Tensor Decomposition for Signal Processing and Machine Learning

Benhammou Nouhayla, Fteri Mghari Khadija, Ait ja Abdellilah
Supervised by: Mr.El Bouanani,
University Mohammed V - ENSIAS

***Abstract*— This study gives an outline of higher-order tensor decompositions, their usage in both signal processing and machine learning.A tensor could be a multidimensional or N-way cluster. Decompositions of higher-order tensors (i.e., N-way clusters with N 3) have applications in psychometrics, chemometrics, flag preparing, numerical straight polynomial math, computer vision, numerical investigation, information mining, neuroscience, chart investigation, and somewhere else. Two particular tensor disintegrations can be considered to be higher-order expansions of the framework particular esteem decomposition:CANDECOMP/PARAFAC (CP) breaks down a tensor as a sum of rank-one tensors, and the Tucker deterioration may be a higher-order shape of principal component examination. There are numerous other tensor deteriorations, counting INDSCAL, PARAFAC2, CANDELINC, DEDICOM, and PARATUCK2 as well as nonnegative variations of all of the over. The N-way Tool kit, Tensor Tool stash, and Multilinear Motor are examples of program bundles for working with tensors.**

Key words: tensor decompositions, multiway arrays, multilinear algebra, parallel factors (PARAFAC), canonical decomposition (CANDECOMP), higher-order principal components analysis (Tucker), higher-order singular value decomposition (HOSVD)

I. INTRODUCTION

A tensor is a multidimensional cluster. More formally, an Nway or Nth-order tensor is an component of the tensor item of N vector spaces, each of which has its possess facilitate system.

Tensor algebra: The difference between signal processing and machine learning

SP analysts (and Chemists) regularly center on the columns of the figure lattices A, B, C and the associated rank-1 components denotes the external item, see area II-C), since they are interested in partition. ML analysts regularly center on the rows of A, B, C, since they think of them as parsimonious latent space representations. For a client × thing × context ratings tensor, for illustration, a push of A could be a representation of the comparing client in idle space, and likewise a push of B (C) could be a representation of the comparing thing (setting) in the same inactive space. The internal item of these three vectors is used to foresee that user's rating of the given thing in the given setting. This can be one reason why ML analysts tend to use internal (rather than external) item documentation. SP researchers are curious about demonstrate identifiability since it guarantees separability.SP analysts (and Chemists) regularly center on the columns of the figure lattices A, B, C and the associated rank-1 components of the decay, since they are interested in partition. ML analysts regularly center on the rows of A, B, C, since they think of them as parsimonious latent space representations. For a client × thing × context ratings tensor, for illustration, a push of A could be a representation of the comparing client in idle space, and likewise a push of B (C) could be a representation of the comparing thing (setting) in the same inactive space. The internal item of these three vectors is used to foresee that user's rating of the given thing in the given setting. This can be one reason why ML analysts tend to use internal (rather than external) item documentation. SP researchers are curious about demonstrate identifiability since it guarantees separability.[**1**]

II. PRELIMINARIES

Consider an I × J matrix X, and let colrank(X) := the number of linearly independent columns of X, i.e., the dimension of the range space of X, dim(range(X)). colrank(X) is the minimum k N such that $X = AB^T$, where A is an I × k basis of range(X), and BT is k × J and holds the corresponding coefficients.

Low-rank matrix approximation

In practice X is usually full-rank, e.g., due to measurement noise, and we observe X = L + N, where $L = AB^T$ is low-rank and N represents noise and 'unmodeled dynamics'. If the elements of N are sampled from a jointly continuous distribution, then N will be full rank almost surely – for the determinant of any square submatrix of N is a polynomial in the matrix entries, and a polynomial that is nonzero at one point is nonzero at every point except for a set of measure zero. In such cases, we are interested in approximating X with a low-rank matrix.

