Tensor Basics

But First…
Mining Matrices
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 Dimensions

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

Data \quad \text{Left singular vectors} \quad \text{Right singular vectors}
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
SVD Properties (I)

◦ “Almost unique” decomposition

\[ X = \begin{pmatrix} \sigma_1 & v_1 \\ u_1 \end{pmatrix} + \begin{pmatrix} \sigma_2 & v_2 \\ u_2 \end{pmatrix} + \ldots \]

◦ There are two sources of ambiguity
  ◦ Orientation of singular vectors
    • Permute rows of left singular vector and corresponding rows of left singular vector
  ◦ If I is identity matrix: \( I = U U^T \), for all orthonormal \( U \)
    • “Hypersphere ambiguity”
    • Related to rotational ambiguity of PCA
SVD Properties (II)

- Theorem (Eckart-Young, 1936)
  - $U\Sigma_1 V^T$ is best rank 1 approximation of $X$, that is $|X - U\Sigma_1 V^T|^2 \leq |X - Y|^2$ for every rank 1 matrix $Y$

  - $U\Sigma_1 V^T + U\Sigma_2 V^T$ is the best rank 2 approximation of $X$, that is $|X - U\Sigma_1 V^T - U\Sigma_2 V^T|^2 \leq |X - Y|^2$ for every rank $\leq 2$ matrix $Y$

  - also for $3, 4, \ldots, r$
SVD Properties (III)

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

\[ U^T U = \begin{bmatrix} 1 & \vdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \vdots & 1 \end{bmatrix} \]

\[ V^T V = \begin{bmatrix} 1 & \vdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \vdots & 1 \end{bmatrix} \]

U and V are orthogonal
Understanding SVD singular vectors

- As $U$ and $V$ have orthogonal rows

\[
X^T X = (U \Sigma V^T)^T (U \Sigma V^T) = V \Sigma^2 V^T
\]

\[
X X^T = (U \Sigma V^T)(U \Sigma V^T)^T = U \Sigma^2 U^T
\]

Now you explain: What do $V$ and $U$ represent?
If $X(i,j) = \text{user i buys product j}$

What is $X^TX$?
- Product-to-product similarity matrix
- What does $V$ represent?

$$X^TX = (U\Sigma V^T)^T(U\Sigma V^T) = V\Sigma^2 V^T$$

What is $XX^T$?
- User-to-user similarity matrix
- What does $U$ represent?

$$XX^T = (U\Sigma V^T)(U\Sigma V^T)^T = U\Sigma^2 U^T$$
Suggestion for SVD > 2 dimensions?

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

- Simple extension:

\[ \mathcal{X} = \sum_{i} \sigma_i a_i \otimes b_i \otimes c_i \]

- Representation of \( \mathcal{X} \)?
Tensor Motivation
App: Social Network Analysis

- Traditional focus on static networks and find community structures
- Tensors can monitor change of community structure over time
See word ("apple")

Answer fundamental human question ("Is it edible?")

How do different parts of your brain communicate in the meanwhile?

Functional Connectivity

Ack: Papalexakis et al. SDM 2014
Tensor Decomposition (Kruskal Tensor)

\[
\chi = a_1 b_1 \sigma_1 + a_2 b_2 \sigma_2 + \ldots
\]
Tensor Definitions

- Krooneberg 1983

Figure 2  Slices, the two-way submatrices of $X$

Figure 3  Fibers, the one-way submatrices of $X$
PARAFAC Decomposition (Harshman 1970)

- Parallel Factors Decomposition
- Columns of A, B, C not orthogonal
- If has r factors & r is minimal, then \( \text{rank}(\mathbf{X}) = r \)
- Important: possible that \( \text{rank}(\mathbf{X}) > \min(I,J,K) \)
- Decomposition often unique*
  - *(Kruskal 1977): A, B, C are unique up to rescaling / counterscaling and joint permutations if
    \[
    2r + 2 \leq k_A + k_B + k_C
    \]
    where \( k_A \) is k-th rank of A: max number \( k_A \) such that every set of \( k \) columns of A are linearly independent

© 2015 Bruno Ribeiro
Tucker Tensor Decomposition (Tucker 1966)

\[ \chi_{I \times J \times K} = A_{I \times r} \times G_{r \times s \times t} \times B_{K \times s} \]
Tucker Decomposition Interpretation

- user x product x time
- \( A = \) user x user factors
- \( B = \) products x products factors
- \( C = \) time x time factors
- \( G \) = how these groups are mixed

Common Properties:
- A, B, C orthonormal
- \( G \) is not diagonal
- Decomposition not unique
Useful “Rearrangements”: Matricize tensor

