Scalable Tensor Factorizations with Incomplete Data

Tamara G. Kolda & Daniel M. Dunlavy
Sandia National Labs

Evrim Acar
Information Technologies Institute, TUBITAK-UEKAE, Turkey

Morten Mørup
Technical University of Denmark
A Nice Way to Start the Day

To: “Kolda, Tamara G” <tgkolda@sandia.gov>
From: “Engineer, Joe S” <jsengin@sandia.gov>
Subject: Help with a math problem

Factor Me!
To: “Engineer, Joe S” <jsengin@sandia.gov>
From: “Kolda, Tamara G” <tgkolda@sandia.gov>
Subject: Re: Help with a math problem

Dear Joe,
No problem thanks to the handy Singular Value Decomposition (SVD)!
–Tammy

p.s. Would you prefer Nonnegative Matrix Factorization (NMF), Independent Component Analysis (ICA), etc.?
A Nicer Way to Start the Day

To: “Kolda, Tamara G” <tgkolda@sandia.gov>
From: “Engineer, Joe S” <jsengin@sandia.gov>
Subject: Help with a math problem
To: “Engineer, Joe S” <jsengin@sandia.gov>
From: “Kolda, Tamara G” <tgkolda@sandia.gov>
Subject: Re: Help with a math problem

Dear Joe,

I factored your tensor using CANDECOMP/PARAFAC (CP) tensor decomposition. Though determining tensor rank is NP hard, I used heuristic methods to make a reasonable determination of the number of components.

–Tammy
Dear Tammy,

Our measurement device had some snafus, and so we couldn’t collect all the data. But we still need to factorize this (incomplete) matrix. Maybe we can just fill in the missing entries with zeros?

-Joe
To: “Engineer, Joe S” <jsengin@sandia.gov>
From: “Kolda, Tamara G” <tgkolda@sandia.gov>
Subject: Re: Help with a math problem

Dear Joe,

Per Ruhe (1974): “When the number of missing, observations is small, compared to the sample, it is common practice to replace the missing observations by means or extreme values, but when larger parts of the data matrix are empty this is no longer a sensible approach.”

However, many modern methods exist for factorizing matrices with missing data. A favorite reference is Buchanan & Fitzgibbon (CVPR’05). See also the work on matrix completion by Candès, Recht, and others. Most problems can be solved so long as there are enough known entries and they are distributed reasonably.

–Tammy
Dear Tammy,

Our measurement device had some snafus, and so we couldn’t collect all the data. But we still need to factorize this (incomplete) tensor.

-Joe
Factor Example: Latent Semantic Indexing

Berry, Dumais, O’Brien (1995)

Matrix Representation

\[
X = \begin{bmatrix}
1 & 0 & 1 \\
1 & 1 & 0 \\
0 & 1 & 0 \\
1 & 0 & 1 \\
\end{bmatrix} = \begin{bmatrix}
.91 & -.38 \\
.72 & .75 \\
.19 & .75 \\
.91 & -.38 \\
\end{bmatrix} \\
\begin{bmatrix}
1.15 & 0 \\
.41 & 1.06 \\
.83 & -.53 \\
\end{bmatrix}^T + E
\]

2-Factor Approximation

\[
AB^T = \hat{A}\hat{B}^T \equiv (AS)(BS^{-T})^T
\]

Terms
- car
- service
- military
- repair

Documents
- d1
- d2
- d3

Gauge Freedom

\[
X = \begin{bmatrix}
.39 & .90 \\
1.04 & -.04 \\
.66 & -.41 \\
.39 & .90 \\
\end{bmatrix} = \begin{bmatrix}
.83 & 0.80 \\
1.04 & -.48 \\
.23 & .96 \\
\end{bmatrix}^T + E
\]
Factor Example: Epilepsy

Data measurements are recorded at multiple sites (channels) over time. The data is transformed via a continuous wavelet transform.

\[ X = a_1 \circ b_1 \circ c_1 + a_2 \circ b_2 \circ c_2 + \varepsilon \]

A Whole Host of Tensor Decompositions

Tensor Factorizations have Numerous Applications

- Modeling fluorescence excitation-emission data
- Signal processing
- Brain imaging (e.g., fMRI) data
- Web graph plus anchor term analysis
- Image compression and classification; texture analysis
- Text analysis, e.g., multi-way LSI
- Approximating Newton potentials, stochastic PDEs, etc.

\[
\begin{align*}
\mathcal{L}(x,t;\omega,u) &= f(x,t;\omega) \quad (x,t) \in \mathcal{D} \times [0,T] \\
\mathcal{B}(x,t;\omega,u) &= g(x,t) \quad (x,t) \in \partial \mathcal{D} \times [0,T] \\
\mathcal{I}(x,0;\omega,u) &= h(x,\omega) \quad x \in \mathcal{D}.
\end{align*}
\]


Furukawa, Kawasaki, Ikeuchi, and Sakauchi, *EGRW ‘02*

Doostan, Iaccarino, and Etemadi, *J. Computational Physics*, 2009

Sun, Tao, and Faloutsos, *KDD’06.*

ERPWAVELAB by Morten Mørup.


