Nested Nonnegative Cone Analysis Method

Lingsong Zhang
lingsong@purdue.edu

May 13, 2014
Collaborators

- This talk, Nested Nonnegative Cone Analysis: J. S. Marron (UNC) and Shu Lu (UNC)
  - Started from the Object-Oriented Data Analysis program at SAMSI

- Related projects:
  - Nested Semi-Definite Cone analysis: Sungkyu Jung (U of Pittsburgh)
  - Nonnegative Dimension Reduction methods: Naomi Altman (Penn State), Xingye Qiao (Binghamton), Yifan Xu (Case Western), Kossi Edoh (NCAT) [SAMSI project]

- Thank Stan Young (NISS) and Jiayang Sun (Case Western) for discussion offline.
Object-Oriented Data Analysis

- Unit of interest
  - Univariate statistics: numbers
  - Multivariate statistics: vectors
  - Functional Data Analysis: functions, curves
  - *All elements are in an Euclidean space*

- Object-Oriented Data Analysis: more complicated data: trees, manifolds, etc.
  - *They can be NOT in an Euclidean space.*

- My recent interests: data lies in structured spaces(/sets)
  - Partially Euclidean, Partially Not.
  - *Nonnegative cone* and Semi-definite cone
Examples of cone-type data

- Nonnegative Cone: data lies in the first orthant
  - Branch length representation of tree data
  - Mass spectrum data
  - Derivatives of monotonic curves.

- Semi-define cone: population of semi-definite matrices
  - Diffusion tensors
  - Population of networks

- Not just for mathematical fun. Expect better interpretability
Visual impression - Nonnegative Cone
2x2 positive definite matrices: inside of surface
Goal of our research

Study population of data that lies in cones

- First step: Identify major modes of variations within a sample of such data

- Second step: Use these data to predict other features (disease type, other important variables)

- We will focus on Nonnegative Cone today, more results are in future conferences.
PCA/SVD

- PCA is a central tool in multivariate analysis and FDA
  - Low dimensional approximation to the original data
  - Further analysis can be based on PCs
  - It reveals useful data structure
  - e.g. Jolliffe (2002), Rice and Silverman (1991), Yao, Müller and Wang (2005), ⋯

- SVD is also useful
  - Mathematical tool for matrix factorization, widely used to calculate PCA
  - Can directly help to identify low dimensional approximation
  - Non-central version of PCA
  - Some examples: Shen and Huang (2005), Zhang et al (2007), ⋯
Let $\mathbf{X}$ be a $d \times n$ matrix, SVD is to identify $\mathbf{U}_{d \times k}$ and $\mathbf{V}_{n \times k}$, such that they minimize

$$\| \mathbf{X} - \mathbf{U} \mathbf{V}^T \|_F^2,$$

with some unit norm constraints and orthogonality constraints.

- Best rank $k$ approximation, no further constraints

PCA is similar, but with a constant adjustment. PCA is to identify $m_{d \times 1}$, $\mathbf{U}^c_{d \times k}$ and $\mathbf{V}^c_{n \times k}$, such that they minimize

$$\| \mathbf{X} - m \mathbf{1}^T - \mathbf{U}^c (\mathbf{V}^c)^T \|_F^2,$$

where $\mathbf{1}$ is a $n \times 1$ vector which all elements are 1.

- From now on, we may skip $m$ for easy presentation.
Let \( \mathbf{U} = [u_1, \cdots, u_r] \), and \( \mathbf{V} = [v_1, \cdots, v_r] \), then

\[
\mathbf{UV}^T = u_1 v_1^T + \cdots + u_r v_r^T
\]

Each \( u_i v_i^T \) provides a rank 1 approximation to \( \mathbf{X} \).

Cumulative \( \mathbf{A} := \sum_{i=1}^{k} u_i v_i^T \) provides best rank \( k \) approximation to \( \mathbf{X} \) (Eckart and Young, 1936).

\( \mathbf{A} \) can be viewed as a (smoothed or de-noised) version of \( \mathbf{X} \).
Drawbacks for PCA/SVD for nonnegative data

- Usually only the first component is within the nonnegative cone. (Perron-Frobenius Theorem)
- The second and later components tend to leave the nonnegative cone. (Orthogonality)
- This may result in no physical interpretation.
A toy example (nonnegative matrix)

Six observations in $\mathbb{R}^3$.

$$X = \begin{pmatrix}
0.0939 & 0.9018 & 0.6182 & 0 & 0 & 0 \\
0.8457 & 0.0165 & 0.4746 & 0.2046 & 0.7476 & 0 \\
0 & 0 & 0 & 0.4511 & 0.6986 & 0.8007
\end{pmatrix},$$
PCA on the toy example
## PCA approximations

<table>
<thead>
<tr>
<th>Rank 1</th>
<th>Rank 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\begin{pmatrix} 0.2801 &amp; 0.3794 &amp; 0.3154 \ 0.8175 &amp; 0.2769 &amp; -0.1536 \ 0.6111 &amp; 0.3163 &amp; 0.0265 \ 0.0770 &amp; 0.4181 &amp; 0.4926 \ -0.1008 &amp; 0.4520 &amp; 0.6477 \ -0.0710 &amp; 0.4464 &amp; 0.6217 \end{pmatrix}$</td>
<td>$\begin{pmatrix} 0.1814 &amp; 0.9007 &amp; 0.0883 \ 0.8705 &amp; -0.0032 &amp; -0.0316 \ 0.5851 &amp; 0.4538 &amp; -0.0334 \ 0.1049 &amp; 0.2706 &amp; 0.5569 \ -0.1401 &amp; 0.6596 &amp; 0.5573 \ 0.0121 &amp; 0.0076 &amp; 0.8129 \end{pmatrix}$</td>
</tr>
</tbody>
</table>

