Big Tensor Data Reduction

Nikos Sidiropoulos
Dept. ECE
University of Minnesota

NSF/ECCS Big Data, 3/21/2013

STAR Group, Collaborators, Credits

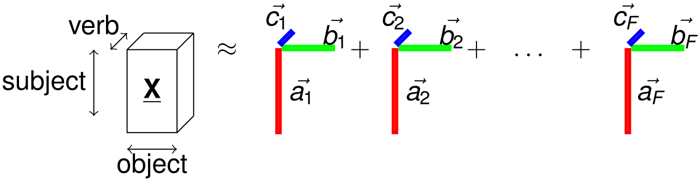
- Signal and Tensor Analytics Research (STAR) group
  `https://sites.google.com/a/umn.edu/nikosgroup/home`
  - Signal processing
  - Big data
  - Preference measurement
  - Cognitive radio
  - Spectrum sensing
- Christos Faloutsos, Tom Mitchell, Vaggelis Papalexakis (CMU), George Karypis (UMN), NSF-NIH/BIGDATA: Big Tensor Mining: Theory, Scalable Algorithms and Applications
- Timos Tsakonas (KTH)
- Tasos Kyrillidis (EPFL)

Tensor? What is this?

- Has different formal meaning in Physics (spin, symmetries)
- Informally adopted in CS as shorthand for *three-way* array: dataset $\underline{\mathbf{X}}$ indexed by three indices, (i, j, k)-th entry $\underline{\mathbf{X}}(i, j, k)$.
- For two vectors $\mathbf{a}$ $(I \times 1)$ and $\mathbf{b}$ $(J \times 1)$, $\mathbf{a} \circ \mathbf{b}$ is an $I \times J$ rank-one matrix with (i, j)-th element $\mathbf{a}(i)\mathbf{b}(j)$; i.e., $\mathbf{a} \circ \mathbf{b} = \mathbf{a}\mathbf{b}^T$.
- For three vectors, $\mathbf{a}$ $(I \times 1)$, $\mathbf{b}$ $(J \times 1)$, $\mathbf{c}$ $(K \times 1)$, $\mathbf{a} \circ \mathbf{b} \circ \mathbf{c}$ is an $I \times J \times K$ rank-one three-way array with (i, j, k)-th element $\mathbf{a}(i)\mathbf{b}(j)\mathbf{c}(k)$.
- The *rank of a three-way array* $\underline{\mathbf{X}}$ is the smallest number of outer products needed to synthesize $\underline{\mathbf{X}}$.
- 'Curiosities':
  - Two-way $(I \times J)$: row-rank = column-rank = rank $\leq \min(I, J)$;
  - Three-way: row-rank $\neq$ column-rank $\neq$ "tube"-rank $\neq$ rank
  - Two-way: rank(randn(I,J))=min(I,J) w.p. 1;
  - Three-way: rank(randn(2,2,2)) is a RV (2 w.p. 0.3, 3 w.p. 0.7)

NELL @ CMU / Tom Mitchell

- Crawl web, learn language 'like children do': encounter new concepts, learn from context
- NELL triplets of "subject-verb-object" naturally lead to a 3-mode tensor



- Each rank-one factor corresponds to a *concept*, e.g., 'leaders' or 'tools'
- E.g., say $\mathbf{a}_1$, $\mathbf{b}_1$, $\mathbf{c}_1$ corresponds to 'leaders': subjects/rows with high score on $\mathbf{a}_1$ will be "Obama", "Merkel", "Steve Jobs", objects/columns with high score on $\mathbf{b}_1$ will be "USA", "Germany", "Apple Inc.", and verbs/fibers with high score on $\mathbf{c}_1$ will be 'verbs', like "lead", "is-president-of", and "is-CEO-of".