Matrix & Tensor Factor Analysis

Singular Value Decomposition (SVD) or Principal Component Analysis (PCA) expresses a matrix as the sum of component rank-1 matrices.

\[ X = \sum_{r=1}^{R} a_r b_r^T \]

\[ X = A B^T \]

CANDECOMP/PARAFAC Decomposition (CP) expresses a tensor as the sum of component rank-1 tensors. (Hitchcock ’27, Harshman ’70, Carroll & Chang ’70)

\[ X = \sum_{r=1}^{R} a_r \circ b_r \circ c_r \]

\[ X = [A, B, C] \]

\[ A = [a_1 \ldots a_R] \quad B = [b_1 \ldots b_R] \quad C = [c_1 \ldots c_R] \]
Tensors have an Advantage for Recovering Factors

Matrices

\[ X = A B^T \]

- Scale & sign ambiguity
  \[ \hat{a}_p = \gamma a_p \quad \hat{b}_p = 1/\gamma b_p \]

- Permutation ambiguity
  \[ \hat{a}_p, \hat{b}_p = a_q, b_q \]
  \[ \hat{a}_q, \hat{b}_q = a_p, b_p \]

- Gauge freedom for any invertible matrix \( S \):
  \[ \hat{A} = AS \]
  \[ \hat{B} = BS^{-T} \]

Tensors

\[ X = [A, B, C] \]

- Scale and sign ambiguity
  \[ \hat{a}_p = \alpha a_p \quad \hat{b}_p = \beta b_p \quad \hat{c}_p = \gamma c_p \]
  \[ \alpha \beta \gamma = 1 \]

- Permutation ambiguity
  \[ \hat{a}_p, \hat{b}_p, \hat{c}_p = a_q, b_q, c_q \]
  \[ \hat{a}_q, \hat{b}_q, \hat{c}_q = a_p, b_p, c_p \]

- Otherwise components unique under mild conditions!
Kruskal’s Uniqueness Condition

The \( k \)-rank of a matrix \( A \), denoted \( k_A \), is the maximum value of \( k \) such that any \( k \) columns are linearly independent.

An \( R \)-component factorization \([A, B, C]\) is essentially unique* if:

\[
k_A + k_B + k_C \geq 2R + 2
\]

An \( R \)-component factorization \([A^{(1)}, ..., A^{(N)}]\) is essentially unique* if:

\[
k_A^{(1)} + \cdots + k_A^{(N)} \geq 2R + (N - 1)
\]


* Essentially unique means up to permutation and scaling:

\[
[A, B, C] = [A\Pi, B\Pi, C\Pi] \quad [A, B, C] = \sum_{r=1}^{R} (\alpha_r a_r \circ (\beta_r b_r \circ (\gamma_r c_r))
\]

\( \Pi = R \times R \) permutation matrix

so long as: \( \alpha_r \beta_r \gamma_r = 1 \)
Biomedical signal processing

- EEG (electroencephalogram) signals can be recorded using electrodes placed on the scalp
- Missing data problem occurs when...
  - Electrodes get loose or disconnected, causing the signal to be unusable
  - Different experiments have overlapping but not identical channels

Can we still do this calculation if data are missing?
The Missing Data Problem

Standard Problem:
Given tensor $X$, find $A$, $B$, and $C$ such that...

$$X = [A, B, C]$$

Typically formulated as a least squares problem.

$$\min_{A, B, C} \frac{1}{2} \| X - [A, B, C] \|^2$$

Missing Data Problem:
Given a subset of the entries of $X$, find $A$, $B$, and $C$ such that...

$$(X)_{ijk} = ([A, B, C])_{ijk}$$

for the known entries.
Define the “weight” tensor $W$ such that

$$ w_{ijk} = \begin{cases} 1 & \text{if entry } (i, j, k) \text{ of } \mathbf{X} \text{ is known} \\ 0 & \text{if entry } (i, j, k) \text{ of } \mathbf{X} \text{ is missing} \end{cases} $$

Then the least squares problem is...

$$ \min_{A,B,C} \frac{1}{2} \left\| W \ast (\mathbf{X} - [A, B, C]) \right\|^2 $$

Elementwise product ($\ast$ in MATLAB)

With a solution, the tensor can be “completed” via...

