery.

- Our nonconvex approach achieves much higher performance than its convex counterpart—Robust PCA.
The $\ell_1$-norm over-penalizes large variable in regression problems, and the solution is biased. [Fan & Li, 2001].

The nonconvex MCP function is a better relaxation of the $\ell_0$-norm, and it alleviates over-penalization. [Zhang, 2010].

The nuclear norm over-penalizes large singular values because $\|L\|_* = \|\sigma(L)\|_1$.

We propose to use the MCP function on singular values in low-rank matrix recovery.

Our nonconvex approach achieves much higher performance than its convex counterpart—Robust PCA.
Reference


