CSC 576: Variants of Sparse Learning

Ji Liu
Department of Computer Science, University of Rochester

October 27, 2015

1 Introduction

Our previous note basically suggests using ℓ_1 norm to enforce sparsity in vectors. In this note, we introduce several variants of sparse learning including Dictionary sparsity, Group sparsity, Rank sparsity, and Robust PCA.

Notation

Let us define some important notations used in this note:

- The ℓ_0 norm is defined as $\|x\|_0 :=$ the number of nonzero elements in x. For example,

  $$\|[1\ 0\ 2\ 1\ 0]^T\|_0 = 3.$$

  Please note that the ℓ_0 norm is not a real "norm", because it does not satisfy the definition of norms;
- $x \in \mathbb{R}^p$ is a p dimension vector;
- $x^* \in \mathbb{R}^p$ is a p dimension sparse vector;
- $\text{supp}(x^*)$ returns a set including indexes of nonzeros in x^*;
- $A \in \mathbb{R}^{n \times p}$ is a $n \times p$ matrix;
- $b \in \mathbb{R}^n$ is constructed from $b = Ax^* + \epsilon$ where ϵ_i's are i.i.d. Gaussian noise;
- $X \in \mathbb{R}^{m \times n}$ is a $m \times n$ matrix.

2 Dictionary Sparsity [Liu et al., 2013b]

In many cases, the signal (or the variable) is not sparse, but sparse under a linear transformation. For example, a stepwise signal

$$x = [1, 1, 1, 1, 3, 3, 3, 4, 4, 4, 4, 2, 2, 2, 2]$$

is not sparse. However, if apply a linear transformation (called dictionary matrix D)

$$D = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 \\ & & \cdots & & & \\ 0 & 0 & \cdots & 0 & 1 & -1 \end{bmatrix}$$

to x, we obtain

$$Dx = [0, 0, 0, -2, 0, 0, -1, 0, 0, 0, 2, 0, 0, 0],$$

which is sparse. Thus, a natural idea is to enforce the sparsity on Dx if we know the dictionary matrix D. A popular formulation recover x^* from $b = Ax^* + \epsilon$ is

$$\hat{x} := \underset{x}{\operatorname{argmin}} \quad \frac{1}{2}\|Ax - b\|^2 + \lambda\|Dx\|_1, \tag{1}$$

where λ is the user defined positive penalty parameter and the loss function is chosen as the least square loss, which certainly could be some other loss functions in different scenarios. Under some conditions, we still have a similar property for the estimate error as the Lasso estimator:

$$\|\hat{x} - x^*\| \leq O\left(\sqrt{\frac{\|x^*\|_0 \log p}{n}}\right)$$

where n is the number of samples and p is the dimension of the variable.

3 Group Sparsity [Huang and Zhang, 2010]

Define some groups $g_1, g_2, \cdots, g_m$. Each $g_i (i = 1, \cdots, m)$ is a subset of $\{1, 2, \cdots, p\}$. x_{g_i} is sub vector taking elements of x with indices in g_i. If we know that the nonzeros in x^* only appear in a few number of groups, it is natural to enforce the sparsity in the group level:

$$\|x\|_{gs} := \sum_{g \in \mathcal{G}} \|x_g\| = \left\| \begin{bmatrix} \|x_{g_1}\| \\ \|x_{g_1}\| \\ \cdots \\ \|x_{g_m}\| \end{bmatrix} \right\|_1 .$$

This term can be combined with any loss functions. Say, for the least squares loss, we obtain

$$\hat{x} := \underset{x}{\operatorname{argmin}} \quad \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_{gs}. \tag{2}$$

Under some conditions, the estimate error of (2) is bounded by

$$\|\hat{x} - x^*\| \leq O\left(\sqrt{\frac{|\cup_{g \in \mathcal{G}^*} g| + |\mathcal{G}^*| \log |\mathcal{G}|}{n}}\right),$$

where $\mathcal{G}^*$ is a superset satisfying $\text{supp}(x^*) \subset \cup_{g \in \mathcal{G}^*} g$.

4 Matrix Rank Sparsity [Recht, 2011]

An important extension of sparse learning is the rank sparsity. How to enforce a low rank structure on a matrix? Apparently, we can penalize the rank of the matrix. However, the function $\text{Rank}(\cdot)$ is a nonsmooth function. It usually leads to a NP-hard problem. Can we find some convex approximation to the rank function like what we did in the ℓ_0 norm function? Yes!

