CONVEX OPTIMIZATION FOR TENSOR DECOMPOSITION

Ryota Tomioka[1]

[1]Department of Mathematical Informatics, The University of Tokyo,
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, JAPAN, tomioka@mist.i.u-tokyo.ac.jp

ABSTRACT

Tensor is a natural extension of matrix and it appears widely in many application areas. Conventionally tensor decomposition has been tackled through non-convex optimization leaving the optimality of the solution unclear. In this talk, we introduce a class of structural regularization terms that extends nuclear norm and enables estimation of low-rank tensors through convex optimization problems. We will talk about different formulations, algorithms, and performance guarantees for them. Furthermore, We will discuss the limitations of the current approach and possible solutions.

Joint work with Taiji Suzuki, Kohei Hayashi, and Hisashi Kashima.

1. INTRODUCTION

Brain signals (EEG, fMRI), climate data, and other multivariate spatio-temporal signals can be naturally regarded as tensors (or multi-dimensional array). Tensors arises naturally in modelling relational data (e.g., collaborative filtering) when contextual or temporal dimensions are relevant.

Tensor decomposition techniques are now widely used to uncover hidden components, improve interpretability, impute missing values, and remove noise [1]. The two major approaches for tensor decomposition are CP (CANDECOMP/PARAFAC) decomposition and Tucker decomposition. However, these techniques are non-convex in nature and estimation performance guarantee for these types of approaches have been widely open (the matrix case has been analyzed in [2] and an approximation error guarantee for a variant of Tucker decomposition has been shown in [3]).

In this talk, we review two convex optimization based approaches for tensor decomposition and discuss performance guarantees we have obtained for them.

2. NOTATIONS

Let $\mathcal{W} \in \mathbb{R}^{n_1 \times \cdots n_K}$ be a K-way tensor. We denote the total number of entries in $\mathcal{W}$ by $N = \prod_{k=1}^{K} n_k$. The dot product between two tensors $\mathcal{W}$ and $\mathcal{X}$ is defined as $\langle \mathcal{W}, \mathcal{X} \rangle = \mathrm{vec}(\mathcal{W})^\top \mathrm{vec}(\mathcal{X})$; i.e., the dot product as vectors in $\mathbb{R}^N$. The Frobenius norm of a tensor is defined as $\left\|\mathcal{W}\right\|_F = \sqrt{\langle \mathcal{W}, \mathcal{W} \rangle}$. Each dimensionality of a tensor is called a *mode*. The mode k *unfolding* $\boldsymbol{W}_{(k)} \in \mathbb{R}^{n_k \times N/n_k}$ is a matrix that is obtained by concatenating the mode-k fibers along columns; here a mode-k fiber is an n_k dimensional vector obtained by fixing all but the kth index of $\mathcal{W}$. The mode-k rank r_k of $\mathcal{W}$ is the rank of the mode-k unfolding $\boldsymbol{W}_{(k)}$. We say that a tensor $\mathcal{W}$ has multilinear rank $(r_1, \ldots, r_K)$ if the mode-k rank is r_k for $k = 1, \ldots, K$ [1].

3. CONVEX TENSOR DECOMPOSITION

In this section we review two recently proposed tensor decomposition algorithms based on convex optimization [4, 5, 6, 7].

3.1. "Overlapped" regularization

Let $\mathcal{W}^*$ be the true low-rank tensor with multilinear rank $(r_1, \ldots, r_K)$. Let us assume that we are given M linear observations

$$y_i = \langle \mathcal{X}_i, \mathcal{W}^* \rangle + \epsilon_i \quad (i = 1, \ldots, M), \tag{1}$$

where the noise ϵ_i is an independent Gaussian random variable with mean zero and variance σ^2.

Since the multilinear rank is based on the matrix rank of the unfoldings, it is natural to regularize the unfoldings to be low-rank as follows [4, 5, 6, 7]

$$\underset{\mathcal{W}}{\text{minimize}} \quad \frac{1}{2M}\|\boldsymbol{y} - \mathfrak{X}(\mathcal{W})\|_2^2 + \lambda_M \sum_{k=1}^{K} \|\boldsymbol{W}_{(k)}\|_{S_1}, \tag{2}$$

where $\boldsymbol{y} = (y_1, \ldots, y_M)^\top$, $\mathfrak{X} : \mathbb{R}^N \to \mathbb{R}^M$ is a linear operator such that $\mathfrak{X}(\mathcal{W}) = (\langle \mathcal{X}_1, \mathcal{W} \rangle, \ldots, \langle \mathcal{X}_M, \mathcal{W} \rangle)^\top$, $\|\cdot\|_{S_1}$ is the Schatten 1-norm (also known as the trace norm and the nuclear norm) defined as the sum of singular values, and λ_M is a regularization constant.

