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 \( L_0 \) and \( S_0 \) be the underlying low-rank matrix and sparse matrix respectively, and let \( D = L_0 + S_0 \) be the observation. It is intuitive to recover \( L_0 \) and \( S_0 \) by solving the following optimization problem:

\[
\min_{L,S} \|L\|_* + \lambda \|S\|_1 ; \quad \text{s.t. } L + S = D. \tag{1}
\]

However, this problem is intractable.

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

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

This problem is convex and can be efficiently solved. When both \( \|L_0\|_* \) and \( \|S_0\|_1 \) are reasonably small and some other mild assumptions are held, RPCA exactly recovers \( L_0 \) and \( S_0 \).

Revisiting \( \ell_1 \)-Norm Penalized Regression

The \( \ell_1 \)-norm penalized regression problem

\[
\min_{w \in \mathbb{W}} \| y - Xw \|_2^2 + \lambda \|w\|_1, \tag{3}
\]

is an effective alternative of the intractable \( \ell_0 \)-norm penalized problem

\[
\min_{w \in \mathbb{W}} \| y - Xw \|_2^2 + \lambda \|w\|_0.
\]

The \( \ell_1 \)-norm penalty contributes to the shrinkage and selection of the entries of \( w \), and the solution to (3) is sparse.

However, \( \ell_1 \)-norm is not a good surrogate of the \( \ell_0 \)-norm. Fan & Li (2001) [2] pointed out that the \( \ell_1 \)-norm over-penalizes large entries in the model \( w \) 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 \( L \) be a matrix and \( \sigma_L \) be a vector containing the singular values of \( L \), then

\[
\text{rank}(L) = \|\sigma_L\|_0 \quad \text{and} \quad \|L\|_* = \|\sigma_L\|_1. \tag{4}
\]

Fan & Li (2001) [2] pointed out that the \( \ell_1 \)-norm over-penalizes large entries in the model \( w \), 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 \( \ell_0 \)-norm than the \( \ell_1 \)-norm does, and their resulting solutions are nearly unbiased. For example, the minimax concave penalty (MCP) [3] is defined by

\[
M_\gamma(w) = \sum_{i=1}^{n} \psi_\gamma(w_i), \tag{5}
\]

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

When \( \gamma \to \infty \), \( \psi_\gamma(t) \to |t| \) and \( M_\gamma(.) \) becomes the \( \ell_1 \)-norm; when \( \gamma \to 0_+ \), it gives rise to a hard threshold operator corresponding to the \( \ell_0 \)-norm. Thus, \( M_\gamma(.) \) bridges the \( \ell_1 \) and \( \ell_0 \) norm.

A better low-rank inducing penalty function. \( M_\gamma(\sigma_L) \) is a tighter approximation to \( \text{rank}(L) \) than \( \|L\|_* \), and it alleviates the over-penalization on large singular values. For example, \( M_\gamma(\sigma_L) \) bridges the matrix rank and the nuclear norm; when \( \gamma \to \infty \), \( M_\gamma(\sigma_L) \to \|L\|_* \), and when \( \gamma \to 0_+ \), \( M_\gamma(\sigma_L) \) corresponds to \( \text{rank}(L) \).

Nonconvex Optimization Model

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

\[
\min_{L,S} f(L, S) = M_\gamma(\sigma_L) + \lambda M_\gamma(S) ; \quad \text{s.t. } L + S = D, \tag{6}
\]

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 \( Q(L, S|L_0, S_0) \) which locally approximates \( f(L, S) \) at \( (L_0, S_0) \):

\[
Q(L, S|L_0, S_0) = \sum_{i} \left( 1 - \sigma_L(l_0)/\gamma \right) \left( \sigma_L(l) - \sigma_L(l_0) \right) + \sum_{i,j \neq i} \left( 1 - \|S_{ij}\|/\gamma \right) \left( \|S_{ij}\| - \|S_{ij}\| \right) + \text{Const.}
\]

This is known as local linear approximation (LLA) [5].

2. **Minimization.** Solving the problem:

\[
\min_{L,S} Q(L, S|L_0, S_0) , \quad \text{s.t. } L + S = D.
\]

which is weighted RPCA. Then it updates \( L_0 \) and \( S_0 \) 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-and-pepper noises. We use relative square error (RSE) to evaluate recovery accuracy:

\[
\text{RSE} = \frac{||L - L_0|| / ||L||_F}{f(L, S)}. \tag{7}
\]

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

References