Low-rank tensor decomposition / approximation

$$\underline{\mathbf{X}} \approx \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f,$$

- Parallel factor analysis (PARAFAC) model [Harshman '70-'72], a.k.a. canonical decomposition [Carroll & Chang, '70], a.k.a. CP; cf. [Hitchcock, '27]
- PARAFAC can be written as a system of matrix equations $\mathbf{X}_k = \mathbf{A}\mathbf{D}_k(\mathbf{C})\mathbf{B}^T$, where $\mathbf{D}_k(\mathbf{C})$ is a diagonal matrix holding the k-th row of $\mathbf{C}$ in its diagonal; or in compact matrix form as $\mathbf{X} \approx (\mathbf{B} \odot \mathbf{A})\mathbf{C}^T$, using the Khatri-Rao product.
- In particular, employing a property of the Khatri-Rao product,

$$\mathbf{X} \approx (\mathbf{B} \odot \mathbf{A})\mathbf{C}^T \Longleftrightarrow \text{vec}(\mathbf{X}) \approx (\mathbf{C} \odot \mathbf{B} \odot \mathbf{A})\,\mathbf{1},$$

where $\mathbf{1}$ is a vector of all 1's.

Uniqueness

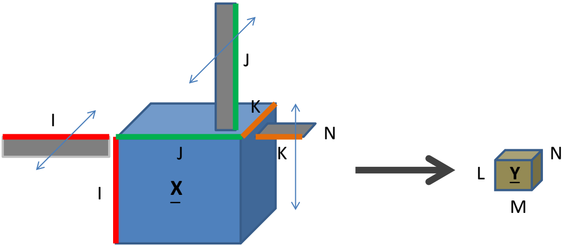
- The distinguishing feature of the PARAFAC model is its essential uniqueness: under certain conditions, $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ can be identified from $\mathbf{X}$, i.e., they are unique up to permutation and scaling of columns [Kruskal '77, Sidiropoulos *et al* '00 - '07, de Lathauwer '04-, Stegeman '06-]
- Consider an $I \times J \times K$ tensor $\underline{\mathbf{X}}$ of rank F. In vectorized form, it can be written as the $IJK \times 1$ vector $\mathbf{x} = (\mathbf{A} \odot \mathbf{B} \odot \mathbf{C})\,\mathbf{1}$, for some $\mathbf{A}$ $(I \times F)$, $\mathbf{B}$ $(J \times F)$, and $\mathbf{C}$ $(K \times F)$ - a PARAFAC model of size $I \times J \times K$ and order F parameterized by $(\mathbf{A}, \mathbf{B}, \mathbf{C})$.
- The *Kruskal-rank* of $\mathbf{A}$, denoted $k_{\mathbf{A}}$, is the maximum k such that *any* k columns of $\mathbf{A}$ are linearly independent ($k_{\mathbf{A}} \leq r_{\mathbf{A}} := \text{rank}(\mathbf{A})$).
- Given $\underline{\mathbf{X}}$ ($\Leftrightarrow \mathbf{x}$), if $k_{\mathbf{A}} + k_{\mathbf{B}} + k_{\mathbf{C}} \geq 2F + 2$, then $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ are unique up to a common column permutation and scaling

Big data: need for compression

- Tensors can easily become really big! - size exponential in the number of dimensions (‘ways’, or ‘modes’).
- Cannot load in main memory; can reside in cloud storage.
- Tensor compression?
- Commonly used compression method for ‘moderate’-size tensors: fit orthogonal Tucker3 model, regress data onto fitted mode-bases.
- Lossless if exact mode bases used [CANDELINC]; but Tucker3 fitting is itself cumbersome for big tensors (big matrix SVDs), cannot compress below mode ranks without introducing errors
- If tensor is sparse, can store as $[i, j, k, value]$ + use specialized sparse matrix / tensor alorithms [(Sparse) Tensor Toolbox, Bader & Kolda]. Useful if sparse representation can fit in main memory.

