数理解析研究所講究録
第2027巻 2017年 144-149

リーマン計量調整に基づくTucker多様体の幾何の提案と最適化問題への応用

電気通信大学・大学院情報理工学研究科情報ネットワーク工学専攻　笠井 裕之
Amazon Development Centre India, Bamdev Mishra

Hiroyuki Kasai
Department of Computer and Network Engineering,
The University of Electro-Communications

Bamdev Mishra
Amazon Development Centre India

概 要

本稿では，低ランク・テンソルTucker分解のための新しい幾何空間“Scaled Tucker Manifold”による“テンソル補完問題”の効率的な手法を提案した論文[1]の概要を記す．提案手法は，一般的なテンソル回帰問題に対して，Scaled Tucker Manifoldにより効率的な解法を確立することが可能となる．Scaled Tucker Manifolの導出にあたっては，Tucker分解の対称構造と回帰問題の最小自乗構造に着目した新しいリーマン計量を提案し，幾何空間を定義する数々の構成要素を導出している．

1 Introduction

This paper addresses the problem of low-rank tensor completion when the rank is a priori known or estimated. Without loss of generality, we focus on 3-order tensors. Given a tensor $\boldsymbol{\mathcal{X}}^{n_1\times n_2\times n_3}$, whose entries $\boldsymbol{\mathcal{X}}^{\star}_{i_1,i_2,i_3}$ are only known for some indices $(i_1, i_2, i_3) \in \Omega$, where Ω is a subset of the complete set of indices $\{(i_1, i_2, i_3) : i_d \in \{1, \ldots, n_d\}, d \in \{1, 2, 3\}\}$, the *fixed-rank tensor completion problem* is formulated as

$$\begin{array}{ll}\displaystyle\min_{\boldsymbol{\mathcal{X}}\in\mathbb{R}^{n_1\times n_2\times n_3}} & \displaystyle\frac{1}{|\Omega|}\|\boldsymbol{\mathcal{P}}_{\Omega}(\boldsymbol{\mathcal{X}}) - \boldsymbol{\mathcal{P}}_{\Omega}(\boldsymbol{\mathcal{X}}^{\star})\|_F^2 \\ \text{subject to} & \mathrm{rank}(\boldsymbol{\mathcal{X}}) = \mathbf{r},\end{array}$$

where the operator $\boldsymbol{\mathcal{P}}_{\Omega}(\boldsymbol{\mathcal{X}})_{i_1i_2i_3} = \boldsymbol{\mathcal{X}}_{i_1i_2i_3}$ if $(i_1, i_2, i_3) \in \Omega$ and $\boldsymbol{\mathcal{P}}_{\Omega}(\boldsymbol{\mathcal{X}})_{i_1i_2i_3} = 0$ otherwise and (with a slight abuse of notation) $\|\cdot\|_F$ is the Frobenius norm. $\mathrm{rank}(\boldsymbol{\mathcal{X}})$ $(= \mathbf{r} = (r_1, r_2, r_3))$, called the *multilinear rank* of $\boldsymbol{\mathcal{X}}$, is the set of the ranks of for each of mode-d unfolding matrices. $r_d \ll n_d$ enforces a low-rank structure. The *mode* is a matrix obtained by concatenating the mode-d fibers along column and mode-d *unfolding* of $\boldsymbol{\mathcal{X}}$ is $\mathbf{X}_d \in \mathbb{R}^{n_d\times n_{d+1}\cdots n_D n_1\cdots n_{d-1}}$ for $d = \{1, \ldots, D\}$.

The optimization problem (1) has many variants, and one of those is extending the nuclear norm regularization approach from the matrix case [2] to the tensor case. While this generalization leads to good results [3–5], its scalabilityto large-scale instances is not trivial, especially due to the necessity of high-dimensional singular value decomposition computations. A different approach exploits *Tucker decomposition* [6, Section 4] of a low-rank tensor $\mathcal{X}$ to develop large-scale algorithms for (1), e.g., in [7, 8]. The present paper exploits both the *symmetry* present in Tucker decomposition and the *least-squares* structure of the cost function of (1) by using the concept of *preconditioning*. While preconditioning in unconstrained optimization is well studied [9, Chapter 5], preconditioning on constraints with *symmetries*, owing to non-uniqueness of Tucker decomposition [6, Section 4.3], is not straightforward. We build upon the recent work [10] that suggests to use *Riemannian preconditioning* with a *tailored metric* (inner product) in the Riemannian optimization framework on quotient manifolds [11–13]. Our proposed preconditioned nonlinear conjugate gradient algorithm is implemented in the Matlab toolbox Manopt [14] and it outperforms state-of-the-art methods. In the supplementary material section, we show concrete mathematical derivations and additional numerical comparisons. We also provide a *generic* Manopt factory (a manifold description Matlab file) with additional support for second-order implementations, e.g., the trust-region method.