Some useful products and their properties

- The Kronecker product: The Kronecker product is an operation that changes two lattices into a bigger framework that contains all the conceivable items of the passages of the two frameworks. It has a few properties that are regularly utilized to unravel troublesome issues in direct variable based math and its applications.The Kronecker product of A (I × K) and B (J × L) is the IJ×KL matrix is:

$$\mathbf{A} \otimes \mathbf{B} := \begin{bmatrix} \mathbf{BA}(1,1) & \mathbf{BA}(1,2) & \cdots & \mathbf{BA}(1,K) \\ \mathbf{BA}(2,1) & \mathbf{BA}(2,2) & \cdots & \mathbf{BA}(2,K) \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{BA}(I,1) & \mathbf{BA}(I,2) & \cdots & \mathbf{BA}(I,K) \end{bmatrix}$$

- Khatri–Rao product: the Khatri–2 Rao (column-wise Kronecker)product of two matrices with the same number of columns the Khatri–Rao product of A and B is:

$$A \odot B = [\, a1 \otimes b1, a2 \otimes b2 \ldots\ldots, aN \otimes bN]$$

- Tensor (outer) product: The tensor product or outer product of vectors a (I × 1) and b (J × 1) is defined as the I × J matrix, it is expressed as bellow:

$$(\mathbf{a} \circledcirc \mathbf{b})(i,j) = \mathbf{a}(i)\mathbf{b}(j),\ \forall i,j$$

III. RANK AND RANK DECOMPOSITION FOR TENSORS: CPD / PARAFAC

In multilinear variable based math, the tensor rank decay or canonical polyadic deterioration (CPD) is one generalization of the network solitary esteem deterioration (SVD) to tensors, which have found application in insights, flag handling, computer vision, computer design, psychometrics, etymology and chemometrics. The tensor rank decay was presented by Straight to the point Lauren Hitchcock in 1927 and afterward rediscovered a few times, outstandingly in psychometrics.For this reason, the tensor rank decay is regularly alluded to as CANDECOMP, PARAFAC,or CANDECOMP/PARAFAC (CP). Another well known generalization of the framework SVD is known as the higher-order particular esteem deterioration.

A. CPD

The polyadic decomposition (PD) approximates a tensor with a sum of R rank-one tensors. If the number of rank-1 terms R is minimal, the decomposition is called canonical (CPD). Let $A \otimes B$, denote the outer product between an Nth-order tensor A and an Mth-order tensor B, then AB is the (N+M)th-order tensor defined by $(A \otimes B)_{(i_1 i_N j_1 j_M = a_{i1} i_N b_{j1} j_M)}$. For example, let a, b and c be nonzero vectors in Rn, then $a \otimes b \equiv a.b^T$ is a rank-one matrix and abc is defined to be a rank-one tensor. Let T be a tensor of dimensions $I_1 I_2 I_N$, and let U(n) be matrices of size In×R and u(n)r the r^{th} column of U(n), then

$$T \approx \sum r{=}1 R u(1)r \otimes u(1)r \otimes \cdots \otimes u(N)r$$

A visual representation of this decomposition in the third-order case is shown in the figure below:

B. Typical, generic, and border rank of tensors

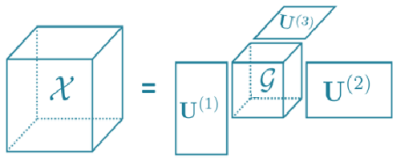
Typical rank: One can define typical ranks as the ranks that are associated with subsets of nonzero volume in the latter partition. If there is a single typical rank, then it may be called the generic rank. There may exist Euclidean-open sets of tensors of rank strictly higher than the generic rank. All ranks appearing on open sets in the Euclidean topology are called typical ranks. The smallest typical rank is called the generic rank; this definition applies to both complex and real tensors

Generic rank: We see that the rank of a tensor for decomposition over R is a random variable that can take more than one value with positive probability. These values are called typical ranks. For decomposition over C the situation is different: rank(randn(2,2,2)) = 2 with probability 1, so there is only one typical rank. When there is only one typical rank (that occurs with probability 1 then) we call it generic rank.