Let $X = U\Sigma V^T$ be the compact SVD of X. We know that $\text{Rank}(X) = \text{Rank}(\Sigma) = \|\text{diag}(\Sigma)\|_0$. Recall that in the ℓ_0 norm case, we use the ℓ_1 norm to approximate the ℓ_0 norm. Therefore, it is a natural idea to use $\|\text{diag}(\Sigma)\|_1$ to approximate $\|\text{diag}(\Sigma)\|_0$ (or $\text{Rank}(X)$). Recall the definition of the nuclear norm $\|X\|_*$. Indeed, it is exactly equivalent to $\|\text{diag}(\Sigma)\|_1$. Due to the convexity of the nuclear norm, it is much easier to solve a nuclear norm minimization problem than directly minimizing the rank function $\text{Rank}(X)$. The nuclear norm regularization has been widely used (for million times) in recent few years. Here, we only mention the most famous application in matrix completion. Let $M \in \mathbb{R}^{m\times n}$ be a low matrix with many missing elements. Ω denotes the index set of observed elements. We minimize the rank of M (or approximately $\|X\|_*$) to estimate the missing elements in M.

$$\hat{X} := \underset{X}{\text{argmin}} \quad \|X\|_* \quad \text{s.t.} \quad X_{ij} = M_{ij} \quad \forall (i,j) \in \Omega. \tag{3}$$

An very interesting theoretical result suggests that under certain conditions, if the number of observed elements in M is large enough, particularly,

$$|\Omega| \geq O\left(\text{Rank}(M)(m+n)\log^2 \max\{m,n\}\right),$$

the estimate provided by (3) is exactly equivalent to the original M with high probability, that is,

$$\hat{X} = M.$$

Roughly say, we only need $O(\text{Rank}(M)n\log^2 n)$ observations to recover the whole matrix M (assuming $m = O(n)$). If the rank of M is low, then the required number of elements to perfectly recover M can be much less than the total number of elements in M.

If the observed elements contain noise, then it is reasonable to formulate it into

$$\min_X \quad \frac{1}{2}\|(X-M)_\Omega\|_F^2 + \lambda\|X\|_*, \tag{4}$$

which is still a convex problem. If we consider the factorization of X by $X = HW^\top$, then $\|X\|_*$ can be written as

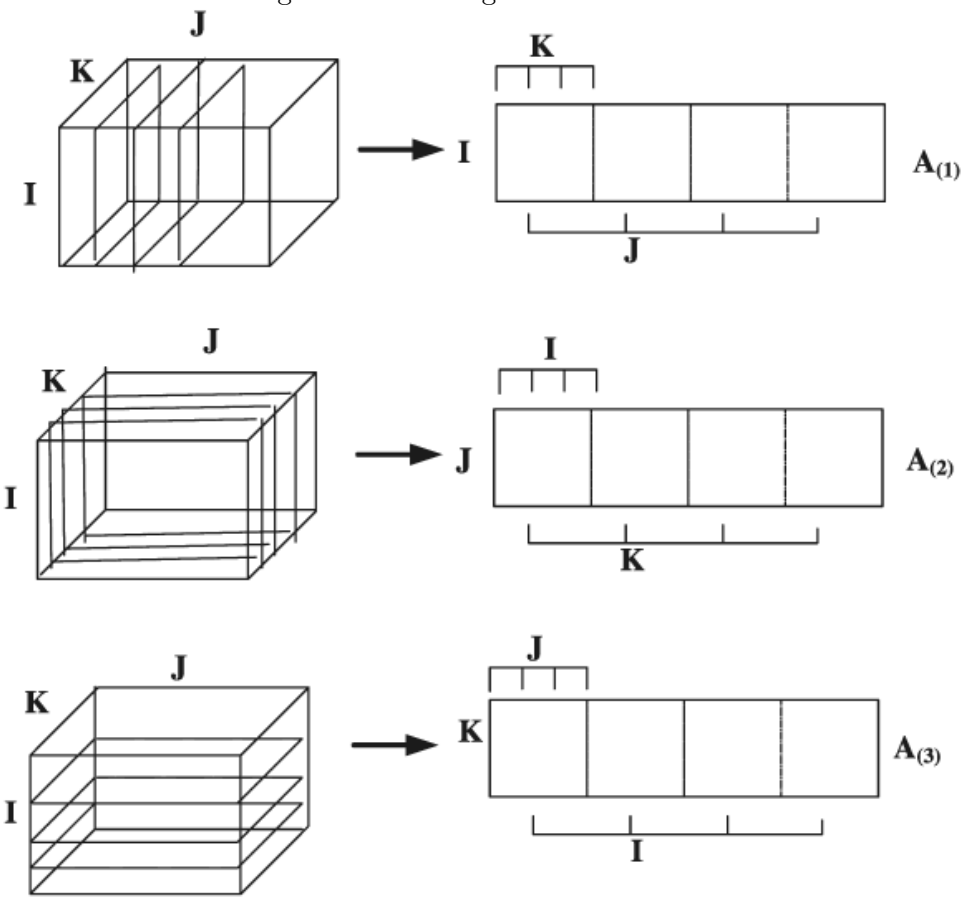
$$\|X\|_* := \min_{H,W} \ \frac{1}{2}\|H\|_F^2 + \frac{1}{2}\|W\|_F^2.$$

