The Role of Multi-linear Constrained Factorization in Image Coding, Clustering and Visual Learning

Amnon Shashua

School of Computer Science & Eng.
The Hebrew University

Tamir Hazan, Ron Zass

Shashua & Hazan - ICML'05
Hazan, Polak & Shashua - ICCV'05
Zass & Shashua - ICCV'05
Shashua, Zass & Hazan - ECCV'06
Zass & Shashua - NIPS06
Low Rank Factorization Algorithms

Matrix Factorization Algorithms:

- Principle Component Analysis (PCA) and its probabilistic versions (Buntine & Perttu, 2003; Tipping & Bishop, 1999),
- Latent Semantic Analysis (Deerwester et al., 1990),
- Probabilistic Latent Semantic Analysis (Hofmann, 1999),
- Maximum Margin Matrix Factorization (Srebro, Rennie & Jaakola, 2005)

All methods factorize the data into a lower dimensional space in order to introduces a compact basis which if set up appropriately can describe the original data in a concise manner.

What are the possible interpretations of high dimensional (tensor) non-negative decompositions?
Outline

• Conditional Independence, tensor-rank, latent class models

• Conditional Independence and Clustering

• High-order affinity clustering (model selection, multiple 3D body segmentation, varying illumination segmentation).
A rank=1 matrix $G$ is represented by an outer-product of two vectors:

$$G_{ij} = u_i v_j \quad \quad G = uv^T = u \otimes v$$

A rank=1 tensor (n-way array) $G \in R^{d_1 \times d_2 \times \ldots \times d_n}$ is represented by an outer-product of $n$ vectors $u_1, \ldots, u_n$

$$G_{i_1,i_2,\ldots,i_n} = u_{1,i_1} u_{2,i_2} \ldots u_{n,i_n} \quad \quad G = u_1 \otimes u_2 \otimes \ldots \otimes u_n$$
N-way array decompositions

A matrix $G$ is of (at most) rank=$k$ if it can be represented by a sum of $k$ rank-1 matrices:

$$G \in \mathbb{R}^{d_1 \times d_2} \quad U \in \mathbb{R}^{d_1 \times k} \quad V \in \mathbb{R}^{k \times d_2}$$

$$G = \sum_{j=1}^{k} u^j \otimes v^j$$

A tensor $G$ is (at most) rank=$k$ if it can be represented by a sum of $k$ rank-1 tensors:

$$G = \sum_{j=1}^{k} u_1^j \otimes \ldots \otimes u_n^j$$

Example:

$$G = \sum_{j=1}^{k} u^j \otimes v^j \otimes w^j$$

$$A_i = w_1^i (u_1 \otimes v_1) + w_2^i (u_2 \otimes v_2) + \cdots + w_k^i (u_k \otimes v_k)$$
N-way array Symmetric Decompositions

A symmetric rank=1 matrix $G$:

$$G_{ij} = u_i u_j \quad G = uu^\top = u \otimes u$$

A symmetric rank=k matrix $G$:

$$G = \sum_{j=1}^{k} u_j \otimes u_j = \sum_{j=1}^{k} u_j u_j^\top = UU^\top$$

A super-symmetric rank=1 tensor (n-way array) $G \in \mathbb{R}^D$, $D = [m] \times \cdots [m] = [m]^{\times n}$ is represented by an outer-product of $n$ copies of a single vector $u \in \mathbb{R}^m$

$$G_{i_1, \ldots, i_n} = u_{i_1} \cdots u_{i_n} \quad G = u \otimes \cdots \otimes u = u^{\otimes n}$$

A super-symmetric tensor described as sum of $k$ super-symmetric rank=1 tensors:

$$G = \sum_{j=1}^{k} u_j \otimes u_j \otimes \cdots \otimes u_j = \sum_{j=1}^{k} u_j^{\otimes n}$$

is (at most) rank=k.
Let $X_1$ and $X_2$ be two random variables taking values in the sets $[d_i] = \{1,\ldots,d_i\}$.

