Towards better computation-statistics trade-off in tensor decomposition

Ryota Tomioka
TTI Chicago

Joint work with: T. Suzuki, K. Hayashi, & H. Kashima
Matrices and Tensors in machine learning

Multivariate time-series

<table>
<thead>
<tr>
<th>Time</th>
<th>Fz</th>
<th>Cz</th>
<th>Pz</th>
<th>Oz</th>
</tr>
</thead>
</table>

Collaborative filtering

<table>
<thead>
<tr>
<th>Movies</th>
<th>Star Wars</th>
<th>Titanic</th>
<th>Blade Runner</th>
</tr>
</thead>
</table>
| Users
| User 1  | 5        | 2       | 4            |
| User 2  | 1        | 4       | 2            |
| User 3  | 5        | ?       | ?            |

Spatio-temoral data

Tensors

Multiple relations

Watch
Buy
Like
Matrices and Tensors in machine learning

- Multivariate time-series
  - Sensors
  - Time

- Collaborative filtering
  - Users
  - Movies

- Spatio-temporal data
  - Sensors
  - Space
  - Time

- Multiple relations
  - Users
  - Movies
  - Relations
From matrices to tensors

- Trace norm: convex relaxation of matrix rank

\[ \| W \|_{S_1} = \sum_{j=1}^{r} \sigma_j(W) \]

- It works like L1 regularization on the singular values

- Performance guarantees [Srebro & Schraibman 2005; Candes & Recht 2009; Candes & Tao 2010; Negahban & Wainwright 2011]

Similar relaxation possible for tensor rank?
From matrices to tensors

• Spectral norm of random Gaussian matrix
  \[ \mathbb{E} \| X \|_{S_{\infty}} \leq \sigma (\sqrt{m} + \sqrt{n}) \]

• Marchenko-Pastur distribution
  [Marchenko & Pastur 1967]

Random tensor theory?
Outline

• Tensor ranks and decompositions

• Overlapped trace norm (moderate computation)
  – Limitations: requires $O(rn^{K-1})$ samples

• Balanced trace norm (heavy computation) [Mu et al. 2013]
  – requires $O(r^{K/2}n^{K/2})$ samples

• Tensor trace norm (probably intractable)
  – requires only $O(rn)$ samples
Tensor rank

• Minimum number $R$ such that

$$X_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr}$$

(for 3rd order tensor)

• Known as CP (canonical polyadic) decomposition

[Hitchcock 27; Carroll & Chang 70; Harshman 70]

• Comutation of the above decomposition is NP hard!
Tucker decomposition

\[ X_{ijk} = \sum_{a=1}^{r_1} \sum_{b=1}^{r_2} \sum_{c=1}^{r_3} C_{abc} U^{(1)}_{ia} U^{(2)}_{jb} U^{(3)}_{kc} \]

- Factors can be obtained by unfolding operation + SVD
- In practice no unfolding is low-rank --- Common solution: iterate truncated SVD (HOSVD, HOOI); non-convex
Unfolding (matricization)

Mode-1 unfolding $X_{(1)}$

Mode-2 unfolding $X_{(2)}$

$\begin{array}{c}
 n_1 \\
 n_2 \\
 n_3 
\end{array}$

$\begin{array}{c}
 n_2 \\
 n_2 \\
 n_2 
\end{array}$

$\begin{array}{c}
 n_1 \\
 n_2 \\
 n_3 
\end{array}$

$\begin{array}{c}
 n_3 \\
 n_3 \\
 n_3 
\end{array}$

rank $r_1$

$\begin{array}{c}
 n_2 \cdot n_3 
\end{array}$

rank $r_2$

$\begin{array}{c}
 n_1 \cdot n_3 
\end{array}$
Core idea

Unfolding (Matricization)

Tensor X is low rank
\( \exists k, r_k < n_k \)
(in the sense of Tucker decomposition)

Unfolding \( X^{(k)} \)
is low-rank
(as a matrix)