Border rank: A rank-s tensor A is called a border tensor if there exists a sequence of tensors of rank at most $r < s$ whose limit is A. If r is the least value for which such a convergent sequence exists, then it is called the border rank of A.

IV. TUCKER DECOMPOSITION

Tucker decomposition decomposes a tensor into a set of matrices and one small core tensor.In hone, Tucker decomposition is utilized as a demonstrating apparatus. For occasion, it is utilized to show three-way (or higher way) information by implies of moderately little numbers of components for each of the three or more modes, and the components are connected to each other by a three- (or higher-) way center cluster. The demonstrate parameters are evaluated in such a way that, given settled numbers of components, the displayed information ideally take after the real information within the slightest squares sense. The demonstrate gives a summary of the information within the information, within the same way as central components investigation does for two-way information.



V. OTHER DECOMPOSITIONS

A. Compression

In Section before we have emphasized the use of -Tucker/MLSVD for tensor approximation and compression. This use was in fact limited to tensors of moderate order. Let us consider the situation at order N and let us assume for simplicity that $r_1 = r_2 = ... = r_N = r > 1$. Then the core tensor G has r^N entries. The exponential dependence of the number of entries on the tensor order N is called the Curse of Dimensionality: in the case of large N (e.g. N = 100), r^N is large, even when r is small, and as a result -Tucker/MLSVD cannot be used. In such cases one may resort to a Tensor Train (TT) representation or a hierarchical Tucker (hTucker) decomposition instead . A TT of an N-th order tensor X is of the for

$$\mathbf{X}(i_1, i_2, \ldots, i_N) = \sum_{r_1 r_2 \ldots r_{N-1}} u^{(1)}_{i_1 r_1} u^{(2)}_{r_1 i_2 r_2} u^{(3)}_{r_2 i_3 r_3} \cdots u^{(N)}_{i_N r_{N-1}}$$

in which one can see $U^{(1)}$ as the train and the next factors as the carriages. Note that each carriage "transports" one tensor measurement, which two sequential carriages are connected through the summation over one common index. Since each record shows up at most twice and since there are no file cycles, the TT-format is "matrix-like", i.e. a TT approximation can be computed utilizing set up techniques from numerical straight variable based math, essentially to MLSVD. Like for MLSVD, fiber testing plans have been created too. On the other hand, the number of passages is presently $O(NIr^2)$, so the Revile of Dimensionality has been broken. hTucker is the extension in which the files are organized in a double tree.

B. Analysis

As we have know the uniqueness of CPD under mellow conditions as a significant advantage of tensors over matrices within the setting of flag partition and information analysis – limitations such as orthogonality or triangularity are not necessary per se. An indeed more significant advantage is the possibility to have a interesting decay in terms that are not indeed rank-1. Piece Term Disintegrations (BTD) type in a tensor as a entirety of terms that have moo multilinear rank. Note that rank-1 structure of information components is indeed an presumption that must be justified. As in CPD, uniqueness of a BTD is up to a permutation of the terms. The scaling/counterscaling ambiguities inside a rank-1 term generalize to the indeterminacies in a Tucker representation. Extending the piece terms into wholes of rank1 terms with rehashed vectors.

C. Fusion

Different information sets may be together analyzed by implies of coupled disintegrations of a few frameworks and/or tensors, possibly of diverse estimate. An early variation, in which coupling was forced through a shared covariance matrix, is Harshman's PARAFAC2. In a coupled setting, particular disintegrations may acquire uniqueness from other decompositions; in specific, the deterioration of a data matrix may ended up one of a kind much obliged to coupling.

VI. ALGORITHMS

A. Gradient descent

Gradient descent may be a first-order iterative optimization calculation for finding a neighborhood least of a differentiable work. The idea is to require rehashed steps within the inverse direction of the gradient of the work at the current point, since this is often the course of steepest descent. On the other hand, venturing within the heading of the angle will lead to a neighborhood most extreme of that work; the strategy is at that point known as gradient climb.