We call the above regularization term *overlapped Schatten 1-norm*, because the Schatten 1-norm is applied to unfoldings of the same tensor along different modes; note that an unfolding is a permutation and thus a linear operation.

The above minimization problem is convex, and can be minimized by the alternating direction method of multipliers (ADMM); see [7] for details.

3.2. "Latent" regularization

The problem with the above overlapped approach is that regularizing every mode to be jointly low rank could be too strong an assumption.

To this end, we assume that the true tensor $\mathcal{W}^*$ is a mixture of tensors that each are low rank in a specific mode as follows:

$$\mathcal{W}^* = \sum_{k=1}^{K} \mathcal{W}^{*(k)},$$

where $\mathcal{W}^{*(k)}$ is assumed to be low-rank in the kth mode.

The resulting minimization problem can be written as follows:

$$\underset{\mathcal{W}^{(1)},\ldots,\mathcal{W}^{(K)}}{\text{minimize}} \frac{1}{2M} \left\| \boldsymbol{y} - \mathfrak{X}\left(\sum_k \mathcal{W}^{(k)}\right)\right\|_2^2 + \lambda_M \sum_{k=1}^{K} \|\boldsymbol{W}_{(k)}^{(k)}\|_{S_1}. \tag{3}$$

Now since each component $\mathcal{W}^{(k)}$ is independently regularized, it can have arbitrarily high rank for mode $k' \neq k$. Moreover, due to the fact that the linear sum of the norms are regularized (as in *group lasso*), the solution tends to consist of small number (or only one) of non-zero $\mathcal{W}^{(k)}$, which corresponds to identifying the mode with the minimal rank.

4. THEORETICAL BOUNDS

4.1. Bound for "overlapped" regularization

Let the entries of $\mathcal{X}_i$ be drawn independently and identically from the standard Gaussian distribution (random Gaussian design). Then there are constants c_1 and c_2 such that for a sample size $M \geq c_1 \left(\frac{1}{K}\sum_{k=1}^{K}(\sqrt{n_k} + \sqrt{N/n_k})\right)^2 \left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{r_k}\right)^2$, any solution $\hat{\mathcal{W}}$ of the minimization problem (2) with regularization constant $\lambda_M = 2\sigma\sum_{k=1}^{K}\left(\sqrt{n_k} + \sqrt{N/n_k}\right)/(K\sqrt{M})$ satisfies the following bound (see [8] for details):

$$\frac{1}{N}\left\|\hat{\mathcal{W}} - \mathcal{W}^*\right\|_F^2 \leq c_2 \frac{\sigma^2 \|\boldsymbol{n}^{-1}\|_{1/2}\|\boldsymbol{r}\|_{1/2}}{M}$$

where we define $\|\boldsymbol{n}^{-1}\|_{1/2} := \left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{1/n_k}\right)^2$ and $\|\boldsymbol{r}\|_{1/2} := \left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{r_k}\right)^2$ and assume $n_k \ll N/n_k$ to simplify the bound.

We call the quantity $\|\boldsymbol{n}^{-1}\|_{1/2}\|\boldsymbol{r}\|_{1/2}$ *normalized rank*, because it equals r/n when all the modes have the same dimension n and rank r.

Note that the condition for the sample size M does not involve the noise variance σ^2, whereas that for the regularization constant λ_M and the upper-bound does. Thus we can theoretically predict that when the noise is (close to) zero, the estimation error drops to nearly zero as soon as the sample size condition (M/N greater than the normalized rank) becomes valid. Figure 1 empirically shows that this is indeed the case.

4.2. Bound for "Latent" regularization

For the analysis, let us consider the following simpler generative model

$$\mathcal{Y} = \sum_{k=1}^{K} \mathcal{W}^{*(k)} + \mathcal{E},$$

where $\mathcal{Y}$ is the observed tensor and $\mathcal{E}$ is the noise. This model corresponds to the case $\mathfrak{X}$ is identity in (1).