Tensor compression

- Consider compressing $\mathbf{x}$ into $\mathbf{y} = \mathbf{S}\mathbf{x}$, where $\mathbf{S}$ is $d \times IJK$, $d \ll IJK$.
- In particular, consider a specially structured compression matrix $\mathbf{S} = \mathbf{U}^T \otimes \mathbf{V}^T \otimes \mathbf{W}^T$
- Corresponds to multiplying (every slab of) $\underline{\mathbf{X}}$ from the I-mode with $\mathbf{U}^T$, from the J-mode with $\mathbf{V}^T$, and from the K-mode with $\mathbf{W}^T$, where $\mathbf{U}$ is $I \times L$, $\mathbf{V}$ is $J \times M$, and $\mathbf{W}$ is $K \times N$, with $L \leq I$, $M \leq J$, $N \leq K$ and $LMN \ll IJK$



Key

- Due to a property of the Kronecker product

$$\left(\mathbf{U}^T \otimes \mathbf{V}^T \otimes \mathbf{W}^T\right)(\mathbf{A} \odot \mathbf{B} \odot \mathbf{C}) =$$

$$\left((\mathbf{U}^T\mathbf{A}) \odot (\mathbf{V}^T\mathbf{B}) \odot (\mathbf{W}^T\mathbf{C})\right),$$

from which it follows that

$$\mathbf{y} = \left((\mathbf{U}^T\mathbf{A}) \odot (\mathbf{V}^T\mathbf{B}) \odot (\mathbf{W}^T\mathbf{C})\right)\mathbf{1} = \left(\tilde{\mathbf{A}} \odot \tilde{\mathbf{B}} \odot \tilde{\mathbf{C}}\right)\mathbf{1}.$$

i.e., the compressed data follow a PARAFAC model of size $L \times M \times N$ and order F parameterized by $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}})$, with $\tilde{\mathbf{A}} := \mathbf{U}^T\mathbf{A}$, $\tilde{\mathbf{B}} := \mathbf{V}^T\mathbf{B}$, $\tilde{\mathbf{C}} := \mathbf{W}^T\mathbf{C}$.

Random multi-way compression can be better!

- Sidiropoulos & Kyrillidis, IEEE SPL Oct. 2012
- Assume that the columns of $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are sparse, and let n_a (n_b, n_c) be an upper bound on the number of nonzero elements per column of $\mathbf{A}$ (respectively $\mathbf{B}$, $\mathbf{C}$).
- Let the mode-compression matrices $\mathbf{U}$ ($I \times L, L \leq I$), $\mathbf{V}$ ($J \times M, M \leq J$), and $\mathbf{W}$ ($K \times N, N \leq K$) be randomly drawn from an absolutely continuous distribution with respect to the Lebesgue measure in $\mathbb{R}^{IL}$, $\mathbb{R}^{JM}$, and $\mathbb{R}^{KN}$, respectively.
- If

$$\min(L, k_{\mathbf{A}}) + \min(M, k_{\mathbf{B}}) + \min(N, k_{\mathbf{C}}) \geq 2F + 2, \quad \text{and}$$

$$L \geq 2n_a, \quad M \geq 2n_b, \quad N \geq 2n_c,$$

  then the original factor loadings $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are almost surely identifiable from the compressed data.

Proof rests on two lemmas + Kruskal

- Lemma 1: Consider $\tilde{\mathbf{A}} := \mathbf{U}^T\mathbf{A}$, where $\mathbf{A}$ is $I \times F$, and let the $I \times L$ matrix $\mathbf{U}$ be randomly drawn from an absolutely continuous distribution with respect to the Lebesgue measure in $\mathbb{R}^{IL}$ (e.g., multivariate Gaussian with a non-singular covariance matrix). Then $k_{\tilde{\mathbf{A}}} = \min(L, k_{\mathbf{A}})$ almost surely (with probability 1).
- Lemma 2: Consider $\tilde{\mathbf{A}} := \mathbf{U}^T\mathbf{A}$, where $\tilde{\mathbf{A}}$ and $\mathbf{U}$ are given and $\mathbf{A}$ is sought. Suppose that every column of $\mathbf{A}$ has at most n_a nonzero elements, and that $k_{\mathbf{U}^T} \geq 2n_a$. (The latter holds with probability 1 if the $I \times L$ matrix $\mathbf{U}$ is randomly drawn from an absolutely continuous distribution with respect to the Lebesgue measure in $\mathbb{R}^{IL}$, and $\min(I, L) \geq 2n_a$.) Then $\mathbf{A}$ is the unique solution with at most n_a nonzero elements per column [Donoho & Elad, '03]