B. Quasi-Newton and Nonlinear Least Squares

NLS algorithm that has several favorable properties. Briefly, the inexact NLS algorithm uses a "parallel version" of one ALS iteration as a preconditioner for solving the linear system of equations.After preconditioning,is solved inexactly by a truncated conjugate gradient algorithm. That is, the set of equations is not solved exactly and neither is the matrix $D_\theta\varphi(\theta)^T D_\theta\varphi(\theta)$ computed or storage.The algorithm has near to quadratic convergence, particularly when the residuals are little. NLS has been watched to be more vigorous for troublesome decompositions than plain ALS. The activity of $D_\theta\varphi(\theta)^T D_\theta\varphi(\theta)$ can effortlessly be part into littler matrix-vector items(N^2 in the N^{th} arrange case), which makes vague NLS by and large wellsuited for parallel execution. Variations for moo multilinear rank estimation are examined in,and references therein

C. Exact line search

the line search methodology is one of two essential iterative approaches to discover a neighborhood least of an objectif function. The line look approach to begin with finds a plummet course along which the objective work f will be decreased and after that computes a step measure that decides how distant x ought to move along that course. The plunge heading can be computed by different strategies, such as slope plunge or quasi-Newton strategy. The step measure can be decided either precisely or inexactly.

D. Missing values

Mssing values, happen when no information esteem is put away for the variable in an perception. Missing data are a common event and can have a critical impact on the conclusions that can be drawn from the information.

E. Stochastic gradient descent

Stochastic gradient descent (SGD) has become popular in the machine learning community for many types of convex and, very recently, non-convex optimization problems as well. In its simplest form, SGD randomly picks a data point X(i, j, k) from the available ones, and takes a gradient step only for those model parameters that have an effect on X(i, j, k); that is, only the i-th row of A, the j-th row of B and the k-th row of C. We have

$$\frac{\partial}{\partial \mathbf{A}(i,f)} \left(\mathbf{X}(i,j,k) - \sum_{f=1}^{F} \mathbf{A}(i,f)\mathbf{B}(j,f)\mathbf{C}(k,f) \right)^2 = -2 \left(\mathbf{X}(i,j,k) - \sum_{f'=1}^{F} \mathbf{A}(i,f')\mathbf{B}(j,f')\mathbf{C}(k,f') \right) \times \mathbf{B}(j,f)\mathbf{C}(k,f),$$

so that

$$\frac{\partial}{\partial \mathbf{A}(i,:)} = -2\left(\mathbf{X}(i,j,k) - \sum_{f=1}^{F} \mathbf{A}(i,f)\mathbf{B}(j,f)\mathbf{C}(k,f)\right) \times (\mathbf{B}(j,:) * \mathbf{C}(k,:)) .$$

F. Constraints

In hone, we are regularly fascinated by forcing constraints on a CPD show. One may address the require for this – after all, CPD is basically special beneath moderately mellow conditions. Constraints are all things considered valuable in

- Restoring identifiability
- Ensuring interpretability of the results

There are many types of constraints that are relevant in many applications, including :

- Symmetry or Hermitian (conjugate) symmetry
- Element-wise non-negativity
- Real-valued parameters
- Orthogonality
- Probability simplex constraints
- Linear constraints
- Parametric constraints
- Data model constraints
- Sparsity
- Smoothness

VII. APPLICATIONS

A. Blind Multiuser CDMA

Tensors permits us the utilize of multidimensional information, permitting distant better;a much better;a higher;a stronger;an improved"¿a much better understanding and precision for a multidimensional viewpoint. Due to its effective signal processing capabilities, tensors can be found connected to many fields, for case, in chemometrics and others.The Parallel Figure (PARAFAC) decay was to begin with utilized in remote communications frameworks in, where a daze recipient was proposed for a DS-CDMA system and a tensor was utilized to show the gotten flag as a multidimensional variable.