2 Exploiting the problem structure

We focus on the two fundamental structures present in (1): *symmetry* in the constraints, and the *least-squares structure* of the cost function. Finally, a novel metric is proposed.

The quotient and least-squares structures. The Tucker decomposition of a tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ of rank $\mathbf{r}$ $(=(r_1, r_2, r_3))$ is [6, Section 4.1] $\mathcal{X} = \mathcal{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3$, where $\mathbf{U}_d \in \mathrm{St}(r_d, n_d)$ for $d \in \{1, 2, 3\}$ belongs to the *Stiefel manifold* of matrices of size $n_d \times r_d$ with orthogonal columns and $\mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$. Here, $\mathcal{W} \times_d \mathbf{V} \in \mathbb{R}^{n_1 \times \cdots n_{d-1} \times m \times n_{d+1} \times \cdots n_N}$ computes the *d-mode product* of a tensor $\mathcal{W} \in \mathbb{R}^{n_1 \times \cdots \times n_N}$ and a matrix $\mathbf{V} \in \mathbb{R}^{m \times n_d}$. Tucker decomposition is *not unique* as $\mathcal{X}$ remains unchanged under the transformation $(\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathcal{G}) \mapsto (\mathbf{U}_1\mathbf{O}_1, \mathbf{U}_2\mathbf{O}_2, \mathbf{U}_3\mathbf{O}_3, \mathcal{G} \times_1 \mathbf{O}_1^T \times_2 \mathbf{O}_2^T \times_3 \mathbf{O}_3^T)$ for all $\mathbf{O}_d \in \mathcal{O}(r_d)$, which is the set of orthogonal matrices of size of $r_d \times r_d$. The classical remedy to remove this indeterminacy is to have additional structures on $\mathcal{G}$ like sparsity or restricted orthogonal rotations [6, Section 4.3]. In contrast, we encode the transformation in an abstract search space of *equivalence classes*, defined as, $[(\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathcal{G})] := \{(\mathbf{U}_1\mathbf{O}_1, \mathbf{U}_2\mathbf{O}_2, \mathbf{U}_3\mathbf{O}_3, \mathcal{G} \times_1 \mathbf{O}_1^T \times_2 \mathbf{O}_2^T \times_3 \mathbf{O}_3^T) : \mathbf{O}_d \in \mathcal{O}(r_d)\}$. The set of equivalence classes is the quotient manifold [15, Theorem 9.16]

$$\mathcal{M}/\sim \; := \; \mathcal{M}/(\mathcal{O}(r_1) \times \mathcal{O}(r_2) \times \mathcal{O}(r_3)),$$

where $\mathcal{M}$ is called the *total space* (computational space) that is the product space $\mathcal{M} := \mathrm{St}(r_1, n_1) \times \mathrm{St}(r_2, n_2) \times \mathrm{St}(r_3, n_3) \times \mathbb{R}^{r_1 \times r_2 \times r_3}$. Due to the invariance of the Tucker decomposition, the local minima of (1) in $\mathcal{M}$ are not isolated, but they become isolated on $\mathcal{M}/\sim$. Consequently, the problem (1) is an optimization problem on a quotient manifold for which systematic procedures are proposed in [11–13] by endowing $\mathcal{M}/\sim$ with a Riemannian structure. We call $\mathcal{M}/\sim$ the *Tucker manifold.*

