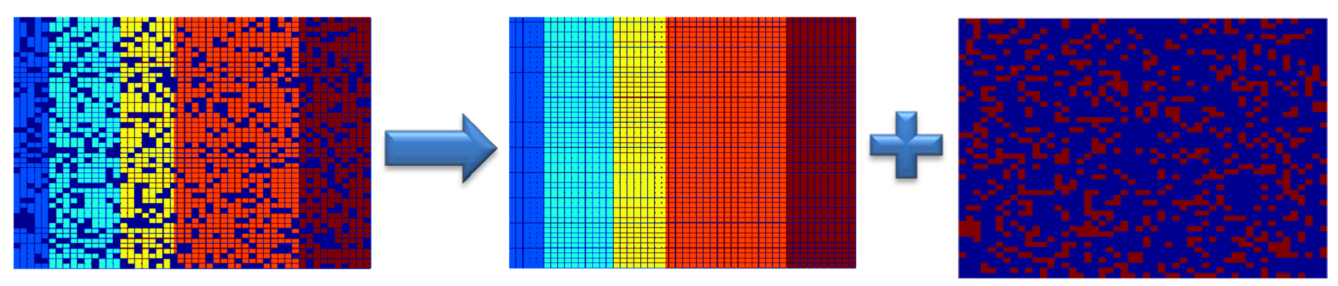
Nonconvex Relaxation Approaches to Robust Matrix Recovery

Shusen Wang, Dehua Liu, and Zhihua Zhang — Zhejiang University, Hangzhou, China

Low-Rank and Sparse Matrices Recovery

In many computer vision and machine learning problems, a data matrix can be represented as a low-rank component plus a sparse component.



Applications:

- Video surveillance is the superposition of low-rank background and sparse foreground;
- Corrupted image can be approximated by low-rank image plus sparse noises;
- In collaborative filtering problems the predicted rating matrix is the sum of a low-rank matrix and a sparse noise matrix.

Optimization Models for Matrix Recovery

Let $\mathbf{L}_0$ and $\mathbf{S}_0$ be the underlying low-rank matrix and sparse matrix respectively, and let $\mathbf{D} = \mathbf{L}_0 + \mathbf{S}_0$ be the observation. It is intuitive to recover $\mathbf{L}_0$ and $\mathbf{S}_0$ by solving the following optimization problem:

$$\min_{\mathbf{L},\mathbf{S}} \ \mathrm{rank}(\mathbf{L}) + \lambda\|\mathbf{S}\|_0; \qquad \text{s.t. } \mathbf{L}+\mathbf{S}=\mathbf{D}. \tag{1}$$

However, this problem is intractable.

A classical method called *robust principal component analysis (RPCA)* [1] tackles the problem by solving an alternative problem:

$$\min_{\mathbf{L},\mathbf{S}} \ \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1; \qquad \text{s.t. } \mathbf{L}+\mathbf{S}=\mathbf{D}. \tag{2}$$

This problem is convex and can be efficiently solved. When both $\mathrm{rank}(\mathbf{L}_0)$ and $\|\mathbf{S}_0\|_0$ are reasonably small and some other mild assumptions are held, RPCA exactly recovers $\mathbf{L}_0$ and $\mathbf{S}_0$.

Revisiting ℓ_1-Norm Penalized Regression

The ℓ_1-norm penalized regression problem

$$\min_{\mathbf{w}\in\mathcal{W}} \ \|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1. \tag{3}$$

is an effective alternative of the intractable ℓ_0-norm penalized problem

$$\min_{\mathbf{w}\in\mathcal{W}} \ \|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_0.$$

The ℓ_1-norm penalty contributes to the shrinkage and selection of the entries of $\mathbf{w}$, and the solution to (3) is sparse.

However, ℓ_1-norm is not a good surrogate of the ℓ_0-norm. Fan & Li (2001) [2] pointed out that the ℓ_1-norm over-penalizes large entries in the model $\mathbf{w}$ and the resulting solution is biased estimation.

Motivation

Is the nuclear norm a good relaxation of matrix rank? To answer this question, we draw an analogy between sparse and low-rank learning:

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

$$\mathrm{rank}(\mathbf{L}) = \|\boldsymbol{\sigma}_\mathbf{L}\|_0 \qquad \text{and} \qquad \|\mathbf{L}\|_* = \|\boldsymbol{\sigma}_\mathbf{L}\|_1. \tag{4}$$

Fan & Li (2001) [2] pointed out that the ℓ_1-norm over-penalizes large entries in the model $\mathbf{w}$, so the nuclear norm over-penalizes large singular values and is thus not a good relaxation of matrix rank.