Tensorization
Overlapped trace norm

\[ [T+10; \text{Signoretto+10; Gandy+11; Liu+09}] \]

- Convex optimization problem

\[
\min_{\mathcal{W} \in \mathbb{R}^{n_1 \times \cdots \times n_K}} \frac{1}{2} \| \mathbf{y} - \mathbf{X}(\mathcal{W}) \|^2 + \lambda_M \| \mathcal{W} \|_{S_1/1}
\]

where

\[
\| \mathcal{W} \|_{S_1/1} := \sum_{k=1}^{K} \| \mathcal{W}^{(k)} \|_{S_1}
\]

- the same tensor is regularized to be

simultaneously low-rank w.r.t. all modes.
Empirical performance

• True tensor: 50x50x20, rank 7x8x9. No noise (λ=0).

• Random train/test split.

Tucker = EM algo (non-convex)  
[Andersson & Bro 00]
Analysis: Problem setting

Observation

\[ y_i = \langle x_i, \mathcal{W}^* \rangle + \epsilon_i \quad (i = 1, \ldots, M) \]

\( \mathcal{W}^* \) : true tensor with rank \((r_1, \ldots, r_K)\)

Gaussian noise \( N(0, \sigma^2) \)

Optimization

\[ \hat{\mathcal{W}} = \arg\min_{\mathcal{W} \in \mathbb{R}^{n_1 \times \cdots \times n_K}} \left( \frac{1}{2} \| y - \mathcal{X}(\mathcal{W}) \|^2 + \lambda_M \| \mathcal{W} \|_{S_{1/1}} \right) \]

\((N = \prod_{k=1}^{K} n_k)\)

Regularization

\(\mathcal{X} : \mathbb{R}^N \rightarrow \mathbb{R}^M\)

Observation operator

\[ \mathcal{X}(\mathcal{W}) = (\langle \chi_1, \mathcal{W} \rangle, \ldots, \langle \chi_M, \mathcal{W} \rangle)^T \]
Theorem ("overlapped" approach) [T, Suzuki, Hayashi, Kashima 11]

Assume that the elements of the design $X$ are independently and identically Gaussian distributed.

Moreover, if

\[
\frac{\text{#samples (} M \text{)}}{\text{#variables (} N \text{)}} \geq c_1 \| n^{-1} \|_{1/2} \| r \|_{1/2} \approx \frac{r}{n}
\]

normalized rank

\[
\| n^{-1} \|_{1/2} := \left( \frac{1}{K} \sum_{k=1}^{K} \sqrt{1/n_k} \right)^2, \quad \| r \|_{1/2} := \left( \frac{1}{K} \sum_{k=1}^{K} \sqrt{r_k} \right)^2
\]
Theorem (random Gauss design) [T, Suzuki, Hayashi, Kashima 11]

Assume that the elements of the design $X$ are independently and identically Gaussian distributed.

Moreover, if

$$\frac{\# \text{samples} (M)}{\# \text{variables} (N)} \geq c_1 \|n^{-1}\|_{1/2} \|r\|_{1/2} \approx \frac{r}{n}$$

Convergence!

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

(with appropriate choice of $\lambda_M$)

$$\|n^{-1}\|_{1/2} := \left( \frac{1}{K} \sum_{k=1}^K \sqrt{1/n_k} \right)^2, \quad \|r\|_{1/2} := \left( \frac{1}{K} \sum_{k=1}^K \sqrt{r_k} \right)^2$$
Tensor completion

size = 50x50x20 true rank 7x8x9 or 40x9x7

Convex [7 8 9]
Covex [40 9 7]
Optimization tolerance

Fraction of observed elements

Estimation error

Fraction at Error<=0.01

#samples (M)
#variables (N)

Normalized rank $\|n_{1/2}\|$ $\|r_{1/2}\|$
Theory vs. Experiments (4\textsuperscript{th} order)

Fraction at err$\leq 0.01$

- size=[50 50 20]
- size=[100 100 50]
- size=[50 50 20 10]
- size=[100 100 20 10]
Limitation: exponentially many samples required!