Another structure that is present in (1) is the least-squares structure of the cost function. A way to exploit it is to endow the search space with a metric (inner product) induced by the Hessian of the cost function [9]. This induced metric (or its approximation) resolves convergence issues of first-order optimization algorithms. Specifically for the case of quadratic optimization with rank constraint (matrix case), Mishra and Sepulchre [10, Section 5] propose a family of Riemannian metrics from the Hessian of the cost function. Since applying this approach directly for (1) is computationally costly, we consider a simplified cost function by assuming that Ω contains the full set of indices, i.e., we focus on $\|\mathcal{X} - \mathcal{X}^\star\|_F^2$ to propose a metric candidate. A good candidate is by considering only the *block diagonal* elements of the Hessian of $\|\mathcal{X} - \mathcal{X}^\star\|_F^2$. It should emphasized that the cost function $\|\mathcal{X} - \mathcal{X}^\star\|_F^2$ is *convex and quadratic* in $\mathcal{X}$. Consequently, it is also convex and quadratic in the arguments $(\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathcal{G})$ individually. The block diagonal approximation of the Hessian of $\|\mathcal{X} - \mathcal{X}^\star\|_F^2$ in $(\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathcal{G})$ is $((\mathbf{G}_1\mathbf{G}_1^T)\otimes \mathbf{I}_{n_1}, (\mathbf{G}_2\mathbf{G}_2^T)\otimes \mathbf{I}_{n_2}, (\mathbf{G}_3\mathbf{G}_3^T)\otimes \mathbf{I}_{n_3}, \mathbf{I}_{r_1r_2r_3})$, where $\mathbf{G}_d$ is the mode-d unfolding of $\mathcal{G}$ and is assumed to be full rank. The terms $\mathbf{G}_d\mathbf{G}_d^T$ for $d \in \{1,2,3\}$ are *positive definite* when $r_1 \leq r_2r_3$, $r_2 \leq r_1r_3$, and $r_3 \leq r_1r_2$.

A novel Riemannian metric and its motivation. An element x in the total space $\mathcal{M}$ has the matrix representation $(\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathcal{G})$. Consequently, the tangent space $T_x\mathcal{M}$ is the Cartesian product of the tangent spaces of the individual manifolds, i.e., $T_x\mathcal{M}$ has the matrix characterization [13] $T_x\mathcal{M} = \{(\mathbf{Z}_{\mathbf{U}_1}, \mathbf{Z}_{\mathbf{U}_2}, \mathbf{Z}_{\mathbf{U}_3}, \mathbf{Z}_{\mathcal{G}}) \in \mathbb{R}^{n_1\times r_1} \times \mathbb{R}^{n_2\times r_2} \times \mathbb{R}^{n_3\times r_3} \times \mathbb{R}^{r_1\times r_2\times r_3} : \mathbf{U}_d^T\mathbf{Z}_{\mathbf{U}_d} + \mathbf{Z}_{\mathbf{U}_d}^T\mathbf{U}_d = 0, \text{ for } d \in \{1,2,3\}\}$. The earlier discussion on symmetry and least-squares structure leads to the novel metric $g_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$

$$\begin{aligned} g_x(\xi_x, \eta_x) \; &= \langle \xi_{\mathbf{U}_1}, \eta_{\mathbf{U}_1}(\mathbf{G}_1\mathbf{G}_1^T)\rangle + \langle \xi_{\mathbf{U}_2}, \eta_{\mathbf{U}_2}(\mathbf{G}_2\mathbf{G}_2^T)\rangle \\ &+ \langle \xi_{\mathbf{U}_3}, \eta_{\mathbf{U}_3}(\mathbf{G}_3\mathbf{G}_3^T)\rangle + \langle \xi_{\mathcal{G}}, \eta_{\mathcal{G}}\rangle, \end{aligned}$$

where $\xi_x, \eta_x \in T_x\mathcal{M}$ are tangent vectors with matrix characterizations, $(\xi_{\mathbf{U}_1}, \xi_{\mathbf{U}_2}, \xi_{\mathbf{U}_3}, \xi_{\mathcal{G}})$ and $(\eta_{\mathbf{U}_1}, \eta_{\mathbf{U}_2}, \eta_{\mathbf{U}_3}, \eta_{\mathcal{G}})$, respectively and $\langle \cdot, \cdot \rangle$ is the Euclidean inner product. As contrasts to the classical Euclidean metric, the metric (2) *scales* the level sets of the cost function on the search space that leads a preconditioning effect on the algorithms developed on the Tucker manifold.

