Information Sciences 415–416 (2017) 298–318

Contents lists available at ScienceDirect

Information Sciences

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

Enhanced joint sparsity via iterative support detection

Ya-Ru Fan[a], Yilun Wang[a,b,c,*], Ting-Zhu Huang[a]

[a] School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, Sichuan 611731 China
[b] PrinceTechs LLC., Shenzhen, Guangdong 518101 PR China
[c] Center for Applied Mathematics, Cornell University, Ithaca, NY 14853, USA

ARTICLE INFO

Article history:
Received 20 October 2015
Revised 19 June 2017
Accepted 24 June 2017
Available online 28 June 2017

Keywords:
Iterative support detection
Joint sparsity
$\ell_{2,1}$-norm minimization
Non-convex optimization
Compressive sensing
Multi-task feature learning

ABSTRACT

Joint sparsity has attracted considerable attention in recent years in many fields including sparse signal recovery in compressive sensing, statistics, and machine learning. Traditional convex models with joint sparsity suffer from the suboptimal performance though enjoying tractable computation. In this paper, we propose a new non-convex joint sparsity model, and develop a corresponding multi-stage adaptive convex relaxation algorithm. This method extends the idea of iterative support detection (ISD) from the single vector estimation to the multi-vector estimation by considering the joint sparsity prior. We provide some preliminary theoretical analysis including convergence analysis and a sufficient recovery condition. Numerical experiments from both compressive sensing and multi-task feature learning show the better performance of the proposed method in comparison with several state-of-the-art alternatives. Moreover, we demonstrate that the extension of ISD from the single vector to multi-vector estimation is not trivial. While ISD does not well reconstruct the single channel sparse Bernoulli signal, it does achieve significantly improved performance when recovering the multi-channel sparse Bernoulli signal thanks to its ability of natural incorporation of the joint sparsity structure.

1. Introduction and contributions

In the last decade, sparsity made it possible for us to reconstruct the high dimensional data with just few samples or measurements. The key of the sparse estimation problem is to stress the identification of the support, which denotes the indices of the nonzeros. If the support is known, the estimation of the sparse vectors reduces to a standard overdetermined linear inverse problem [24].

In order to enhance the estimation, many recent studies tend to consider the structure information of the solutions. For example, group sparsity structure [8,14] widely appears in many applications [31,32,35,43], where the components of solutions are likely to be either all zero or all nonzero in a group. By employing the grouping prior, ones aim to decrease the dispersion to facilitate recovering a much better solution. Here, we focus on joint sparsity, which is a special case of the group sparsity. Joint sparsity means that multiple unknown sparse vectors share a common unknown nonzero support set. Unlike the many group sparsity situations where the grouping information is unknown, the joint sparsity provides us the group information. In the following section, we will introduce the background of joint sparsity via two important applications, i.e., compressive sensing and multi-task feature learning.

* Corresponding author at: School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, Sichuan 611731 China.
E-mail addresses: yarufanfan@163.com (Y.-R. Fan), yilun.wang@rice.edu (Y. Wang), tingzhuhuang@126.com (T.-Z. Huang).

http://dx.doi.org/10.1016/j.ins.2017.06.034
0020-0255/

In compressive sensing, joint sparsity aims to reconstruct unknown signals from m measurement vectors based on a common measurement matrix. This is also called the multiple measurement vectors (MMV) problem [7,9,36]. Given the observation vectors $b_j \in \mathbb{R}^m$ $(j = 1, \ldots, \ell)$ and a measurement matrix $A \in \mathbb{R}^{m\times n}$, we want to recovery the signal $x_j \in \mathbb{R}^n$ from the noisy underdetermined systems $b_j = Ax_j + e_j$, where $e_j \in \mathbb{R}^m$ is the noise. All the signal vectors $x_1, \ldots, x_\ell$ share the sparsity pattern M, which implies the nonzero entries of $x_1, \ldots, x_\ell$ almost appear on the same position. A common signal recovery model is

$$\min_{x_j} \quad |M| \quad s.t. \quad b_j = Ax_j + e_j, \quad j = 1, \ldots, \ell, \tag{1}$$

where $|M|$ is the cardinality of M [17]. In theory, we can recover the signals $X = [x_1, \ldots, x_\ell]$ with rank $rank(X) = K$ if and only if