The statement $X_1$ is independent of $X_2$, denoted by $X_1 \perp X_2$, means:

$$P(X_1,X_2) = P(X_1)P(X_2)$$

$P(X_1,X_2)$ is a 2D array (a matrix) \hspace{1cm} $G_{ij} = P(X_1=i,X_2=j)$

$P(X_1)$ is a 1D array (a vector) \hspace{1cm} $u_i = P(X_1=i)$

$P(X_2)$ is a 1D array (a vector) \hspace{1cm} $v_j = P(X_2=j)$

$X_1 \perp X_2$ means that $G_{ij} = u_i v_j$ is a rank=1 matrix

$G = u \otimes v$
Reduced Rank in Statistics

Let \( X_1, \ldots, X_n \) be random variables taking values in the sets \([d_i] = \{1, \ldots, d_i\}\).

The statement \( X_1 \perp \ldots \perp X_n \) means:

\[
P(X_1, \ldots, X_n) = P(X_1) \cdots P(X_n)
\]

\( P(X_1, \ldots, X_n) \) is a \( n \)-way array (a tensor) \( G_{i_1, \ldots, i_n} = P(X_1 = i_1, \ldots, X_n = i_n) \).

\( P(X_j) \) is a 1D array (a vector) \( \vec{u}_j \) whose entries \( u_{j,i} = P(X_j = i) \).

\( X_1 \perp \ldots \perp X_n \) means that \( G_{i_1, \ldots, i_n} = u_{1,i_1}u_{2,i_2} \cdots u_{n,i_n} \) is a rank=1 tensor

\[
G = \vec{u}_1 \otimes \vec{u}_2 \otimes \cdots \otimes \vec{u}_n
\]
Let $X_1, X_2, X_3$ be three random variables taking values in the sets $[d_i] = \{1, \ldots, d_i\}$.

The conditional independence statement $X_1 \perp X_2 \mid X_3$ means:

$$P(X_1, X_2 \mid X_3 = i) = P(X_1 \mid X_3 = i)P(X_2 \mid X_3 = i)$$

Slice $X_3 = i$ is a rank=1 matrix.
Reduced Rank in Statistics: Latent Class Model

Let $X_1,\ldots,X_n$ be random variables taking values in the sets $[d_i] = \{1,\ldots,d_i\}$.

Let $Y$ be a “hidden” random variable taking values in the set $\{1,\ldots,k\}$.

The “observed” joint probability n-way array is:

$$P(X_1,\ldots,X_n) = \sum_{j=1}^{k} P(X_1,\ldots,X_n, Y = j) = \sum_{j=1}^{k} P(X_1,\ldots,X_n \mid Y = j)P(y = j)$$

A statement of the form $X_1 \perp X_2 \perp \ldots \perp X_n \mid Y$ translates to the algebraic statement

About the n-way array $P(X_1,\ldots,X_n)$ having tensor-rank equal to $k$. 
Reduced Rank in Statistics: Latent Class Model

\[ P(X_1, \ldots, X_n) \] is a n-way array \[ G \geq 0 \quad \|G\|_1 = 1 \]

\[ P(X_i \mid Y = j) \] is a 1D array (a vector) \[ u^i_j \quad \|u^i_j\|_1 = 1 \]

\[ P(X_1, \ldots, X_n \mid Y = j) \] is a rank-1 n-way array \[ \otimes_{i=1}^n u^i_j \]

\[ P(Y) \] is a 1D array (a vector) \[ \sigma \quad \|\sigma\|_1 = 1 \]

\[
\min_{u^i_j, \sigma} \|G - \sum_{j=1}^k \sigma_j \otimes_{i=1}^n u^i_j\|^2 \quad s.t. \quad u^i_j \geq 0, \quad \sigma \geq 0, \quad \|u^i_j\|_1 = 1, \quad \|\sigma\|_1 = 1
\]

reduced to repeated application of “projection onto probability simplex”

\[
\min_x \|x - b\|^2 \quad s.t. \quad x \geq 0, \quad \|x\|_1 = 1
\]
(non-negative) Low Rank Decompositions

\[
\min_{\mathbf{u}_i, \mathbf{\sigma}} \| \mathbf{G} - \sum_{j=1}^{k} \mathbf{\sigma}_j \otimes \mathbf{u}_i \|^2 \quad \text{s.t} \quad \mathbf{u}_i \geq 0, \mathbf{\sigma} \geq 0, \| \mathbf{u}_i \|_1 = 1, \| \mathbf{\sigma} \|_1 = 1
\]

