Pattern Recognition Letters 71 (2016) 52–58

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Pathwise component descent method with MC+ penalty for low rank matrix recovery☆

Xiaofan Lin*, Gang Wei

School of Electronic and Information Engineering, South China University of Technology, Wushan Road, Tianhe District, Guangzhou 510641, PR China

ARTICLE INFO

Article history:
Received 4 August 2015
Available online 30 December 2015

Keywords:
Low rank matrix recovery
Low rank regularization
Non-convex optimization
Matrix completion

ABSTRACT

We consider the low rank matrix recovery (LRM) problem, which recovers an unknown low rank matrix from very limited information. Although recent study has shown that non-convex models for LRM can serve as better approximation of low rank regularization and outperform their convex counterpart, these models suffer from getting trapped into a bad local minimum. We propose an algorithm, named PIC-LR, to address this problem. PIC-LR is inspired by a recent algorithm, called SparseNet, which addresses analogical problem in the study of sparse optimization. Specifically, we take advantage of the properties of MC+ penalty and employ path-following technique. We also generalize coordinate descent to better imitate SparseNet. In numerical experiment, we apply PIC-LR to matrix completion problem, and the results show that PIC-LR outperforms several state-of-the-art solvers in terms of precision.

1. Introduction

There is a rapidly growing interest in the low rank matrix recovery (LRM) problem, which recovers an unknown low rank matrix from very limited information [7,12,15,20]. One common formulation of this problem can be given by

$$\min_X \operatorname{rank}(X), \quad s.t.\ \mathcal{A}(X) = B, \tag{1}$$

where the linear map $\mathcal{A}$ and the matrix B are known.

The above problem aims to find a matrix of minimum rank that satisfies a given system of linear equality constraints, which is a useful idea that can be applied to various applications in pattern recognition/machine learning, such as matrix completion [15], image processing [24], subspace segmentation [20], collaborative filtering/recommender system [7]. Although (1) is simple in form, it is a challenging problem due to the discrete nature of the rank function.

A commonly used heuristic introduced in [3] is replacing the rank function with the nuclear norm, which is the sum of the singular values of the matrix. This technique is based on the fact that the nuclear norm is the tightest convex relaxation of the matrix rank over the unit ball of matrices. The new formula can be given by

$$\min_X ||X||_*, \quad s.t.\ \mathcal{A}(X) = B, \tag{2}$$

where $||X||_* := \sum_{i=1}^{r} \sigma_i(X)$ denotes the nuclear norm. It is well-known that (2) can be rewritten as follow under some conditions[1]

$$\min_X \lambda ||X||_* + \frac{1}{2}||\mathcal{A}(X) - B||_F^2, \tag{3}$$

where $\lambda > 0$ is a Lagrange multiplier and $||\cdot||_F$ denotes the Frobenius norm of a matrix.

However, recent literature shows that using convex regularizer often leads to a biased and sub-optimal solution, and by contrast, some non-convex regularizers can sometimes serve as better approximation of low rank regularization and outperform their convex counterpart [11,21]. To take advantage of non-convex regularizers, several new regularizers such as the ℓ_p-norm and Minimax Concave Penalty have been proposed [11,22]. These theories and techniques usually first appear in the study of sparse regularizers, then generalized to low rank regularizers.

Unfortunately, many of these non-convex solvers often get trapped into a bad local minimum [13,23]. In this paper, we propose an algorithm, named PIC-LR, to address this problem. PIC-LR is inspired by a recent algorithm, called SparseNet, which addresses analogical problem in the study of sparse optimization. It has been shown in [13] that SparseNet successfully avoids bad local minimum and achieves better precision. It is thus desirable to

☆ This paper has been recommended for acceptance by Egon L. van den Broek.
* Corresponding author. Tel.: +86 150 175 25536.
E-mail address: walkerlin@foxmail.com (X. Lin).

[1] Recall the Lagrangian relaxation. Usually one use (3) as a more robust estimator, since the corresponding constraint is $||\mathcal{A}(X) - B||_F^2 \le \epsilon$, where ϵ is a certain noise level.

http://dx.doi.org/10.1016/j.patrec.2015.12.010
0167-8655/

investigate analogical algorithms that applicable to the more general LRM problem.

In SparseNet, three crucial techniques are applied: coordinate descent, path following, and recalibration. Our main contributions lies in generalizing the first two techniques to our problem setting.

In the first two techniques we applied, coordinate descent is an old and classical one, and is known as a highly scalable method [18,19]. The basic idea of coordinate descent method (CDM) is to partition the variable into disjoint (block) coordinates and handle a much smaller subproblem instead at each iteration. Applying this idea usually leads to better convergence rate and less computational cost [17]. For LRM however, the iterative procedure requires singular value decomposition (SVD), which is very tricky to perform with the detached submatrices of the matrix variable X. This obstacle indicates that the traditional CDM cannot be directly applied to LRM. Although several variants of CDM have been proposed for LRM, most of them categorize to greedy techniques and are not suitable for our framework. Therefore, we propose a new variants of CDM for PIC-LR, which is more natural and simple, and makes PIC-LR more analogous to SparseNet.

Another important technique, path following, can be interpreted as the process of tracing a sequence of solutions when the optimization problem is gradually modified from convex to non-convex. With this technique, we expect to find a good local optimal solution because such a process can be interpreted as simulated annealing [6].

Path following over the parameter has been studied in the context of ℓ_1-regularized feature selection [8], ℓ_1-MKL [1] and boosting [25]. However, it seems to us that this technique is rare in the study of LRM. To the best of our knowledge, this is the first work that path following with MC+ penalty is applied to the LRM problem. The popular competing technique to ours is the non-convex regularization, and we present comparative results in Section 6.