$$ \bar{x}_{ijk} = \begin{cases} x_{ijk} & \text{if entry } (i, j, k) \text{ of } \mathbf{X} \text{ is known} \\ ([A, B, C])_{ijk} & \text{if entry } (i, j, k) \text{ of } \mathbf{X} \text{ is missing} \end{cases} $$
Brain dynamics can be captured even extensive missing channels

![Diagram showing brain dynamics and missing channels]

<table>
<thead>
<tr>
<th>Number of Missing Channels</th>
<th>Replace Missing Entries with Zero</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.98</td>
</tr>
<tr>
<td>10</td>
<td>0.82</td>
</tr>
<tr>
<td>20</td>
<td>0.67</td>
</tr>
<tr>
<td>30</td>
<td>0.45</td>
</tr>
<tr>
<td>40</td>
<td>0.24</td>
</tr>
</tbody>
</table>

http://www.madehow.com/
Brain dynamics can be captured even extensive missing channels

28 exps. 4392 time-freq.

64 channels = +

<table>
<thead>
<tr>
<th>Number of Missing Channels</th>
<th>Replace Missing Entries with Zero</th>
<th>More Sensible Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.98</td>
<td>1.00</td>
</tr>
<tr>
<td>10</td>
<td>0.82</td>
<td>0.98</td>
</tr>
<tr>
<td>20</td>
<td>0.67</td>
<td>0.95</td>
</tr>
<tr>
<td>30</td>
<td>0.45</td>
<td>0.89</td>
</tr>
<tr>
<td>40</td>
<td>0.24</td>
<td>0.65</td>
</tr>
</tbody>
</table>
Brain dynamics can be captured even extensive missing channels

http://www.madehow.com/

No Missing Data

<table>
<thead>
<tr>
<th>channel</th>
<th>time-freq</th>
<th>experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Channel 1" /></td>
<td><img src="image2.png" alt="Time-Freq 1" /></td>
<td><img src="image3.png" alt="Experiments 1" /></td>
</tr>
<tr>
<td><img src="image4.png" alt="Channel 2" /></td>
<td><img src="image5.png" alt="Time-Freq 2" /></td>
<td><img src="image6.png" alt="Experiments 2" /></td>
</tr>
</tbody>
</table>

30 Chan./Exp. Missing

<table>
<thead>
<tr>
<th>channel</th>
<th>time-freq</th>
<th>experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image7.png" alt="Channel 3" /></td>
<td><img src="image8.png" alt="Time-Freq 3" /></td>
<td><img src="image9.png" alt="Experiments 3" /></td>
</tr>
<tr>
<td><img src="image10.png" alt="Channel 4" /></td>
<td><img src="image11.png" alt="Time-Freq 4" /></td>
<td><img src="image12.png" alt="Experiments 4" /></td>
</tr>
</tbody>
</table>

July 15, 2010
Fluorescence measurements of 5 samples containing 3 amino acids

- Tyrosine
- Tryptophan
- Phenylalanine

Each amino acid corresponds to a rank-one component

Tensor of size $5 \times 51 \times 201$

- 5 samples
- 51 excitations
- 201 emissions
No Missing Data

[Movie]
No Missing Data
Replacing 50% Missing Values with Mean

[Movie]
Replacing 50% Missing Values with Mean
50% Missing Data using Sensible Approach

[Movie]
50% Missing Data using Sensible Approach
75% Missing Data using Sensible Approach

[Movie]
75% Missing Data using Sensible Approach
Sensible Approach: Direct Optimization

Goal is to minimize the following nonlinear optimization function:

\[
f(A, B, C) = \frac{1}{2} \| y - z \|^2
\]

\[y = \mathcal{W} \ast \mathcal{X} \text{ and } z = \mathcal{W} \ast [A, B, C]\]

Fits only the known values.

• 2\textsuperscript{nd} order optimization
  • Gauss-Newton as implemented in \textsc{INDAFAC}
  • Size of Jacobian is (1-M) IJK [# equations] by R(I+J+K) [# variables]
  • Tomasi & Bro, 2005

• 1\textsuperscript{st} order optimization
  • Nonlinear CG as implemented in \textsc{CP-WOPT (available in Tensor Toolbox)}
  • Acar, Dunlavy, K., Mørup, 2010

• Alternative to direct optimization: \textbf{EM-ALS}
  • Repeat: Impute missing values from model, update model
  • Implementation in parafac from N-Way Toolbox (Bro et al.)
EM-ALS Algorithm

Implementation: parafac in N-way Toolbox by Bro et al.