In addition, we assume that the truth $\mathcal{W}^{*(k)}$ ($k = 1,\ldots,K$) satisfies the following "incoherence" assumption

$$\|\boldsymbol{W}_{(k)}^{*(l)}\|_{S_\infty} \leq \alpha \quad (\forall l \neq k,\, k, l = 1,\ldots,K) \tag{4}$$

Then there are universal constants c_0 and c_1, such that any solution of the minimization problem

$$\underset{\mathcal{W}^{(1)},\ldots,\mathcal{W}^{(K)}}{\text{minimize}} \quad \frac{1}{2}\left\|\mathcal{Y} - \textstyle\sum_k \mathcal{W}^{(k)}\right\|_F^2 + \lambda \sum_{k=1}^{K} \|\boldsymbol{W}_{(k)}^{(k)}\|_{S_1},$$
$$\text{s.t.} \quad \|\boldsymbol{W}_{(k)}^{(l)}\|_{S_\infty} \leq \alpha \quad (\forall l \neq k),$$

with regularization constant $\lambda = c_0\sigma\left(\sqrt{N/\min_k n_k} + \sqrt{\max_k n_k} + \sqrt{\log K}\right) + \alpha(K-1)$ satisfies the following bound:

$$\frac{1}{N}\left\|\hat{\mathcal{W}} - \mathcal{W}^{*(k)}\right\|_F^2 \leq c_1 K F \sigma^2 \frac{\min_k r_k}{\min_k n_k}, \tag{5}$$

where $\hat{\mathcal{W}} := \sum_k \hat{\mathcal{W}}^{(k)}$ and $F = \left(\left(1 + \sqrt{\frac{n_1 n_K}{N}}\right) + \left(\sqrt{\log K} + \frac{\alpha(K-1)}{c_0\sigma}\right)\sqrt{\frac{n_K}{N}}\right)^2$ is a factor that mildly depends on the dimensionalities; see [9] for details.

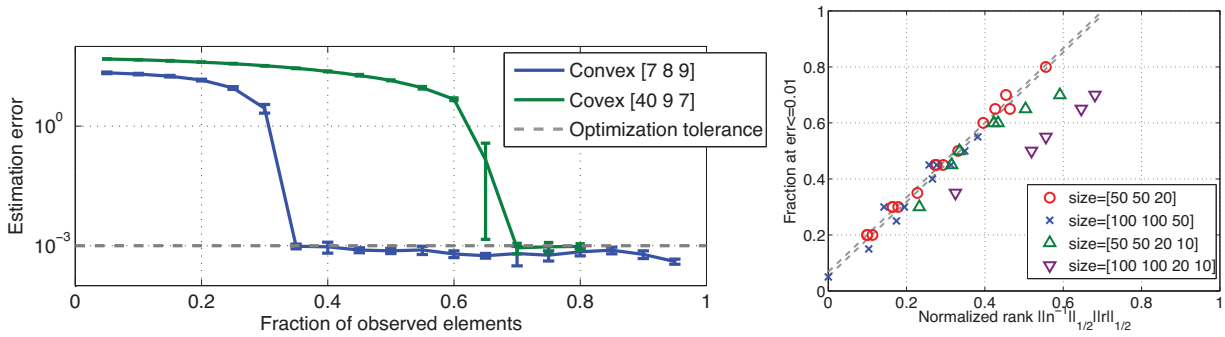
4.3. Comparison between the two approaches

In the same setting in which all entries are observed with noise, we have the following bound for the "overlapped" approach

$$\frac{1}{N}\left\|\hat{\mathcal{W}} - \mathcal{W}^*\right\|_F^2 \leq c_1'\sigma^2\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\tfrac{1}{n_k}}\right)^2\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{r_k}\right)^2. \tag{6}$$

Comparing inequalities (5) and (6), we notice that the complexity of the overlapped approach depends on the average (square root) of the multilinear rank $r_1,\ldots,r_K$, whereas that of the latent approach only grows linearly against the *minimum* multilinear rank. Interestingly, the latent approach performs *as if it knows the mode with the minimum rank*, although such information is not available to it.

This is empirically confirmed in Figure 2. Figure 2 compares the two approaches for recovering a $r \times r \times 3$ tensor of size $50\times50\times20$ from noisy measurements. The estimation errors of the two approaches are plotted against the rank of the first two modes. The error of the overlapped approach grows continuously as the rank of the



(a) Phase transition curves for size $50\times50\times20$ tensor with two different multilinear ranks. (b) M/N at the phase transition vs. normalized rank.