• Simplify by setting $n_k=n$ and $r_k=r$

• Then there are constants $c_0$, $c_1$, $c_2$ such that

  – #samples $M \geq c_1 n^{K-1} r$
  – reg. const. $\lambda_M = c_0 \sigma \sqrt{n^{K-1}/M}$

  $\| \hat{\mathcal{W}} - \mathcal{W}^* \|_F^2 \leq c_2 \frac{\sigma^2 r n^{K-1}}{M}$

  with high probability.
Why?

- Key steps in the analysis
  - Relation between the norm and the rank
    \[ \| \mathcal{W} \|_{S_{1/1}} \leq K \sqrt{r} \| \mathcal{W} \|_F \quad \text{(OK)} \]
  - Dual norm of noise tensor
    \[
    \mathbb{E} \left\| \mathcal{X}^\top (\epsilon) \right\|_{(S_{1/1})^*} \leq \frac{\sigma \sqrt{M}}{K} \left( \sqrt{n^{K-1}} + \sqrt{n} \right)
    \]
    where \( \mathcal{X}^\top (\epsilon) := \sum_{i=1}^{M} \epsilon_i \mathcal{X}_i \)
Balanced unfolding

• For $K>3$, there are $2^{K-1}-1 > K$ ways to unfold a tensor. For example,

$$X_{(1,2;3,4)} = n_1n_2$$

(See also Mu et al. 2013)
Balanced trace norm (for K=4)

• Definition

\[ \| W \|_{\text{balanced}} := \| W_{(1,2;3,4)} \|_{S_1} + \| W_{(1,3;2,4)} \|_{S_1} + \| W_{(1,4;2,3)} \|_{S_1} \]

– Relation between the norm and the rank

\[ \| W \|_{\text{balanced}} \leq 3\sqrt{r^2} \| W \|_F \]

– Dual norm of noise tensor

\[ \mathbb{E} \| \mathbf{X}^\top (\epsilon) \|_{\text{balanced}}^* \leq \frac{\sigma \sqrt{M}}{3} \cdot 2\sqrt{n^2} \]

Sample complexity \( O(r^2n^2) \)
Experiment (K=4)

tensor completion at rank (2,2,2,2)

\[ M_c = 4.5 \cdot n^{2.93} \]

\[ M_c = 23.3 \cdot n^{2.08} \]

Theoretically

- \( O(n^3) \)
- \( O(n^2) \)

Overlapped (balanced)
Overlapped (unbalanced)
Comparison of computational complexity

- Overlapped trace norm (Sample Complex. $O(rn^{K-1})$)
  
  - requires SVD of $n^{K-1} \times n$ matrix:
    
    $O(n^{K+1}+n^3) \Rightarrow O(n^5)$ for $K=4$ OK

- Balanced trace norm (Sample Complex. $O(r^{K/2}n^{K/2})$)
  
  - requires SVD of $n^{K/2} \times n^{K/2}$ matrix:
    
    $O(n^{1.5K}) \Rightarrow O(n^6)$ for $K=4$ Large!

statistically more efficient, computationally more challenging!
Computation-statistics trade-off

Sample complexity

Frobenius norm

Overlapped trace norm

Balanced trace norm

Computational complexity

$n^K$

$n^{K-1}$

$n^{K/2}$

$n^K$

$n^{K+1}$

$n^{3K/2}$

?
Tensor trace norm

For $K=3$

$$\|\mathcal{W}\|_{tr} = \inf \sum_{a \in A} c_a \quad \text{s.t.} \quad \mathcal{W} = \sum_{a \in A} c_a u_a \circ v_a \circ w_a$$

$$c_a \geq 0$$

$$\|u\| \leq 1, \|v\| \leq 1, \|w\| \leq 1$$

Rank-1 tensor (outer prod. of vectors)

Can be seen as an atomic norm [Chandrasekaran 12] with atomic set = set of rank-1 tensors
Tensor trace norm