Measurements
= (non-negative) Low Rank Decompositions

The rank-1 blocks tend to represent local parts of the image class

4 factors 8 factors 12 factors 16 factors 20 factors

Hierarchical build-up of “important” parts
(non-negative) Low Rank Decompositions

$$\min_{u_i, \sigma} \left\| G - \sum_{j=1}^{k} \sigma_j \otimes_{i=1}^{n} u^j_i \right\|^2$$

Example: The swimmer

Sample set (256)

Non-negative Tensor Factorization

NMF
• Conditional Independence, tensor-rank, latent class models

• Conditional Independence and Clustering

• High-order affinity clustering (model selection, multiple 3D body segmentation, varying illumination segmentation).
Clustering data into $k$ groups: Pairwise Affinity

$x_1, \ldots x_m \in \mathbb{R}^d$  \hspace{1cm} \text{input points}

$K_{ij} = e^{-\|x_i-x_j\|^2/\sigma^2}$  \hspace{1cm} \text{input (pairwise) affinity value}

interpret $K_{ij}$ as “the probability that $x_i$ and $x_j$ are clustered together”

$y_1, \ldots, y_m \in \{1, \ldots, k\}$  \hspace{1cm} \text{unknown class labels}
Clustering data into $k$ groups: Pairwise Affinity

$x = x_1, \ldots x_m \in \mathbb{R}^d$ \hspace{1cm} \text{input points}

$K_{ij} = e^{-\|x_i - x_j\|^2/\sigma^2}$ \hspace{1cm} \text{input (pairwise) affinity value}

interpret $K_{ij}$ as “the probability that $x_i$ and $x_j$ are clustered together”

$y_1, \ldots, y_m \in \{1, \ldots, k\}$ \hspace{1cm} \text{unknown class labels}

A probabilistic view:

$G_{ij} = P(y_i = j \mid X)$ \hspace{1cm} \text{probability that $x_i$ belongs to the j’th cluster}

note: $G \geq 0$, $G_1 = 1$

What is the (algebraic) relationship between the input matrix $K$ and the desired $G$?
Clustering data into k groups: 
Pairwise Affinity

Assume the following conditional independence statements:

\[ y_1 \perp \cdots \perp y_m \mid x_1, \ldots, x_m \]

\[ K_{ij} = \sum_{r=1}^{k} P(y_i = r, y_j = r \mid X) \]

\[ = \sum_{r=1}^{k} P(y_i = r \mid X)P(y_j = r \mid X) \]

\[ = \sum_{r=1}^{k} G_{ir}G_{jr} \]

\[ K = GG^T, \quad G \geq 0, \quad G1 = 1 \]
A “hard” assignment requires that: 

\[ G^\top G = D = \text{diag}(n_1, \ldots, n_k) \]

where \( n_1, \ldots, n_k \) are the cardinalities of the clusters.

**Proposition**: the feasible set of matrices \( G \) that satisfy \( G \geq 0, \ G_1 = 1, \ G^\top G = D \)

are of the form:

\[ G_{ij} = \begin{cases} 
1 & x_i \in \psi_j \\
0 & \text{otherwise} 
\end{cases} \]

\[
\max_G \text{tr}(G^\top K G) \quad \text{s.t} \quad G \geq 0, \ G_1 = 1, \ G^\top G = D
\]

are equivalent to:

\[
\max_{\psi_1, \psi_2} \sum_{(i,j) \in \psi_1} K_{ij} + \sum_{(i,j) \in \psi_2} K_{ij}
\]

**Min-cut formulation**

\[
\min_{\psi_1, \psi_2} \sum_{i \in \psi_1, j \in \psi_2} K_{ij}
\]
Relation to Spectral Clustering

\[ K = GG^\top, \quad G \geq 0, \quad G1 = 1 \]

Add a “balancing” constraint: \( G^\top 1 = (n/k) 1 \)

put together \( G1 = 1, \quad G^\top 1 = 1 \) means that \( GG^\top 1 = (n/k) 1 \)
also, \( GG^\top = (n/k) 1 \) and \( G^\top G = D \) means that \( D = (n/k) I \)

\[
\max_G \text{tr}(G^\top KG) \quad s.t \quad G \geq 0, \quad GG^\top 1 = \frac{n}{k} 1, \quad G^\top G = \frac{n}{k} I
\]

\( n/k \) can be dropped.

\[
\max_G \text{tr}(G^\top KG) \quad s.t \quad G \geq 0, \quad GG^\top 1 = 1, \quad G^\top G = I
\]

