Reminder

- As there are only 7 projects
- Project presentation all NEXT Tuesday
- 10 minutes each: 7 presentation, 3 question, will be timed
Review of Last Class
Singular Value Decomposition (SVD)

\[ X = U \Sigma V^T \]

- \( X(i,j) \) = value of user \( i \) for property \( j \)
  - \( X(\text{Alice, cholesterol}) = 10 \)
  - \( X(i,j) \) = number of times \( i \) buys \( j \)
  - \( X(i,j) \) = how much \( i \) pays \( j \)
  - \( X(i,j) \) = 1 if \( i \) and \( j \) are friends, 0 otherwise
  - \( X(i,j) \) = temperature of sensor \( j \) at time \( i \)
SVD Definition

- SVD gives best rank-k approximation of $X$ in $L_2$ and Frobenius norm

$$X = U \Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i \otimes v_i$$

Outer product
Tensor Decomposition (Kruskal Tensor)

\[ \chi = a_1 b_1 c_1 + a_2 b_2 c_2 + \ldots \]
Tucker Tensor Decomposition (Tucker 1966)

\[ X = I \times J \times K = A \times r \times r \times s \times t = B \times K \times s \times t \]
This Class

Probabilistic Interpretation
Two Topics

- Generalized Tensor Factorizations
- Spectral Methods for Latent Variable Models a.k.a. Method of Moments
Generalized Tensor Factorizations
Probabilistic Interpretation of TF

\[
\arg \min_{\sigma, A, B, C} \left\| \mathbf{X} - \sum_r \sigma_r a_r \otimes b_r \otimes c_r \right\|^2
\]

- Equivalent to CP with Euclidean error minimization:

\[
\Lambda(i, j, k) = \sum_r a_r(i)b_r(j)c_r(k)\sigma_r
\]

\[
\mathbf{X}(i, j, k) \sim \text{Normal}(\Lambda(i, j, k), 1)
\]

- Equivalent to CP with KL divergence error minimization

\[
\Lambda(i, j, k) = \sum_r a_r(i)b_r(j)c_r(k)\sigma_r
\]

\[
\mathbf{X}(i, j, k) \sim \text{Poisson}(\Lambda(i, j, k))
\]

Exponential R.V. gives CP with Itakura-Saito-Divergence

Non-negative tensor factorization
Coupled Tensors (and Matrices)

- In real life we often have information from multiple data sources
- What to do?
- Really, what should we do?

![Diagram of coupled tensors](Figure credit: Ermis et al. 2013)

Figure 1: UCLAF dataset represented in the form of a third-order tensor coupled with two matrices in two different modes.

- UCLAF dataset: Extracted from the GPS data that includes information of three types of entities: user, location, and activity (see Fig. 1 for an illustration of the data). The relations between user–location–activity triplets are used to construct a three-way tensor $X_1$.

In tensor $X_1$, an entry $X_1(i, j, k)$ indicates the frequency of user $i$ visiting location $j$ and doing activity $k$ there; otherwise, it is 0. Since we address the link prediction problem in this study, we define the user–location–activity tensor $X_1$ as:

$$X_1(i, j, k) = \begin{cases} 1 & \text{if user } i \text{ visits location } j \text{ and performs activity } k \text{ there}, \\ 0 & \text{otherwise}. \end{cases}$$

2. http://www.public.esu.edu/~ylin56/kdd09sup.html
Coupled Factorization

\[ X_1^{i,j,k} \sim \text{Normal}(\sum A(i,r)B(j,r)C(k,r)\sigma_r, 1) \]
\[ X_2^{i,h} \sim \text{Normal}(\sum A(i,r)F(h,r)\sigma_r, 1) \]
\[ X_3^{k,m} \sim \text{Normal}(\sum_r^r C(k,r)G(m,r)\sigma_r, 1) \]

\[ A = \text{Latent groups of users} \]
\[ B = \text{Latent groups of activities} \]
\[ C = \text{Latent groups of visited locations} \]
\[ F = \text{Latent groups of home locations} \]
\[ G = \text{Latent groups of features} \]
Solution

- Via Fisher scoring the update rule is:

\[
\tilde{z}_\alpha \leftarrow z_\alpha + \left(\nabla_{\alpha}^T D \nabla_{\alpha}\right)^{-1} \nabla_{\alpha}^T D (\text{vec}(X) - \text{vec}(\hat{X}))
\]
Two Topics