1. Initialize matrices A, B, C
2. Complete X using the current model:
   \[
   \bar{X} = W \times X + (1 - W) [A, B, C]
   \]
3. Compute a new A, B, C that best fit the completed X via one or more iterations of alternating least squares (ALS):
   \[
   A = \arg \min_A \| \bar{X} - [A, B, C] \|
   \]
   \[
   B = \arg \min_B \| \bar{X} - [A, B, C] \|
   \]
   \[
   C = \arg \min_C \| \bar{X} - [A, B, C] \|
   \]
4. Go back to step 2 (until convergence)
Observe that $Y$ and $Z$ are sparse if $W$ is sparse:

$$f(A, B, C) = \frac{1}{2} \| Y - Z \|^2$$

$Y = W \times X$ and $Z = W \times [A, B, C]$

Gradient can be calculated as follows:

$$\frac{\partial f}{\partial A} = (Z^{(1)} - Y^{(1)}) (C \odot B)
\frac{\partial f}{\partial B} = (Z^{(2)} - Y^{(2)}) (C \odot A)
\frac{\partial f}{\partial C} = (Z^{(3)} - Y^{(3)}) (B \odot A)$$

Sparse calculation between two tensors with the same set of structural nonzeros.

More missing data yields a sparser problem!
Numerical Experiment
Set-up for Simulated Data

Step 1: Generate factor matrices with random entries drawn from \( N(0,1) \). Each matrix has \( R=5 \) columns. Normalize each column to length 1.

\[
X_{\text{full}} = [A, B, C] + \eta \frac{||X||}{||N||} \mathcal{N}
\]

Step 2: Generate tensors and add noise \((\eta = 10\%)\).

Step 3: Randomly remove entries or entire fibers

Step 4: Factorize incomplete tensor
- CP-WOPT – Dense or Sparse
- INDAFAC (Gauss-Newton)
- EM-ALS

Step 5: Compute \( FMS = \) factor match score of computed factors against truth. Assume columns are normalized and \( \xi_r \) is the product of the norms.

\[
\frac{1}{R} \sum_r \left( 1 - \frac{|\xi_r - \hat{\xi}_r|}{\max(\xi_r, \hat{\xi}_r)} \right) a_r^T \hat{a}_r b_r^T \hat{b}_r c_r^T \hat{c}_r
\]

Best FMS is 1
No Difference in Accuracy Between Methods

Missing Entries

50 x 40 x 30, R=5

80% Missing

90% Missing

95% Missing

CPWOPT-D/S
INDAFAC
EM-ALS
No Difference in Accuracy Between Methods

Missing Entries
50 x 40 x 30, R=5

80% Missing
90% Missing
95% Missing

95% missing is pretty hard at this size.

Number of starts ➔
Larger Problems are Easier

ρ = \frac{\# \text{ known tensor entries}}{\# \text{ variables}}

95% Missing Entries, R=5

50 \times 40 \times 30

ρ = 5

100 \times 80 \times 60

ρ = 20

150 \times 120 \times 90

ρ = 45

Number of starts →
Structured Missing Data is Harder

80% Missing
50 x 40 x 30, R=5

Missing Entries

Missing Fibers

FMS

CPWOPT-D/S
INDAFAC
EM-ALS

July 15, 2010
T. G. Kolda - SIAM Annual
Exploiting Sparsity for Speed

- 80% Missing
- 90% Missing
- 95% Missing

- 50x40x30
- 100x80x60
- 150x120x90

- CPWOPT-D
- CPWOPT-S
- INDAFAC
- EM-ALS
CP-WOPT scales to larger problems!

500 x 500 x 500 with M=99% (1.25 million nonzeros)

- Dense storage = 1GB
- Sparse storage = 40MB

1000 x 1000 x 1000 with M = 99.5% (5 million nonzeros)

- Dense storage = 8GB
- Sparse storage = 160MB

Dense storage = 1GB
Sparse storage = 40MB

Dense storage = 8GB
Sparse storage = 160MB

Failed

Success!

Restarted and converged to correct solution!
To: “Engineer, Joe S” <jsengin@sandia.gov>
From: “Kolda, Tamara G” <tgkolda@sandia.gov>
Subject: Re: Help with a math problem

Joe,

Here you go. There’s still a lot of work to do on this topic, but I think you’ll find this to be a workable solution in the short term. Thanks for the opportunity to work on such an interesting problem.

-Tammy
Challenges for Missing Data Problems

• Ubiquitous Real-World Applications
  – Data doesn’t perfectly fit (multi-) linear model
  – Missing data patterns may be structured

• Theory
  – Many recent discoveries in matrix case
  – Need similar ideas (e.g., “coherence”, “nuclear norm”) for tensors!

• Algorithms
  – Extensions to other tensor factorizations
  – Purposely dropping data for memory/speed of computation or maybe to avoid local minima
  – Rank determination

For more info:
Tammy Kolda
tgkolda@sandia.gov

Back-Up Slides
Network traffic data
• Comprises source-destination information over time (tensor)
• Missing data arises due to various collection issues

Our data set
• One month of Géant data
• 23 x 23 x 2756 (15 min intervals)
• Modeled using two components