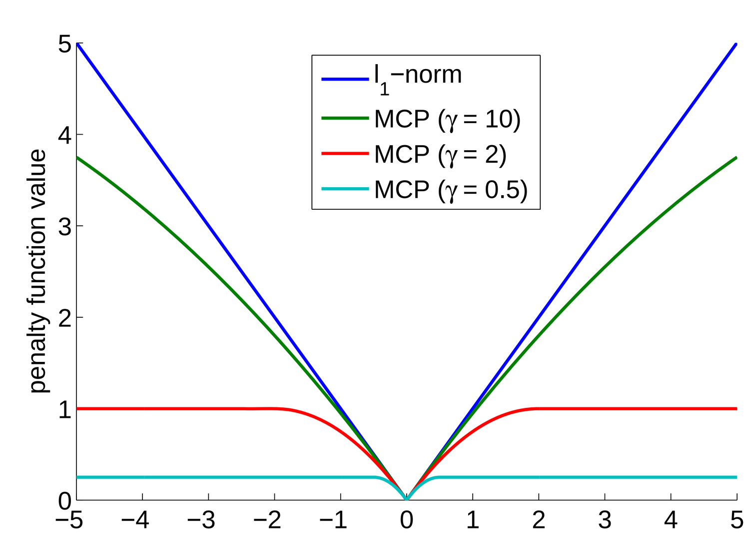
Methodology

Better sparse-inducing penalty functions. Several nonconvex sparse-inducing penalties better approximate the ℓ_0-norm than the ℓ_1-norm does, and their resulting solutions are nearly unbiased. For example, the minimax concave penalty (MCP) [3] is defined by

$$M_\gamma(\mathbf{w}) \;=\; \sum_{i=1}^{p} \psi_\gamma(w_i), \tag{5}$$

where

$$\gamma > 0 \quad \text{and} \quad \psi_\gamma(t) = \begin{cases} \gamma/2 & \text{if } |t| \geq \gamma, \\ |t| - \frac{t^2}{2\gamma} & \text{otherwise.} \end{cases}$$



When $\gamma \to \infty$, $\psi_\gamma(t) \to |t|$ and $M_\gamma(\cdot)$ becomes the ℓ_1-norm; when $\gamma \to 0_+$, it gives rise to a hard threshold operator corresponding to the ℓ_0-norm. Thus, $M_\gamma(\cdot)$ bridges the ℓ_1 and ℓ_0 norm.

A better low-rank inducing penalty function. $M_\gamma(\boldsymbol{\sigma}_\mathbf{L})$ is a tighter approximation to $\mathrm{rank}(\mathbf{L})$ than $\|\mathbf{L}\|_*$ is, and it alleviates the over-penalization on large singular values.

- $M_\gamma(\boldsymbol{\sigma}_\mathbf{L})$ bridges the matrix rank and the nuclear norm;
- When $\gamma \to \infty$, $M_\gamma(\boldsymbol{\sigma}_\mathbf{L}) \to \|\mathbf{L}\|_*$;
- When $\gamma \to 0_+$, $M_\gamma(\boldsymbol{\sigma}_\mathbf{L})$ corresponds to $\mathrm{rank}(\mathbf{L})$.

Nonconvex Optimization Model

Model. With the MCP function defined in (5), we establish the following nonconvex optimization model for matrix recovery.

$$\min_{\mathbf{L},\mathbf{S}} \ f(\mathbf{L},\mathbf{S}) = M_{\gamma_1}(\boldsymbol{\sigma}_\mathbf{L}) + \lambda M_{\gamma_2}(\mathbf{S}); \qquad \text{s.t. } \mathbf{L}+\mathbf{S}=\mathbf{D}, \tag{6}$$

Model (6) is a tighter approximation to (1) than (2) is.