Vectorization

\[ \text{vec}(\mathbf{X}) = \]
Products

- 3-way outer product
  \[ \mathbf{X} = a \otimes b \otimes c \]

- Kronecker product (generalization of outer product)
  \[ \mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11} \mathbf{B} & \cdots & a_{1n} \mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1} \mathbf{B} & \cdots & a_{mn} \mathbf{B} \end{bmatrix}, \text{ wikipedia} \]

- Khatri-Rao product (same elements but as a thinner matrix)
  \[ \mathbf{A} \odot \mathbf{B} = \begin{bmatrix} a_1 \otimes b_1 & a_2 \otimes b_2 & \cdots & a_r \otimes b_r \end{bmatrix} \]
  \[ M \times r \quad N \times r \quad MN \times r \]
Solving Tucker:

- Tucker Decomposition in Matrix form:
  \[
  X_{(1)} = AG_{(1)}(C \otimes B) \\
  X_{(2)} = BG_{(2)}(C \otimes A) \\
  X_{(3)} = CG_{(3)}(B \otimes A) \\
  \text{vec}(X) = (A \otimes B \otimes C) \text{vec}(G)
  \]

- Error is then:
  \[
  \left\| \chi - \sum_{r} g_{rst} a_r \otimes b_s \otimes w_t \right\|^2 = \| \chi \|^2 - \| \text{vec} ((C^T \otimes B^T \otimes A^T) \text{vec}(\chi)) \|^2
  \]

Minimize = Maximize \\
\text{s.t. } A, B, C \text{ orthonormal}
Solving Tucker:

\[
\begin{aligned}
\arg\min_{\mathcal{G}, A, B, C} \left\| \mathcal{X} - \sum_r g_{rst} a_r \otimes b_s \otimes w_t \right\|^2
\end{aligned}
\]

- A, B, C have to be orthonormal

- Ideas?
  - Minimize \( \|X_1 - A G_1 (C \otimes B)\|^2 \)
  - This minimization can be reduced by solving first for \( G \) as
    \[ G^* = A^T X_1 (C^T \otimes B^T), \]
    and substituting \( G^* \) into the loss function to obtain
    \[ \|X_1 - A^T A X_1 (C^T \otimes B^T)(C \otimes B)\|^2 \]
A simpler way to write it

$$
\left\| \chi - \sum_{r} g_{rst} a_r \otimes b_s \otimes w_t \right\|^2 = \| \chi \|^2 - \| \text{vec} \left( (C^T \otimes B^T \otimes A^T) \text{vec}(\chi) \right) \|^2
$$

$$
= \| \chi \|^2 - \| A^T X_{(1)}(C \otimes B) \|^2
$$

$$
= \| \chi \|^2 - \| B^T X_{(2)}(C \otimes A) \|^2
$$

$$
= \| \chi \|^2 - \| C^T X_{(3)}(B \otimes A) \|^2
$$
Solving Tucker (III)

- If B and C were fixed we could solve for A:
  \[
  \arg\max_A \| A^T X^{(1)} (C \otimes B) \|^2
  \]

- A has to be orthonormal

- Algorithm?
  - Optimal A from leading r left singular vectors of \( X^{(1)} (C \otimes B) \)
  - Same for B and C (Alternating Least Squares)

- Algorithm (at step \( k \)): (Kroonenberg & De Leeuw 1980)
  - \( A_k = r \) leading left singular vectors of \( X^{(1)} (C_{k-1} \otimes B_{k-1}) \)
  - \( B_k = s \) leading left singular vectors of \( X^{(2)} (C_{k-1} \otimes A_{k-1}) \)
  - \( C_k = t \) leading left singular vectors of \( X^{(3)} (B_{k-1} \otimes A_{k-1}) \)

- Final step (\( K \)): get \( \mathcal{G} = A_K^T X^{(1)} (C_K \otimes B_K) \)

Not unique as \( EE^T = I \) can be inserted before \( X^{(\cdot)} \)

© 2015 Bruno Ribeiro
Solving PARAFAC:

- Same idea: Alternating Least Squares
  \[ X_{(1)} = A\Lambda (C \odot B)^T \]
- With Fixed B and C Solve
  \[ A = X_{(1)}((C \odot B)^T)^\dagger \Lambda^{-1} \]
- Noting that
  \[ (C \odot B)^\dagger = (C^T C \odot B^T B)^\dagger (C \odot B)^T \]
Next Class

- Probabilistic interpretation
- Non-negative tensor factorization
- How tensors can be used for mixture model inference, topic modeling, etc.
Acknowledgments: Faloutsos, Kolda, Sun tutorial @ ICML'07