Complexity

- First fitting PARAFAC in compressed space and then recovering the sparse **A**, **B**, **C** from the fitted compressed factors entails complexity $O(LMNF + (I^{3.5} + J^{3.5} + K^{3.5})F)$.
- Using sparsity first and then fitting PARAFAC in raw space entails complexity $O(IJKF + (IJK)^{3.5})$ - the difference is huge.
- Also note that the proposed approach does not require computations in the uncompressed data domain, which is important for big data that do not fit in memory for processing.

Further compression - down to $O(\sqrt{F})$ in 2/3 modes

- Sidiropoulos & Kyrillidis, IEEE SPL Oct. 2012
- Assume that the columns of $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are sparse, and let n_a (n_b, n_c) be an upper bound on the number of nonzero elements per column of $\mathbf{A}$ (respectively $\mathbf{B}$, $\mathbf{C}$).
- Let the mode-compression matrices $\mathbf{U}$ ($I \times L, L \leq I$), $\mathbf{V}$ ($J \times M, M \leq J$), and $\mathbf{W}$ ($K \times N, N \leq K$) be randomly drawn from an absolutely continuous distribution with respect to the Lebesgue measure in $\mathbb{R}^{IL}$, $\mathbb{R}^{JM}$, and $\mathbb{R}^{KN}$, respectively.
- If

$$r_{\mathbf{A}} = r_{\mathbf{B}} = r_{\mathbf{C}} = F$$

$$L(L-1)M(M-1) \geq 2F(F-1), \; N \geq F, \quad \text{and}$$

$$L \geq 2n_a, \quad M \geq 2n_b, \quad N \geq 2n_c,$$

then the original factor loadings $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are almost surely identifiable from the compressed data up to a common column permutation and scaling.

Proof: Lemma 3 + results on a.s. ID of PARAFAC

- Lemma 3: Consider $\tilde{\mathbf{A}} = \mathbf{U}^T\mathbf{A}$, where $\mathbf{A}$ ($I \times F$) is deterministic, tall/square ($I \geq F$) and full column rank $r_{\mathbf{A}} = F$, and the elements of $\mathbf{U}$ ($I \times L$) are i.i.d. Gaussian zero mean, unit variance random variables. Then the distribution of $\tilde{\mathbf{A}}$ is nonsingular multivariate Gaussian.
- From [Stegeman, ten Berge, de Lathauwer 2006] (see also [Jiang, Sidiropoulos 2004], we know that PARAFAC is almost surely identifiable if the loading matrices $\tilde{\mathbf{A}}$, $\tilde{\mathbf{B}}$ are randomly drawn from an absolutely continuous distribution with respect to the Lebesgue measure in $\mathbb{R}^{(L+M)F}$, $\tilde{\mathbf{C}}$ is full column rank, and $L(L-1)M(M-1) \geq 2F(F-1)$.