(Think about why it is true.) Applying the alternative representation of $\|X\|_*$, Eq. (4) can be rewritten as

$$\|X\|_* := \min_{H,W} \quad \frac{1}{2}\|(X-M)_\Omega\|_F^2 + \frac{\lambda}{2}\|H\|_F^2 + \frac{\lambda}{2}\|W\|_F^2$$

Recall when we introduce the matrix completion problem for the first time, we consider the F-norm regularization term for both factors but do not explain why. Now the reason is clear.

Figure 1: unfolding a 3-Mode tensor



5 Tensor Rank Sparsity [Liu et al., 2013a]

Tensors are natural extension of matrices, but much more complicated than matrices. Many important properties would not hold for the general tensor, for example, calculate the tensor rank is NP hard, while it is quite easy for matrices.

Many practical data is a tensor, for example, color image, video, MRI image, and hyper spectral image. How to enforce the low rank structure on tensors? The following is a commonly used norm – tensor nuclear norm:

$$\|\mathcal{X}\|_{t*} = \frac{1}{N}\sum_{n=1}^{N}\|\mathcal{X}_{(n)}\|_*$$

where $\mathcal{X}_{(n)}$ is a matrix which is the unfolding of tensor $\mathcal{X}$ along the $n-$th dimension. The tensor nuclear norm can be used for many purposes. One example is in tensor completion:

$$\min_{\mathcal{X}} \quad \|\mathcal{X}\|_{t*} \quad \text{s.t.} \quad \mathcal{X}_\Omega = \mathcal{T}_\Omega.$$

A video showing the application of nuclear norm minimization can be found in `https://www.youtube.com/watch?v=kbnmXM3uZFA`.

6 Robust PCA [Candès et al., 2009]

Let M be a matrix, the foreground F is sparse and the background B is of low rank. The goal is to separate the foreground and background from M. The problem is formulated into

$$\begin{aligned} \min_{B,F} \quad & \text{Rank}(B) + \lambda \|F\|_0 \\ \text{s.t.} \quad & B + F = M, \end{aligned} \tag{5}$$

where the matrix ℓ_0 norm of F is defined the number of nonzeros in matrix F. To make the original problem is tractable, it is natural to apply $\|B\|_*$ and $\|F\|_1$ to approximate $\text{Rank}(B)$ and $\lambda\|F\|_0$ respectively

$$\begin{aligned} \min_{B,F} \quad & \|B\|_* + \lambda \|F\|_1 \\ \text{s.t.} \quad & B + F = M. \end{aligned} \tag{6}$$

A video showing the application of Robust PCA can be found in `https://www.youtube.com/watch?v=YpdCvbJI2eg`.

References

E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *CoRR*, abs/0912.3599, 2009. URL `http://arxiv.org/abs/0912.3599`.

J. Huang and T. Zhang. The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978–2004, 2010.

J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):208–220, 2013a.

J. Liu, L. Yuan, and J. Ye. Dictionary lasso: Guaranteed sparse recovery under linear transformation. *ICML*, 2013b.

B. Recht. A simpler approach to matrix completion. *The Journal of Machine Learning Research*, 12:3413–3430, 2011.