The remaining paper is organized as follows. In Section 2, we introduce some preliminaries and notations. In Section 3, we develop a variant of CDM for PIC-LR, and in Section 4 we present the PIC-LR algorithm. We then give convergence analysis in Section 5. Numerical results are reported in Section 6.

2. Preliminaries and notation

Block structure. We first introduce the block structure, which will be used in the development of a variant of CDM. Assume the variable of LRM problem is a matrix $X \in \mathbb{R}^{m\times n}$ of rank r, we model the block structure of the problem by decomposing X into r rank-one matrices. That is, we decompose $X = X_1 + X_2 + \cdots + X_r$, with $X_i, i = 1, 2, \ldots, r$ being rank-one matrices. For a given X, we construct a block diagonal matrix

$$e(X) = \begin{bmatrix} X_1 & & & \\ & X_2 & & \\ & & \ddots & \\ & & & X_r \end{bmatrix}, \tag{4}$$

where $e : \mathbb{R}^{m\times n} \to \mathbb{R}^{mr\times nr}$ maps a matrix to a block diagonal matrix. It will be assumed throughout the paper that the rank r is known or can be estimated.

For an iterative method, image that the order of sequence $\{X_i\}$ has been fixed since the first decomposition. At each iteration, the updated value of X_i is stored back into its original variable. This suggests that X_i plays a similar role as coordinates to CDM. To distinguish between them, we name X_i a "component".

We point out that in our method, it is unnecessary to maintain a huge-size matrix like $e(X)$, as the whole process is on component level. Besides, we can store each component in the form $X_i = u_i v_i^{\mathrm{T}}$, which dramatically reduces the number of primal variables from $m \times n$ to $m + n$. However, the above definitions are essential to understand the relevance between "component" and "coordinate".

For convenience we define the following operators: $e_i(X) : \mathbb{R}^{m\times n} \to \mathbb{R}^{mr\times nr}$ is a matrix whose ith component is X and the rest are zeros, i.e.,

$$e_i(X) = \begin{bmatrix} \ddots & & \\ & X & \\ & & \ddots \end{bmatrix} \in \mathbb{R}^{mr\times nr}. \tag{5}$$

$\mathcal{P}$ is a projector defined as $\mathcal{P}(X) = u_1\sigma_1(X)v_1^{\mathrm{T}}$, that is, keep the first singular value and the corresponding singular value vectors of X and eliminate the rest.

With above definitions, we now can write

$$e(X) = e_1(X_1) + e_2(X_2) + \cdots + e_r(X_r)$$

and

$$\mathcal{P}(X_i) = X_i, i = 1, 2, \ldots, r.$$

We will use above notations and definitions throughout the paper.

Smoothness of ℓ. Corresponding to block structure, we define a new concept "component-wise Lipschitz continuous", which is slightly different from that of coordinate-wise Lipschitz continuous (see, e.g., [10] for the definition of coordinate-wise Lipschitz continuous), as we operate component instead of coordinates, despite their analogy. Formally, we have the following definition.

Definition 1. A function $\ell : \mathbb{R}^{m\times n} \to \mathbb{R}^+$ is called component-wise Lipschitz continuous if

$$||\nabla\ell(X + Z_i) - \nabla\ell(X)||_{\mathrm{F}} \le L_i||Z_i||_{\mathrm{F}} \tag{6}$$

holds for all $X \in \mathbb{R}^{m\times n}$ and $Z_i \in \mathbb{R}^{m\times n}, i = 1, \ldots, r$ satisfying $\mathcal{P}(X_i + Z_i) = X_i + Z_i$.

Note that (6) is also different from the conventional definition of Lipschitz continuity

$$||\nabla\ell(X + Z) - \nabla\ell(X)||_{\mathrm{F}} \le L||Z||_{\mathrm{F}}, \tag{7}$$

since L_i is a single-"component" derivatives of ℓ. That is, from the perspective of block structure, $X + Z_i$ and X agree in all components except the ith component. Besides, (6) has an extra condition $\mathcal{P}(X_i + Z_i) = X_i + Z_i$. This suggests $L_i \le L$.

An important consequence of (6) is the following standard inequality (see, e.g., Lemma 1.2.3 in [16]):

$$\ell(X + Z_i) \le \ell(X) + \langle\nabla\ell(X), Z_i\rangle + \frac{L_i}{2}||Z_i||_{\mathrm{F}}^2, \tag{8}$$

The above inequality can be deduced from (6) with Taylor's theorem.

Note that throughout the rest of this paper, we always have $\ell(X) = \frac{1}{2}||\mathcal{A}(X) - B||_{\mathrm{F}}^2$, which clearly is component-wise Lipschitz continuous.

3. Component descent method with MC+ penalty

As mentioned in Section 1, non-convex models for LRM are popular in recent literature. A general formula for these models can be given by

$$\min_{X\in\mathbb{R}^{m\times n}} f(X) = \ell(X) + \lambda\sum_{i=1}^{r} P(\sigma_i(X)). \tag{9}$$

where $\sigma_i(X)$ is the ith singular value of block X, $\sum_{i=1}^{r} P(\sigma_i(X))$ is a non-convex penalty that tend to encourage the low rank structure of the solution, and $\lambda \in \mathbb{R}^+$ balance the effects of loss ℓ and penalty.

We will start with our new formula and its continuous properties, then develop a fixed parameter method for solving the formula. The fixed parameter method serves as a core step in PIC-LR presented in next section.