Generalization to higher-way arrays

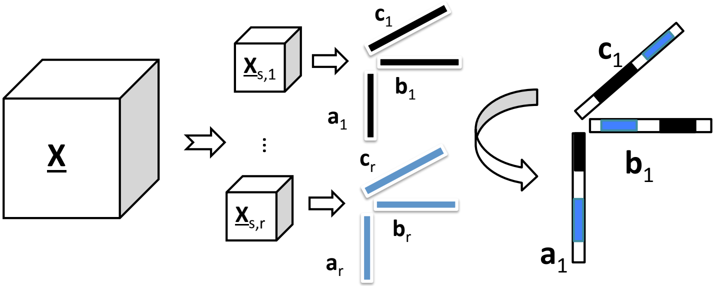
- Theorem 3: Let $\mathbf{x} = (\mathbf{A}_1 \odot \cdots \odot \mathbf{A}_\delta)\,\mathbf{1} \in \mathbb{R}^{\prod_{d=1}^{\delta} I_d}$, where $\mathbf{A}_d$ is $I_d \times F$, and consider compressing it to $\mathbf{y} = \left(\mathbf{U}_1^T \otimes \cdots \otimes \mathbf{U}_\delta^T\right)\mathbf{x} = \left((\mathbf{U}_1^T \mathbf{A}_1) \odot \cdots \odot (\mathbf{U}_\delta^T \mathbf{A}_\delta)\right)\mathbf{1} = \left(\tilde{\mathbf{A}}_1 \odot \cdots \odot \tilde{\mathbf{A}}_\delta\right)\mathbf{1} \in \mathbb{R}^{\prod_{d=1}^{\delta} L_d}$, where the mode-compression matrices $\mathbf{U}_d$ ($I_d \times L_d, L_d \leq I_d$) are randomly drawn from an absolutely continuous distribution with respect to the Lebesgue measure in $\mathbb{R}^{I_d L_d}$. Assume that the columns of $\mathbf{A}_d$ are sparse, and let n_d be an upper bound on the number of nonzero elements per column of $\mathbf{A}_d$, for each $d \in \{1, \cdots, \delta\}$. If

$$\sum_{d=1}^{\delta} \min(L_d, k_{\mathbf{A}_d}) \geq 2F + \delta - 1, \quad \text{and} \quad L_d \geq 2n_d, \quad \forall d \in \{1, \cdots, \delta\},$$

  then the original factor loadings $\{\mathbf{A}_d\}_{d=1}^{\delta}$ are almost surely identifiable from the compressed data $\mathbf{y}$ up to a common column permutation and scaling.
- Various additional results possible, e.g., generalization of Theorem 2.

PARCUBE: Parallel sampling-based tensor decomp

- Papalexakis, Faloutsos, Sidiropoulos, ECML-PKDD 2012



- Challenge: different permutations, scaling
- ‘Anchor’ in small *common* sample
- Hadoop implementation $\rightarrow$ 100-fold improvement (size/speedup)

Road ahead

- Important first steps / results pave way, but simply scratched surface
- Randomized tensor algorithms based on *generalized* sampling
- Other models?
- Rate-distortion theory for big tensor data compression?
- Statistically *and* computationally efficient algorithms - big open issue
- Distributed computations - not all data reside in one place - Hadoop / multicore
- Statistical inference for big tensors
- Applications

Switch gears: Large-scale Conjoint Analysis

- Preference Measurement (PM): Goals
  - Predict responses of individuals based on previously observed preference data (ratings, choices, buying patterns, etc)
  - Reveal utility function - marketing sensitivity
- PM workhorse: Conjoint Analysis (CA)
- Long history in marketing, retailing, health care, ...
- Traditionally offline, assuming rational individuals, responses that regress upon few vars
- No longer true for modern large-scale PM systems, esp. web-based

Conjoint Analysis

- Individual rating J profiles $\{\mathbf{p}_i\}_{i=1}^{J}$, e.g., $\mathbf{p}_i = [\text{screen size, MP, GB, price}]^T$
- **w** is the unknown vector of *partworths*
- Given choice data, $\{\mathbf{d}_i,\ y_i\}_{i=1}^{N}$, $\mathbf{d}_i \in \mathbb{R}^p$, $y_i \in \{-1, +1\}$, $\mathbf{d}_i := \mathbf{p}_i^{(1)} - \mathbf{p}_i^{(2)}$, assumed to obey $y_i = \text{sign}\left(\mathbf{d}_i^{\mathrm{T}}\mathbf{w} + e_i\right), \forall i$
- Estimate partworth vector **w**