3 Notions of optimization on quotient manifolds

Each point on a quotient manifold represents an entire equivalence class of matrices in the total space. Abstract geometric objects on a quotient manifold call for matrix representatives in the total space. Similarly, algorithms are run in the total space $\mathcal{M}$, but under appropriate compatibility between the Riemannian structure of $\mathcal{M}$ and the Riemannian structure of the quotient manifold $\mathcal{M}/\sim$, they define algorithms on the quotient manifold. Once we endow $\mathcal{M}/\sim$ with a Riemannian structure, the constraint optimization problem (1) is conceptually transformed into an unconstrained optimization over the Riemannian quotient manifold (2). When the points x and y in $\mathcal{M}$ belong to the same equivalence class, they represent a single point $[x] := \{y \in \mathcal{M} : y \sim x\}$ on the quotient manifold $\mathcal{M}/\sim$. The abstract tangent space $T_{[x]}(\mathcal{M}/\sim)$ at $[x] \in \mathcal{M}/\sim$ has the matrix representation in $T_x\mathcal{M}$, but restricted to the directions that do not induce a displacement along the equivalence class $[x]$. This is realized by decomposing $T_x\mathcal{M}$ into two complementary subspaces. The vertical space $\mathcal{V}_x$ is the tangent space of the equivalence class $[x]$. On the other hand, the horizontal space $\mathcal{H}_x$ is the *orthogonal subspace* to $\mathcal{V}_x$, i.e., $T_x\mathcal{M} = \mathcal{V}_x \oplus \mathcal{H}_x$. The horizontal subspace provides a valid matrix representation to the abstract tangent space $T_{[x]}(\mathcal{M}/\sim)$ [11, Section 3.5.8]. An abstract tangent vector $\xi_{[x]} \in T_{[x]}(\mathcal{M}/\sim)$ at $[x]$ has a unique element $\xi_x \in \mathcal{H}_x$ that is called its *horizontal lift*. Endowed with the Riemannian metric (2), the quotient manifold $\mathcal{M}/\sim$ is a *Riemannian submersion* of $\mathcal{M}$. The submersion principle then allows to work out concrete matrix representations of abstract object on $\mathcal{M}/\sim$. Particularly, starting from an arbitrary matrix (with appropriate dimensions), two linear projections are needed: the first projection Ψ_x is onto the tangent space $T_x\mathcal{M}$, while the second projection Π_x is onto the horizontal subspace $\mathcal{H}_x$. The computation cost of these projections is $O(n_1r_1^2 + n_2r_2^2 + n_3r_3^2)$.

Finally, we propose a Riemannian nonlinear conjugate gradient algorithm for (1) that scales well to large-scale instances. Specifically, we use the conjugate gradient implementation of Manopt with the ingredients described in Table ??. The convergence analysis of this method follows from [11, 16, 17]. If $f(\mathcal{X}) = \|\mathcal{P}_\Omega(\mathcal{X}) - \mathcal{P}_\Omega(\mathcal{X}^\star)\|_F^2/|\Omega|$, then the Riemannian gradient $\mathrm{grad}_x f$, which has the matrix characterization $\Psi(\mathrm{egrad}_x f)$, where $\mathrm{egrad}_x f$ is the Euclidean gradient of f. We show a way to compute a step-size guess effectively. The total computational cost per iteration of our proposed algorithm is $O(|\Omega| r_1 r_2 r_3)$, where $|\Omega|$ is the number of known entries.

4 Numerical comparisons

We show numerical comparisons of our proposed algorithm with state-of-the-art algorithms that include TOpt [7] and geomCG [8], for comparisons with Tucker decomposition based algorithms, and HaLRTC [3], Latent [4], and Hard [5] as nuclear norm minimization algorithms. All simulations are performed in Matlab on a 2.6 GHz Intel Core i7 machine with 16 GB RAM. For specific operations with unfoldings of $\boldsymbol{\mathcal{S}}$, we use the `mex` interfaces that are provided in geomCG. For large-scale instances, our algorithm is only compared with geomCG as other algorithms cannot handle these instances. We randomly and uniformly select known entries based on a multiple of the dimension, called the *over-*

sampling (OS) ratio, to create the training set Ω. Algorithms (and problem instances) are initialized randomly, as in [8], and are stopped when either the mean square error (MSE) on the training set Ω is below 10^{-12} or the number of iterations exceeds 250. We also evaluate the mean square error on a test set Γ, which is different from Ω. Five runs are performed in each scenario.