Relax the balancing constraint by replacing \( K \) with the “closest” doubly stochastic matrix \( K' \) and ignore the non-negativity constraint:

\[
\max_G \text{tr}(G^\top K' G) \quad s.t \quad G^\top G = I
\]
Relation to Spectral Clustering

\[
\max_G \text{tr}(G^\top K'G) \; s.t \; G^\top G = I
\]

Let \( D = \text{diag}(K1) \)

Then: \( K' = K - D + I \) is the closest D.S. in L1 error norm. \( \text{Ratio-cuts} \)

\[
K' = D^{-1/2}KD^{-1/2}
\]

\( \text{Normalized-Cuts} \)

Proposition: iterating \( K^{(t+1)} = D^{-1/2}K^{(t)}D^{-1/2} \) with \( D = \text{diag}(K^{(t)}1) \) converges to the closest D.S. in KL-div error measure.
\[
\max_G \text{tr}(G^\top K'G) \quad \text{s.t.} \quad G^\top G = I
\]

where

\[
K' = \text{argmin}_F \|K - F\|_F^2 \quad \text{s.t.} \quad F \geq 0, \; F 1 = 1, \; F = F^\top
\]

we are looking for the closest doubly-stochastic matrix in least-squares sense.
New Normalization for Spectral Clustering

\[ K' = \arg\min_F \|K - F\|_F^2 \text{ s.t. } F \geq 0, \ F1 = 1, \ F = F^\top \]

to find \( K' \) as above, we break this into two subproblems:

\[ P_1(X) = \arg\min_F \|X - F\|_F^2 \text{ s.t. } F1 = 1, \ F = F^\top \]

\[ P_2(X) = \arg\min_F \|X - F\|_F^2 \text{ s.t. } F \geq 0 \]

\[ P_1(X) = \frac{1}{n}I + \frac{1}{n^2}X1 - \frac{1}{n}X \]

\[ P_2(X) = th_{\geq0}(X) \]

use the Von-Neumann successive projection lemma:

\[ K' = P_1P_2P_1P_2...P_1(K) \]
New Normalization for Spectral Clustering

successive-projection vs. QP solver

running time for the three normalizations
### New Normalization for Spectral Clustering

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Kernel</th>
<th>Clusters</th>
<th>Size</th>
<th>Dim.</th>
<th>Lowest Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>L1</td>
</tr>
<tr>
<td>SPECTF heart</td>
<td>RBF</td>
<td>2</td>
<td>349</td>
<td>44</td>
<td>27.5</td>
</tr>
<tr>
<td>Pima Indians Diabetes</td>
<td>RBF</td>
<td>2</td>
<td>768</td>
<td>8</td>
<td>36.2</td>
</tr>
<tr>
<td>Wine</td>
<td>RBF</td>
<td>3</td>
<td>178</td>
<td>13</td>
<td>38.8</td>
</tr>
<tr>
<td>SpamBase</td>
<td>RBF</td>
<td>2</td>
<td>4601</td>
<td>57</td>
<td>36.1</td>
</tr>
<tr>
<td>BUPA liver disorders</td>
<td>Poly</td>
<td>2</td>
<td>345</td>
<td>6</td>
<td>37.4</td>
</tr>
<tr>
<td>WDBC</td>
<td>Poly</td>
<td>2</td>
<td>569</td>
<td>30</td>
<td>18.8</td>
</tr>
</tbody>
</table>

**UCI Data-sets**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Kernel</th>
<th>Clusters</th>
<th>Size</th>
<th>Dim.</th>
<th>Lowest Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>L1</td>
</tr>
<tr>
<td>AML/ALL Leukemia</td>
<td>Poly</td>
<td>2</td>
<td>72</td>
<td>5</td>
<td>27.8</td>
</tr>
<tr>
<td>Lung</td>
<td>Poly</td>
<td>2</td>
<td>181</td>
<td>5</td>
<td>15.5</td>
</tr>
<tr>
<td>Prostate</td>
<td>RBF</td>
<td>2</td>
<td>136</td>
<td>5</td>
<td>40.4</td>
</tr>
<tr>
<td>Prostate Outcome</td>
<td>RBF</td>
<td>2</td>
<td>21</td>
<td>5</td>
<td>28.6</td>
</tr>
</tbody>
</table>

**Cancer Data-sets**

Hebrew University
• Conditional Independence, tensor-rank, latent class models

• Conditional Independence and Clustering

• High-order affinity clustering (model selection, multiple 3D body segmentation, varying illumination segmentation).
Clustering data into k groups: Beyond Pairwise Affinity

A model selection problem that is determined by n-1 points can be described by a factorization problem of n-way array (tensor).