Figure 1. Phase transition occurs for the "overlapped" Schatten 1-norm regularization for tensor completion when certain fraction M/N of the entries are observed without noise. After the phase transition, the proposed method recovers the true tensor almost exactly (including the multilinear rank); see [8] for details.

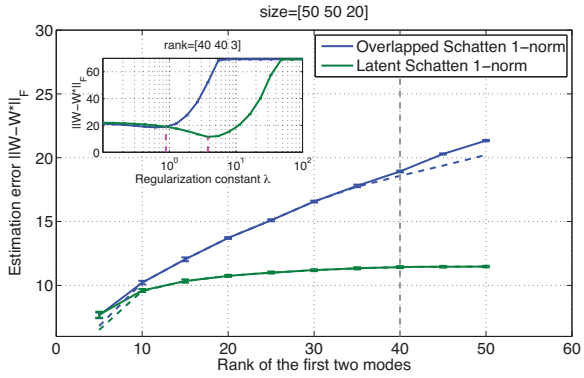


Figure 2. Comparison of the "overlapped" and "latent" Schatten 1-norm regularizations for noisy tensor decomposition. When the underlying tensor is almost full rank in all but one mode, the "latent" approach performs better than the "overlapped" approach because it only requires that the tensor can be represented as a mixture of low-rank tensors, whereas the "overlapped" approach assumes that the underlying tensor is simultaneously low-rank.

first two modes increase, whereas that of the latent approach stays almost constant. This can be explained by the fact that the bound (5) depends on the minimum multilinear rank; it can exploit the low-rank-ness of the last mode even when the first two modes are almost full rank.

5. DISCUSSION

We have analyzed the performance of two recently proposed convex optimization based tensor decomposition algorithms.

The analysis so far is quite preliminary. For example, the sample complexity implied by the above bounds is $O(rn^{K-1})$, which becomes prohibitive when K is large. One possible approach (personally suggested by Nam H. Nguyen and recently carried out in [10]) would be to matricize the tensor evenly; when K is even, this approach would lead to sample complexity $O(r^{K/2}n^{K/2})$ at the cost of computing the SVD of $n^{K/2}\times n^{K/2}$ matrices. Therefore it would be interesting to study the trade-off between the theoretical guarantee and the computational complexity. This is also related to the sequential convex relaxation of tensor rank suggested in [11].

Another interesting direction would be to combine the convex optimization based algorithms for tensors with spectral methods for estimating latent variable models; see [12].

6. ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI 22700138, 23240019, and 25870192.

7. REFERENCES

[1] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.

[2] K. Fukumizu, "Generalization error of linear neural networks in unidentifiable cases," in *Algorithmic Learning Theory*, pp. 51–62. Springer, 1999.

[3] C. Ding, H. Huang, and D. Luo, "Tensor reduction error analysis: Applications to video compression and classification," in *Proc. IEEE CVPR 2008*. IEEE, 2008, pp. 1–8.

[4] M. Signoretto, L. De Lathauwer, and J.A.K. Suykens, "Nuclear norms for tensors and their use for convex multilinear estimation," Tech. Rep. 10-186, ESAT-SISTA, K.U.Leuven, 2010.

[5] S. Gandy, B. Recht, and I. Yamada, "Tensor completion and low-n-rank tensor recovery via convex optimization," *Inverse Problems*, vol. 27, pp. 025010, 2011.

[6] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," in *Prof. ICCV*, 2009.

[7] R. Tomioka, K. Hayashi, and H. Kashima, "Estimation of low-rank tensors via convex optimization," Tech. Rep., arXiv:1010.0789, 2011.

[8] R. Tomioka, T. Suzuki, K. Hayashi, and H. Kashima, "Statistical performance of convex tensor decomposition," in *Advances in NIPS 24*, pp. 972–980. 2011.

[9] R. Tomioka and T. Suzuki, "Convex tensor decomposition via structured schatten norm regularization," Tech. Rep., arXiv:1303.6370, 2013.

[10] C. Mu, B. Huang, J. Wright, and D. Goldfarb, "Square deal: Lower bounds and improved relaxations for tensor recovery," Tech. Rep., arXiv:1307.5870, 2013.

[11] V. Chandrasekaran, B. Recht, PA Parrilo, and A. Willsky, "The convex geometry of linear inverse problems, prepint," Tech. Rep., arXiv:1012.0621v2, 2010.

[12] A. T. Chaganty and P. Liang, "Spectral experts for estimating mixtures of linear regressions," in *Proc. ICML 2013*. 2013.