For $K=3$

\[ \| \mathcal{W} \|_{tr} = \inf \sum_{a \in \mathcal{A}} c_a \quad \text{s.t.} \quad \mathcal{W} = \sum_{a \in \mathcal{A}} c_a u_a \circ v_a \circ w_a \]

\[ c_a \geq 0 \]

\[ \| u \| \leq 1, \quad \| v \| \leq 1, \quad \| w \| \leq 1 \]

Relation between the norm and the orthogonal CP rank

(Kolda 2001)

\[ \| \mathcal{W} \|_{tr} \leq \sqrt{R} \| \mathcal{W} \|_F \]

Dual norm of the noise tensor

\[ \mathbb{E} \| \mathcal{X}^\top (\epsilon) \|_{tr^*} \leq C \sigma \sqrt{M} \sqrt{n} \]

Sample complexity $O(Rn)$
Dual of the trace norm is the tensor operator norm

\[ \| Y \|_{\text{tr}^*} = \| Y \|_{\text{op}} := \sup_{u,v,w} \sum_{i,j,k} Y_{ijk} u_i v_j w_k \]

s.t. \( \| u \| \leq 1, \| v \| \leq 1, \| w \| \leq 1 \)

Greedy algorithm for computing the operator norm
1. Initialize \( u, v, w \).
2. Fix \( u \), maximize over \( v \) and \( w \) (matrix operator norm)
3. Cycle over \( v, w, u, \ldots \) until convergence
(can be improved by incorporating gradient)
Empirical scaling (K=3)

Theoretically

- $O(n)$
- $O(\sqrt{n})$
Low-rank tensor estimation with the tensor trace norm

\[
\min_{\mathcal{W} \in \mathbb{R}^{n_1 \times \cdots \times n_K}} \frac{1}{2} \| \mathbf{y} - \mathcal{X}(\mathcal{W}) \|^2 + \lambda_M \| \mathcal{W} \|_{\text{tr}}
\]

Likelihood

Regularization

Key operation: prox operator

\[
\text{prox}_\lambda(\mathcal{W}) = \arg\min_{\mathcal{Y}} \left( \lambda \| \mathcal{Y} \|_{\text{tr}} + \frac{1}{2} \| \mathcal{Y} - \mathcal{W} \|^2_F \right)
\]

\[
= \mathcal{W} - \text{proj}_\lambda(\mathcal{W}) \quad \text{(Moreau's theorem)}
\]

\[
\text{proj}_\lambda(\mathcal{W}) = \arg\min_{\mathcal{Y}} \| \mathcal{W} - \mathcal{Y} \|_F \quad \text{s.t.} \quad \| \mathcal{Y} \|_{\text{op}} \leq \lambda
\]

Tensor operator norm
Greedy algorithm for prox$_\lambda$ (W)

1. Let R=W.

2. Compute $\|R\|_{op}$
   
   if $\|R\|_{op} \leq \lambda$, done. Return W-R
   
   otherwise, $R=R+(\lambda-\|R\|_{op}) u \cdot v \cdot w$

3. Go to 2.
Tensor completion experiment

size=50x50x20, CP rank=8

As a matrix (mode 1)
Overlap
Atomic
PARAFAC (large)
PARAFAC (exact)

PARAFAC implemented in N-way toolbox

[Andersson & Bro 00]
Balanced vs. unbalanced

size=25x5x5, CP rank=3

(\lambda \rightarrow 0)

As a matrix
Overlap
Atomic
PARAFAC (large)
PARAFAC (exact)
L2Ball

PARAFAC implemented in N-way toolbox
[Andersson & Bro 00]
Summary

• Tensor decomposition via convex optimization
  – Fast and stable algorithm for tensor decomposition
  – Rank selection is replaced by regularization parameter selection

• Limitation of the overlapped trace norm
  – unbalancedness of the unfolding
  – balanced unfolding

• Optimization statistics trade-off
  – balanced trace norm requires less samples but more computation
  – tensor trace norm requires only $O(n)$ samples but seems intractable
References