Robust statistical choice-based CA

- Preference data *can be inconsistent* (unmodeled dynamics, when seeking **w** of ‘population’ averages; ... but also spammers, fraudsters, prankers!)
- Introduce gross errors $\{o_i\}_{i=1}^N$ in response model (before the sign)
- Sensible to assume that gross errors are *sparse*
- Number of attributes p in **w** can be very large (e.g., cellphones), but only few features matter to any given individual
- Can we exploit these two pieces of prior information in CA context?
- Sparse CA model formulation:

$$y_i = \text{sign}\left(\mathbf{d}_i^T \mathbf{w} + o_i + e_i\right) \quad i = 1, \cdots, N$$

  with constraints $||\mathbf{w}||_0 \leq \kappa_w$ and $||\mathbf{o}||_0 \leq \kappa_o$.
- Small ‘typical’ errors e_i modeled as random i.i.d. $\mathcal{N}(0, \sigma^2)$
- Tsakonas, Jalden, Sidiropoulos, Ottersten, 2012

MLE

- Log-likelihood $l(\mathbf{w}, \mathbf{o})$ can be shown to be

$$l(\mathbf{w}, \mathbf{o}) = \log p_y(\mathbf{w}, \mathbf{o}) = \sum_{i=1}^{N} \log \Phi \left(\frac{y_i \mathbf{d}_i^{\mathrm{T}} \mathbf{w} + y_i o_i}{\sigma} \right)$$

  to be maximized over $||\mathbf{w}||_0 \leq \kappa_w$ and $||\mathbf{o}||_0 \leq \kappa_o$.
- $\Phi(\cdot)$ is the Gaussian c.d.f., so ML metric is a *concave* function
- Cardinality constraints are hard, *relaxing* to ℓ_1-norm constraints yields *convex relaxation*
- Identifiability? Best achievable MSE performance (CRB)?
- Turns out sparsity plays key role in both

Algorithms for Big Data

- Huge volumes of preference data, cannot be analyzed in real-time
- Decentralized collection and/or storage of datasets
- Distributed CA algorithms highly desirable
  - Solve large-scale problems
  - Privacy / confidentiality
  - Fault-tolerance
- Relaxed ML problem is of the form

$$\text{minimize} \quad \sum_{i=1}^{M} f_i(\xi)$$

  and we wish to 'split' w.r.t the training examples only
- Many distributed opt. techniques can be used, one appealing (and recently popular) method is the ADMoM.
- Developed fully decentralized MLE for our CA formulation based on ADMoM
- Tsakonas, Jalden, Sidiropoulos, Ottersten, 2012

Experiments

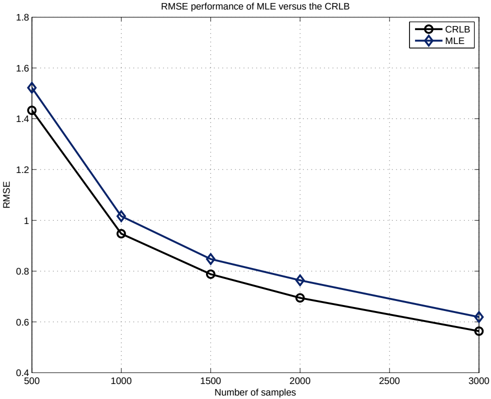


Figure: RMSE comparison of the MLE versus CRLB for different sample sizes N, when outliers are not present in the data.

Experiments

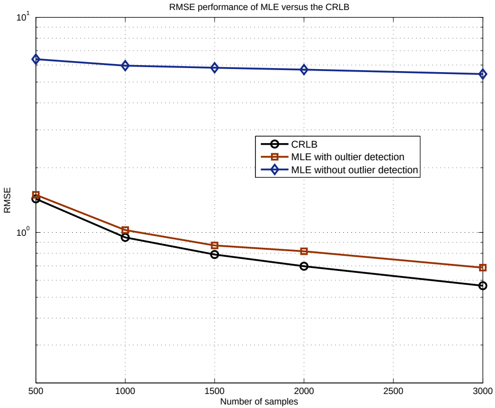


Figure: RMSE comparison of the MLE versus CRLB for different number of samples N, when outliers are present in the data [outlier percentage 4%].