Case 1 considers synthetic small-scale tensors of size $100 \times 100 \times 100$, $150 \times 150 \times 150$, and $200 \times 200 \times 200$ and rank $\mathbf{r} = (10, 10, 10)$ are considered. OS is $\{10, 20, 30\}$. The result shows that the convergence behavior of our proposed algorithm is either competitive or faster than the others. Next, **Case 2** considers large-scale tensors of size $3000 \times 3000 \times 3000$, $5000 \times 5000 \times 5000$, and $10000 \times 10000 \times 10000$ and ranks $\mathbf{r} = (5, 5, 5)$ and $(10, 10, 10)$. OS is 10. Our proposed algorithm outperforms geomCG. **Case 3** considers instances where the dimensions and ranks along certain modes are different than others. Two cases are considered. Case (3.a) considers tensors size $20000 \times 7000 \times 7000$, $30000 \times 6000 \times 6000$, and $40000 \times 5000 \times 5000$ with rank $\mathbf{r} = (5, 5, 5)$. Case (3.b) considers a tensor of size $10000 \times 10000 \times 10000$ with ranks $(7, 6, 6)$, $(10, 5, 5)$, and $(15, 4, 4)$. In all the cases, the proposed algorithm converges faster than geomCG. Finally, **Case 4** considers MovieLens-10M dataset that contains 10000054 ratings corresponding to 71567 users and 10681 movies. We split the time into 7-days wide bins results, and finally, get a tensor of size $71567 \times 10681 \times 731$. The fraction of known entries is less than 0.002%. We perform five random 80/10/10–train/validation/test partitions. The maximum iteration is set to 500. Our proposed algorithm consistently gives lower test errors than geomCG across different ranks.

5 Conclusion and future work

We have proposed a preconditioned nonlinear conjugate gradient algorithm for the tensor completion problem by exploiting the fundamental structures of symmetry, due to non-uniqueness of Tucker decomposition, and least-squares of the cost function. A novel Riemannian metric is proposed that enables to use the versatile Riemannian optimization framework. Numerical comparisons suggest that our proposed algorithm has a superior performance on different benchmarks.

参考文献

[1] H. Kasai and B. Mishra. Low-rank tensor completion: a Riemannian manifold preconditioning approach. In *ICML*, 2016.

[2] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[3] J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):208–220, 2013.

[4] R. Tomioka, K. Hayashi, and H. Kashima. Estimation of low-rank tensors via convex optimization. Technical report, arXiv preprint arXiv:1010.0789, 2011.

[5] M. Signoretto, Q. T. Dinh, L. D. Lathauwer, and J. A. K. Suykens. Learning with tensors: a framework based on convex optimization and spectral regularization. *Mach. Learn.*, 94(3):303–351, 2014.

[6] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Rev.*, 51(3):455–500, 2009.

[7] M. Filipović and A. Jukić. Tucker factorization with missing data with application to low-n-rank tensor completion. *Multidim. Syst. Sign. P.*, 2013.

[8] D. Kressner, M. Steinlechner, and B. Vandereycken. Low-rank tensor completion by Riemannian optimization. *BIT Numer. Math.*, 54(2):447–468, 2014.

[9] J. Nocedal and S. J. Wright. *Numerical Optimization*, volume Second Edition. Springer, 2006.

[10] B. Mishra and R. Sepulchre. Riemannian preconditioning. *SIAM J. Optim.*, 26(1):635–660, 2016.

[11] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

[12] S. T. Smith. Optimization techniques on Riemannian manifold. In A. Bloch, editor, *Hamiltonian and Gradient Flows, Algorithms and Control*, volume 3, pages 113–136. Amer. Math. Soc., Providence, RI, 1994.

[13] A. Edelman, T.A. Arias, and S.T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, 1998.

[14] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt: a Matlab toolbox for optimization on manifolds. *J. Mach. Learn. Res.*, 15(1):1455–1459, 2014.

[15] J. M. Lee. *Introduction to smooth manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 2003.

[16] H. Sato and T. Iwai. A new, globally convergent Riemannian conjugate gradient method. *Optimization*, 64(4):1011–1031, 2015.

[17] W. Ring and B. Wirth. Optimization methods on Riemannian manifolds and their application to shape space. *SIAM J. Optim.*, 22(2):596–627, 2012.