B. Blind source separation

The daze source division show for multivariate time arrangement for the most part accept that the watched arrangement could be a direct change of an imperceptibly arrangement with transiently uncorrelated or free components. Given the perceptions, the objective is to discover a straight change that recoups the inactive arrangement. A few strategies for finishing this exist and three specific ones are the classic SOBI and the as of late proposed generalized FOBI (gFOBI) and generalized JADE (gJADE), each based on the utilize of joint slacked minutes. In this paper we generalize the techniques behind these calculations for tensor-valued time arrangement. We accept that our information comprises of a tensor watched at each time point which the perceptions are direct changes of idle tensors we wish to assess. The tensorial generalizations are appeared to have especially exquisite shapes and we appear that each of them is Fisher steady and orthogonal equivariant.

C. Harmonics

The association between spherical harmonics and symmetric tensors is investigated. For each circular harmonic, a comparing traceless symmetric tensor is developed. These tensors are at that point amplified to incorporate nonzero traces, giving an orthonormal angular-momentum eigenbasis for symmetric tensors of any rank. The relationship between the spherical-harmonic tensors and spin-weighted circular sounds is determined. The results facilitate the spherical-harmonic extension of a huge lesson of tensor-valued capacities. A few straightforward illustrative examples are examined, and the formalism is utilized to determine the leading-order impacts of infringement of Lorentz invariance in Newtonian gravity.

D. Collaborative filtering - based recommender systems

A considerable advance in advancement of modern and proficient tensor factorization procedures has driven to an broad investigate of their appropriateness in recommender frameworks field. Tensor-based recommender models thrust the boundaries of conventional collaborative sifting strategies by taking into consideration a multifaceted nature of genuine situations, which permits to create more precise, situational (e.g. context-aware, criteria-driven) proposals. In spite of the promising comes about, tensor-based strategies are ineffectively secured in existing recommender frameworks studies. This overview points to complement past works and give a comprehensive overview on the subject. To the leading of our information, this is often the primary endeavor to solidify ponders from different application spaces in an effectively lucid, digestible format, which makes a difference to urge a idea of the current state of the field. We moreover give a tall level discourse of long run viewpoints and headings for assist change of tensor-base

E. Gaussian mixture parameter estimation

Gaussian mixture models (GMM) are essential devices in information science. d-th minute of n-dimensional irregular variable, in reality, is symmetric d-way Tensor of estimate n$\hat{\text{d}}$. Subsequently, working with minutes is restrictively costly for any minute degree greater than 2 and bigger values of n There may be a later hypothesis that has been created for certain computations with minute Tensors of GMMs, lessening computational and capacity costs for common covariance matrices. A brief expository expression for minutes in terms of symmetrized Tensor items has been inferred. It depends on the correspondence between symmetric Tensors and homogeneous polynomials. The essential application of this hypothesis is to appraise assess GMM parameters from a set of perceptions when formulated as a moment-matching optimization problem. If there's a known and common covariance network, it is additionally conceivable to de-bias the information perceptions, in which case the issue of evaluating the obscure implies diminishes to symmetric Canonical Polyadic Tensor decomposition.

F. Topic modeling

The trance source division appear for multivariate time course of action for the foremost portion acknowledge that the observed course of action might be a coordinate alter of an subtly course of action with transitorily uncorrelated or free components. Given the discernments, the objective is to find a straight alter that recovers the inert course of action. A couple of procedures for wrapping up this exist and three particular ones are the classic SOBI and the as of late proposed generalized FOBI (gFOBI) and generalized JADE (gJADE), each based on the utilize of joint slacked minutes. The tensorial generalizations are showed up to have particularly dazzling shapes and we show up that each of them is Fisher consistent and orthogonal equivariant.Tensor decompositions give an approach for analyzing both organize activity and geospatial information that has been illustrated to overcome these inadequacies. Tensor disintegrations uncover patterns-of-activity without forthright classification of typical versus anomalous behavior. Assist, tensor investigation works in an self-assertive number of measurements and in this way can distinguish complicated connections between a few information traits simultaneously. While tensor deteriorations have been demonstrated to be a effective instrument for analyzing huge, complex datasets, there stay obstructions to their far reaching utilize. Right now, an examiner who is both a space master and is prepared in tensor arithmetic must characterize the tensor investigation to be performed. Moreover, whereas tensor investigation has been illustrated to be competent of recognizing designs and inconsistencies in huge multidimensional datasets, the investigation right now requires seriously manual review of tensor deterioration comes about.