$$|M| < \frac{spark(A) - 1 + K}{2}, \tag{2}$$

where $spark(A)$ is the smallest set of linearly dependent columns of A [33]. Since problem (1) is NP-hard, it is usually relaxed with a convex alternative which is computationally efficient at the cost of more required measurements. Like ℓ_1-norm being the convex relaxation of ℓ_0-norm [2], the $\ell_{2,1}$-norm is widely used as the convex replacement of $|M|$ as below:

$$\min_{X} \quad ||X||_{2,1} := \sum_{i=1}^{n} ||x^i||_2 \tag{3}$$
$$s.t. \quad b_j = Ax_j + e_j, j = 1, \ldots, \ell,$$

where $x^i \in \mathbb{R}^\ell$ and $x_j \in \mathbb{R}^n$ denote the i-th row and the jth column of X, respectively.

Several fast algorithms have been proposed to solve problem (3) [40] such as greedy pursuit methods, iterative shrinkage algorithm [1] and alternating direction method (ADM) [26]. The greedy pursuit methods such as matching pursuit and orthogonal matching pursuit (OMP) [36] tend to require fewer computations but at the expense of slightly more measurements.

Multi-task learning has attracted much attention in machine learning [13,39]. It aims to learn the shared information among related tasks in order for the improved performance than considering each learning task individually. Recently, multi-task feature learning based on the $\ell_{2,1}$-norm regularization has been studied. An underlying property of the $\ell_{2,1}$-norm regularization is that it urges multiple features from different tasks to share similar sparsity patterns [38]. Given ℓ learning tasks associated with training data $\{(A^1, b_1), \ldots, (A^\ell, b_\ell)\}$, where $A^j \in \mathbb{R}^{m_j\times n}$ is the data matrix of the jth task with each row as a sample and each column as a feature; $b_j \in \mathbb{R}^{m_j}$ is the response of the jth task with biases e_j; n is the number of features; and m_j is the number of samples for the jth task, we would like to learn a weight matrix (sparsity pattern) $X = [x_1, \ldots, x_\ell] \in \mathbb{R}^{n\times\ell}$ $(x_j \in \mathbb{R}^n$ consists of the weight vectors for ℓ linear predictive models $b_j = A^j x_j + e_j)$ by solving the following optimization problem:

$$\min_{X} \quad ||X||_{2,1} := \sum_{i=1}^{n} ||x^i||_2 \tag{4}$$
$$s.t. \quad b_j = A^j x_j + e_j, j = 1, \ldots, \ell,$$

where $x^i \in \mathbb{R}^\ell$ and $x_j \in \mathbb{R}^n$ denote the ith row and the jth column of X, respectively. In this situation we assume that these different tasks share the same significant features, which leads to a joint sparsity problem.

The unconstrained formula corresponding to problems (3) and (4) can be written as the following unified form

$$\min_{X} \quad L(X) + \rho||X||_{2,1}, \tag{5}$$

where $\rho > 0$ is the regularization parameter, and $L(X)$ is a smooth convex loss function such as the least square loss function or the logistic loss function. For example, $L(X) = \sum_{j=1}^{\ell} \|A\mathbf{x}_j - \mathbf{b}_j\|_2^2$ for problem (3), and $L(X) = \sum_{j=1}^{\ell} \frac{1}{\ell m_j} \| A^j\mathbf{x}_j - \mathbf{b}_j\|_2^2$ for problem (4).

While the convexity of $\ell_{2,1}$-norm regularization provides computational efficiency, it also gives rise to the inherited bias issue. Similar with the ℓ_1-norm regularized model which only achieves suboptimal recovery performance compared with the original cardinality regularized model from the theoretical viewpoint [6], $\ell_{2,1}$-norm regularized model also only achieves suboptimal performance compared with the cardinality based model (1).

Recently, several computational advances have considered the non-convex sparse regularization since its performance is better than that of the convex sparse regularization [11,12]. For instance, for the single vector recovery, the non-convex ℓ_p-norm $(0 < p < 1)$ based sparsity regularization usually obtains better performance than ℓ_1-norm based sparsity regularization [5,18,21]. For the joint sparsity, the $\ell_{q,p}$-norm is applied in a similar way, where $0 < p < 1$ and $q \geq 1$. The non-convex sparse regularization needs less strict recovery requirements and usually achieves a better performance than the convex alternatives. While there have existed many algorithms for solving the non-convex sparse regularized models, it is still a challenging problem to obtain the global optimal solution efficiently. The behavior of a local solution is hard to analyze and more seriously structural information of the solution is also hard to be incorporated into these algorithms.