Example: clustering m points into k lines

\[ D = \{x_1, \ldots, x_m\} \]

**Input:** \( K_{i_1,i_2,i_3} \) = the probability that \( x_{i_1}, x_{i_2}, x_{i_3} \) belong to the same model (line).

**Output:** \( g_{rs} = Pr(y_s = r \mid D) \) the probability that the point \( x_s \) belongs to the r’th model (line)

Under the independence assumption:

\( y_1 \perp \ldots \perp y_m \mid x_1, \ldots, x_m \)

\[
    P(y_{i_1} = r, y_{i_2} = r, y_{i_3} = r \mid D) = P(y_{i_1} = r \mid D)P(y_{i_2} = r \mid D)P(y_{i_3} = r \mid D)
\]

\[
    K_{i_1,i_2,i_3} = \sum_{r=1}^{4} P(y_{i_1} = r \mid D)P(y_{i_2} = r \mid D)P(y_{i_3} = r \mid D) = \sum_{r=1}^{4} g_{r,i_1}g_{r,i_2}g_{r,i_3}
\]

\( K = \sum_{r=1}^{K} g_r \otimes g_r \otimes g_r \) is a 3-dimensional super-symmetric tensor of rank=4

\[ G = [g_1, \ldots, g_k] \]
Clustering data into \( k \) groups: Beyond Pairwise Affinity

**General setting**: clusters are defined by \( n-1 \) dim subspaces, then for each \( n \)-tuple of points \( x_{i_1}, \ldots, x_{i_n} \) we define an affinity value \( K_{i_1,\ldots,i_n} = e^{-\Delta} \) where \( \Delta \) is the volume defined by the \( n \)-tuple.

**Input**: \( K_{i_1,\ldots,i_n} \) the probability that \( x_{i_1}, \ldots, x_{i_n} \) belong to the same cluster

**Output**: \( g_{rs} = P(y_s = r \mid D) \) the probability that the point \( x_s \) belongs to the \( r \)'th cluster

Assume the conditional independence: \( y_1 \perp \ldots \perp y_m \mid x_1, \ldots, x_m \)

\[
K_{i_1,\ldots,i_n} = \sum_{r=1}^{k} P(y_{i_1} = r \mid D) \cdots P(y_{i_n} = r \mid D) = \sum_{r=1}^{k} g_{r,i_1} \cdots g_{r,i_n}
\]

\[\rightarrow \quad K = \sum_{r=1}^{k} g_r^{\otimes n}\]

is a \( n \)-dimensional super-symmetric tensor of rank=\( k \)
**Hyper-stochastic constraint:** under balancing requirement

K is (scaled) hyper-stochastic:

\[
\sum_{i_1, \ldots, i_{j-1}, i_{j+1}, \ldots, i_n} K_{i_1, \ldots, i_n} = \left(\frac{m}{k}\right)^{n-1} 1, \quad j = 1, \ldots, n
\]

**Theorem:** for any non-negative super-symmetric tensor \(K^{(0)}\), iterating

\[
K_{i_1, \ldots, i_n}^{(t+1)} = \frac{K_{i_1, \ldots, i_n}^{(t)}}{(a_{i_1} \cdots a_{i_n})^{1/n}}
\]

\[
a_i = \sum_{i_2, \ldots, i_n} K_{i,i_2, \ldots, i_n}, \quad i = 1, \ldots, m
\]

converges to a hyper-stochastic tensor.
Example: **multi-body segmentation**

9-way array, each entry contains
The probability that a choice of 9-tuple of points arise from the same model.

\[ p^T F p = 0 \]

Probability:

\[ e^{-(p^T F p)^2} \]
Model Selection

Example: visual recognition under changing illumination

4-way array, each entry contains
The probability that a choice of 4 images
Live in a 3D subspace.
Previous Work

- In VLSI / PCB Placement - since early 70s
  - Hyper-vertex swapping and multi-level paradigms (heuristic in nature).