- There are 3 observations for each rank approximation that are out of the cone.
- The approximation are not sparse.
SVD on the toy example

![Graph showing SVD](image)
### SVD approximations

<table>
<thead>
<tr>
<th>Rank 1</th>
<th>Rank 2</th>
</tr>
</thead>
</table>
| \[
\begin{pmatrix}
0.2090 & 0.5136 & 0.3828 \\
0.0906 & 0.2228 & 0.1660 \\
0.1717 & 0.4220 & 0.3144 \\
0.1278 & 0.3142 & 0.2342 \\
0.2998 & 0.7370 & 0.5492 \\
0.1411 & 0.3467 & 0.2584 \\
\end{pmatrix}
\] |
| \[
\begin{pmatrix}
0.3298 & 0.5256 & 0.3007 \\
0.7049 & 0.2837 & -0.2511 \\
0.6240 & 0.4668 & 0.0073 \\
-0.0666 & 0.2950 & 0.3662 \\
0.0277 & 0.7100 & 0.7339 \\
-0.2286 & 0.3101 & 0.5094 \\
\end{pmatrix}
\] |

- The rank 1 approximation is always in the nonnegative cone.
- There are 3 observations for the rank 2 approximation that are out of the cone.
- The approximations are not sparse.
A useful approach in image analysis and relate areas. (Paatero and Tapper, 1994; Lee and Seung, 1999)

Very popular algorithm in machine learning and applied fields

Some large-scale computing algorithms are available.

The low rank approximation is nonnegative, (usually has physical interpretation)
There are several formulations of NMF. The following one is more related to SVD/PCA:

Let $\mathbf{X}$ be a $d \times n$ matrix, NMF is to find $\mathbf{W}_{d \times k}$ and $\mathbf{H}_{n \times k}$, such that they minimize

$$\| \mathbf{X} - \mathbf{W} \mathbf{H}^T \|_F^2,$$

while $\mathbf{W} \geq 0$, $\mathbf{H} \geq 0$. 
The factorization may not be unique (Donoho and Stodden, 2004)

The approximation may not be unique (see later simulation)

Different approximations at different ranks are not nested within each other
  - hard to compare between them, and thus model selection may be challenging.
NMF on the toy example
NMF on the toy example

NMF (different angle)
We focus on identifying a unique low rank nonnegative approximation matrix.

Desired properties:
- Non-negativity
- Uniqueness
- Optimality (in terms of approximation)
- Interpretability
Backward algorithm

- Forward type is challenging
  - All the residuals at any ranks are required to be nonnegative

- Backward type is more natural.
  - Corresponding to adding constraints/conditions in learning information
  - See some discussion in Damon and Marron (2013)
  - Multiple-scale view
Let $X$ be $d \times n$ matrix, and rank is $r_0$.

We first identify $A_{r_0-1}$, where

$$\|X - A_{r_0-1}\|_F^2 = \min!$$

and $\text{rank}(A_{r_0-1}) = r_0 - 1$, and $A_{r_0-1} \geq 0$.

Then given $A_{k+1}$, we are going to identify $A_k$, such that

$$\|A_{k+1} - A_k\|_F^2 = \min!$$

and $\text{rank}(A_k) = k$, and $A_k \geq 0$. 

This NNCA provides a sequence of approximation matrices, indexed by the ranks.

Similar to putting additional constraints a time to learn the structure of a data.

A central algorithm is to identify a rank $k$ nonnegative approximation matrix $A$ to another nonnegative matrix $B$, where $\text{rank}(B) = k + 1$. 
NNCA approximation

<table>
<thead>
<tr>
<th>Rank 1</th>
<th>Rank 2</th>
</tr>
</thead>
</table>
| \[
\begin{pmatrix}
0.2676 & 0.5252 & 0.3491 \\
0.2211 & 0.4339 & 0.2884 \\
0.2364 & 0.4641 & 0.3085 \\
0.1641 & 0.3222 & 0.2141 \\
0.3629 & 0.7124 & 0.4735 \\
0.2043 & 0.4010 & 0.2665 \\
\end{pmatrix}
| \[
\begin{pmatrix}
0.3298 & 0.5256 & 0.3007 \\
0.5922 & 0.4365 & 0 \\
0.6240 & 0.4668 & 0.0073 \\
0 & 0.3210 & 0.3417 \\
0.0277 & 0.7100 & 0.7339 \\
0 & 0.3996 & 0.4253 \\
\end{pmatrix}
| Non-negativity.  
| Showing sparsity. |
NNCA on to the toy example
NNCA on to the toy example

NNCA (different angle)
The nested sequence of $\mathbf{A}$'s can lead to a new set of orthonormal bases.

- Help to reveal clusters (similar to PCA/SVD)

The sequence provides multi-resolution/multi-scale view of the original data

- A full spectrum of the data
- Connection between different ranks (working in progress)

Learning by parts

- What additional information we get by increasing the ranks? (working in progress)

- 81 IR spectra, a polymeric material measured over a 27-day cooling period
- 1556 measurements per spectrum
- Target: how different components change over time
NNCA result - a new set of orthogonal basis
NNCA result - NNCA 2

![NNCA Spectrum Graph]

L. Zhang (Purdue)

May 13, 2014 34 / 42
NNCA result - NNCA 3
Summary

- Novel approaches for analysis of data lies in cone
  - A sequence of nested cones
- Uniqueness and Nested Structure
- Multi-level approximation
Future work

- Optimization optimality (current work)
- Statistical theory: consistency, efficiency, · · · (current work)
- Fast algorithms (parallel computing?)
- More results will be reported in future presentations
Thank you