- Generalized Tensor Factorizations
- Spectral Methods for Latent Variable Models a.k.a. Method of Moments
Spectral Methods for Latent Variable Models
One-dimensional Method of Moments (Pearson, 1894)

- Suppose we want to estimate parameters of uniform,

\[ f(x) = \frac{1}{b - a}, \quad a < x < b \]

- Moments are:

\[ E[X] = \frac{(a + b)}{2} \]

\[ E[X^2] = \frac{a^2 + ab + b^2}{3} \]

- Solving,

\[ E[X^2] = \frac{a^2 + a(2E[X] - a) + (2E[X] - a)^2}{3} \]

- For Uniform just first and second moment enough to identify parameters
Method of Moments Worse Than MLE

- Method of moments parameter estimates are, in theory, worst than MLE parameter estimates
Observations with Related Latent Structures

- E.g. Topic Models
  - $k$ topics/word distributions: $\mu_1, \ldots, \mu_k$
  - sample topic $H = i$ with prob. $w_i$
  - observe $m$ (exchangeable) words
    - $x_1, x_2, \ldots, x_m$ sampled i.i.d. from $\mu_i$
  - dataset: multiple points / $m$-word documents
  - how to learn parameters? $\mu_1, \ldots, \mu_k, w_1, \ldots, w_k$

ack: S. M. Kanade’14
Vector notation

- binary word encoding: \( x_1 = [0, 1, 0, \ldots]^{\top} \)
- \( \mu_i \)'s are probability vectors

- for each word, the conditional probabilities are:

\[
P[x_1|\text{topic } i] = E[x_1|\text{topic } i] = \mu_i
\]
Method of Moments

- Topic model moments:

\[ P[x_1], P[x_1,x_2], P[x_1,x_2,x_3], \ldots \]

- Identifiability problem: if we had the exact moments, how many order moments do we need to identify parameters?
  - how many words per document suffice?
First Moment Enough?

- Yes, no?

- I.e. with ONE word per document can we identify topics?

\[ P[x_1] = \sum_{i=1}^{k} w_i \mu_i \]

- There are \( d \) documents

A: not identifiable … only \( d \) equations for \( mk \) words x topics
Second Moment Enough?

- by exchangeability:

\[
P[x_1, x_2] = E[E[x_1|\text{topic } i] \otimes E[x_2|\text{topic } i]]
= \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i
\]

- What is the maximum rank of this matrix?
  - What is the maximum rank of this matrix?
  - Rank \( \leq k \)
  - Matrix is \( m \times m \)
  - If \( k < m \) then many rows are linearly dependent
What about 3\textsuperscript{rd} Moment?

\[
P[x_1, x_2] = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i
\]

\[
P[x_1, x_2, x_3] = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i
\]

Tensor
Whitening (Anandkumar et al. 2012)

- 2nd moment matrix has rank $k$: make orthonormal basis

\[ \tilde{M}_2 = I = \sum_{i=1}^{k} \tilde{w}_i \tilde{\mu}_i \otimes \tilde{\mu}_i \]

- Now the tensor is the only useful information

\[ \tilde{M}_3 = \sum_{i=1}^{k} \tilde{w}_i \tilde{\mu}_i \otimes \tilde{\mu}_i \otimes \tilde{\mu}_i \]

- Unique solution?
  - yes: If $k < d$ + generic params (Kruskal (1977) (Anandkumar et al. 2013))
  - $k > d$ case in (Lathauwer, Castaing, & Cardoso (2007))
E.g.: Latent Dirichlet Allocation

- (Anandkumar et al. 2012)
  
  Prior for topic mixture $\pi$:

  \[ p(\alpha) = \frac{1}{c} \prod_{i=1}^{k} \pi_i^{\alpha_i - 1}, \quad \alpha^T \mathbf{1} = 1 \]

- Implies:

  \[ M_2 := E[x_1 \otimes x_2] - \frac{\sum_j \alpha_j}{\sum_j \alpha_j + 1} E[x_1] \otimes E[x_1] \]

  \[ M_3 := E[x_1 \otimes x_2 \otimes x_3] - \frac{\sum_j \alpha_j}{\sum_j \alpha_j + 2} E[x_1 \otimes x_2 \otimes E[x_1]] - \ldots \]

  Then

  \[ M_2 = \sum_i \tilde{\omega}_i \mu_i \otimes \mu_i \]

  \[ M_3 = \sum_i \tilde{\omega}_i \mu_i \otimes \mu_i \otimes \mu_i \]

ack: S. M. Kanade'14