G. Multilinear discriminative subspace learning

Multilinear subspace learning is an approach to dimensionality reduction.Dimensionality decrease can be performed on a data tensor whose observations have been vectorized and organized into a data tensor, or whose perceptions are lattices that are concatenated into a data tensor.Here are a few illustrations of information tensors whose perceptions are vectorized or whose perceptions are networks concatenated into data tensor pictures (2D/3D), video arrangements (3D/4D), and hyperspectral 3d shapes (3D/4D). The mapping from a high-dimensional vector space to a set of lower dimensional vector spaces may be a multilinear projection.[4] When perceptions are held Multilinear Subspace Learning utilize distinctive sorts of data tensor investigation instruments for dimensionality lessening. Multilinear Subspace learning can be connected to perceptions whose estimations were vectorized and organized into a data tensor,or whose estimations are treated as a network and concatenated into a tensor.

VIII. Conclusion

In this paper we have seen the foundations of tensor algebra; namely the notion of the rank of a tensor, the different tensor decompositions as well as the possible applications for tensors.

References

[1] N. Vervliet, O. Debals, and L. De Lathauwer, "Tensorlab 3.0 — numerical optimization strategies for large-scale constrained and coupled matrix/tensor factorization," in 2016 Conference Record of the 50th Asilomar Conference on Signals, Systems and Computers. IEEE, 2016.

[2] R. Bro and C. Andersson, "Improving the speed of multiway algorithms: Part ii: Compression," Chemometrics and Intelligent Laboratory Systems, vol. 42, no. 12, pp. 105 – 113, 1998.

[3] J. Douglas Carroll, S. Pruzansky, and J. B. Kruskal, "Candelinc: A general approach to multidimensional analysis of many-way arrays with linear constraints on parameters," Psychometrika, vol. 45, no. 1, pp. 3–24, 1980.

[4] B. Savas and L. Elden, "Krylov-type methods for tensor computations ´ i," Linear Algebra and its Applications, vol. 438, no. 2, pp. 891–918, 2013

[5] S. Goreinov, E. Tyrtyshnikov, and N. Zamarashkin, "A theory of pseudoskeleton approximations," Linear Algebra and its Applications, vol. 261, no. 1, pp. 1–21, 1997.

[6] L. Sorber, I. Domanov, M. Van Barel, and L. De Lathauwer, "Exact Line and Plane Search for Tensor Optimization," Computational Optimization and Applications, vol. 63, no. 1, pp. 121–142, Jan. 2016.

[7] R. Bro and N.D. Sidiropoulos, "Least squares regression under unimodality and non-negativity constraints," Journal of Chemometrics, vol. 12, pp. 223–247, 1998.

[8] K. Huang, N. D. Sidiropoulos, and A. P. Liavas, "A flexible and efficient algorithmic framework for constrained matrix and tensor factorization," IEEE Transactions on Signal Processing, vol. 64, no. 19, pp. 5052 – 5065, Oct. 2016.

[9] S. P. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," Foundations and Trends® in Machine Learning, vol. 3, no. 1, pp. 1–122, 2011.

[10] E.E. Papalexakis, N.D. Sidiropoulos, and R. Bro, "From k -means to higher-way co-clustering: Multilinear decomposition with sparse latent factors," IEEE Trans. on Signal Processing, vol. 61, no. 2, pp. 493–506, 2013.

[11] S. Basu and Y. Bresler, "The stability of nonlinear least squares problems and the Cramer-Rao bound," ´ IEEE Transactions on Signal Processing, vol. 48, no. 12, pp. 3426–3436, 2000.

[12] G. Tomasi, Practical and computational aspects in chemometric data analysis, Ph.D. thesis, 2006.