Algorithm

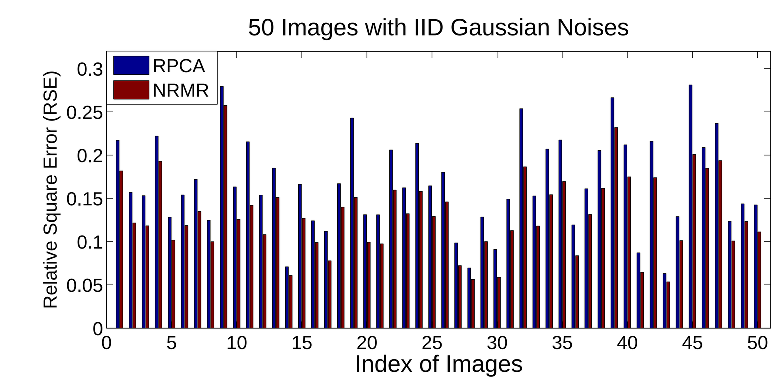
The nonconvex problem can be solved by the majorization-minimization (MM) algorithm [4] by alternating between the two steps until convergence:

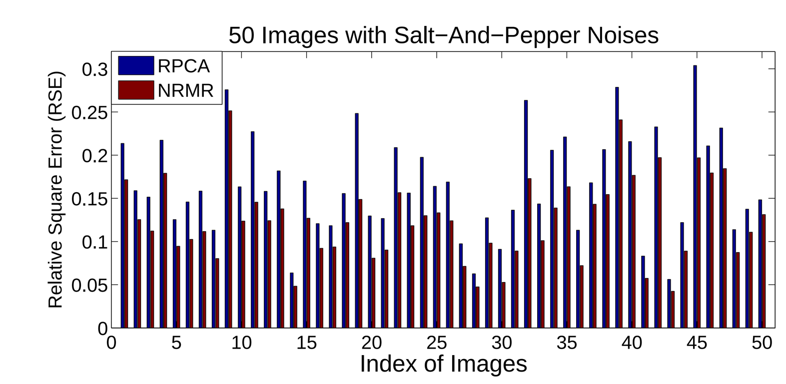
1. **Majorization.** Find a convex function $Q(\mathbf{L},\mathbf{S}|\mathbf{L}_0,\mathbf{S}_0)$ which locally approximates $f(\mathbf{L},\mathbf{S})$ at $(\mathbf{L}_0,\mathbf{S}_0)$:
$$\begin{aligned} Q(\mathbf{L},\mathbf{S}|\mathbf{L}_0,\mathbf{S}_0) \;&=\; \sum_l \Big(1-\sigma_l(\mathbf{L}_0)/\gamma_1\Big)_+ \Big(\sigma_l(\mathbf{L}) - \sigma_l(\mathbf{L}_0)\Big) \\ &+\; \lambda \sum_{i,j} \Big(1-|(S_0)_{ij}|/\gamma_2\Big)_+ \Big(|S_{ij}| - |(S_0)_{ij}|\Big) + \text{Const.} \end{aligned}$$
This is known as local linear approximation (LLA) [5].
2. **Minimization.** Solving the problem:
$$\min_{\mathbf{L},\mathbf{S}} \ Q(\mathbf{L},\mathbf{S}|\mathbf{L}_0,\mathbf{S}_0), \quad \text{s.t. } \mathbf{L}+\mathbf{S}=\mathbf{D},$$
which is weighted RPCA. Then it updates $\mathbf{L}_0$ and $\mathbf{S}_0$ by the optimal solution.

Since waiting for the iteration to converge is time consuming, so we propose to use an efficient strategy called *one-step LLA* studied in [5]. One-step LLA runs the loop only once instead of waiting to converge. According to our experiments, the results of full MM algorithm is only marginally better than one-step LLA. We use one-step LLA in all experiments to solve our nonconvex model.

Experiments

Here we show empirical comparisons between RPCA [1] and our method (**NRMR**) on 50 natural images (from the Berkeley Segmentation Dataset), each contaminated with i.i.d. Gaussian noises or salt-pepper noises. We use relative square error (RSE) to evaluate recovery accuracy: $\mathrm{RSE} = \|\mathbf{L}^* - \mathbf{L}\|_F / \|\mathbf{L}\|_F$.





On average, RCPA conducts 39.0 times SVD, while NRMR conducts 72.5 times SVD.

References

[1] E. Candès, X. Li, Y. Ma, and J. Wright: *Robust Principal Component Analysis?* Journal of the ACM, 2011.

[2] J. Fan and R. Li: *Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties.* Journal of the American Statistical Association, 2001.

[3] C.-H. Zhang: *Nearly Unbiased Variable Selection Under Minimax Concave Penalty.* The Annals of Statistics, 2010.

[4] R. Hunter and R. Li: *Variable Selection Using MM Algorithm.* Annals of Statistics, 2005.

[5] H. Zou and R. Li: *One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models.* The Annals of Statistics, 2008.