- Recent works in vision and learning communities, seeks an approximation as a pairwise problem:

  \[
  A_{i_1,i_2,i_3} = \sum_{i_2,...,i_n} A_{r,i_2,...,i_n} A_{s,i_2,...,i_n}
  \]

- Govindu 2005:

  \[
  A_{r,s} = \sum_{i_2,...,i_n} A_{r,i_2,...,i_n} A_{s,i_2,...,i_n}
  \]

- When the affinity dimension \( n \) is high or when the number of points per cluster is low, the averaging decreases the SNR, and the projection result becomes less informative.
200 points
Arranged in five $3^{\text{rd}}$ order curves (4 coefficients), with added Gaussian noise.
Induces a 5-way affinity tensor.
Error Rate vs. Sampling

- Projection with Normalized Cuts
- SNTF
Error Rate vs. Sigma

\[ K_{i_1, \ldots, i_n} = \frac{e^{\Delta^2}}{\sigma^2} \]

Projection with Normalized Cuts

SNTF
Revisiting Latent Class Model

\[ P(X_1, \ldots, X_n) \] is a n-way array \[ \|G\|_1 = 1 \]

\[ P(X_i \mid Y = j) \] is a 1D array (a vector) \[ u^i_j \| u^i_j \|_1 = 1 \]

\[ P(X_1, \ldots, X_n \mid Y = j) \] is a rank-1 n-way array \[ \otimes_{i=1}^n u^i_j \]

\[ P(Y) \] is a 1D array (a vector) \[ \sigma \| \sigma \|_1 = 1 \]

\[ \min_{u^i_j, \sigma} \| G - \sum_{j=1}^{k} \sigma_j \otimes_{i=1}^n u^i_j \|^2 \quad \text{s.t. } u^i_j \geq 0, \sigma \geq 0, \| u^i_j \|_1 = 1, \| \sigma \|_1 = 1 \]

Why use the Least-Squares (Frobenius) error?
Revisiting Latent Class Model

The Relative Entropy: 
\[ D(p\|q) = \sum_i p_i \log \frac{p_i}{q_i} \]

\[
\min_{u_i^j, \sigma} D(G \| \sum_j \sigma_j \otimes_i u_i^j) \quad s.t \quad u_i^j \geq 0, \sigma \geq 0, \|u_i^j\|_1 = 1, \|\sigma\|_1 = 1
\]

is the Maximum-Likelihood solution.
The Expectation-Maximization algorithm, for example, introduces auxiliary tensors \( W_j \) and alternates among the three sets of variables: \( W_j, \sigma, u^j_i \)

\[
\min_{u^j_i, \sigma, W_j} \sum_{j=1}^{k} D(W_j \odot G \mid \sigma_j \otimes u^j_i) \quad s.t \quad u^j_i \geq 0, \sigma \geq 0, \|u^j_i\|_1 = 1, \|\sigma\|_1 = 1
\]

\( W_j \geq 0, \sum_j W_j = 1 \)

\((A \circ B)_i = A_i B_i \) Hadamard product.

reduces to repeated application of “projection onto probability simplex”

\[
\min_x D(x \mid b) \quad s.t. \quad x \geq 0, \|x\|_1 = 1
\]

This, however, is known to be very sensitive to additive noise!

Hebrew University
Revisiting Latent Class Model

Frobenius error behaves quite well under additive noise:

\[ S = \{ Q : Q \geq 0, \| Q \|_1 = 1, \text{rank}(Q) = k \} \] probability tensors of rank=k

Let \( E \) be a perturbation tensor with bounded infinity-norm \( \| E \|_\infty = \varepsilon \)

**Proposition:** let \( P \in S \) be a rank=k probability tensor and let

\[ Q^* = \arg\min_{Q \in S} \| P + E - Q \|_F^2 \]

then

\[ \| P - Q^* \|_\infty \leq 2\sqrt{\varepsilon} \]
The noise-resilience has to do with allowing solutions on the boundary:

\[ x^* = \frac{1}{\|b\|_1} b \]
conditional independence $\rightarrow$ rank-1 tensor slices constraint

Latent class model $\rightarrow$ low rank NTF

sparse image coding $\rightarrow$ low rank super-symmetric NTF

clustering/model-selection is associated with a graphical model of conditional independence constraints
END