Biomedical Data Mining
with Matrix Models

SDM 2016 Tutorial Part I
May 5, 2016

Fei Wang
Associate Professor
Department of Computer Science and Engineering
School of Engineering
University of Connecticut
feiwang03@gmail.com
Biomedical Data
Electronic Health Record (EHR) is an evolving concept defined as a systematic collection of electronic health information about individual patients or populations.

Medical Imaging

X-ray Computed Tomography (CT)

Positron Emission Tomography (PET)

Magnetic Resonance Imaging (MRI)
A chemical compound is a pure chemical substance consisting of two or more different chemical elements that can be separated into simpler substances by chemical reactions.
The term biological target is frequently used in pharmaceutical research to describe the native protein in the body whose activity is modified by a drug resulting in a desirable therapeutic effect. In this context, the biological target is often referred to as a drug target.

http://www.uniprot.org/help/uniprotkb
Gene

**DNA:** A long molecule that looks like a twisted ladder. It is made of four types of simple units and the sequence of these units carries information, just as the sequence of letters carries information on a page.

**Gene:** A segment of DNA. Genes are like sentences made of the "letters" of the nucleotide alphabet, between them genes direct the physical development and behavior of an organism. Genes are like a recipe or instruction book, providing information that an organism needs so it can build or do something - like making an eye or a leg, or repairing a wound.

**Gene expression:** The process in which the information encoded in a gene is converted into a form useful for the cell. The first step is transcription, which produces a messenger RNA molecule complementary to the DNA molecule on which a gene is encoded. For protein-coding genes, the second step is translation, in which the messenger RNA is read by the ribosome to produce a protein.

**A Single Nucleotide Polymorphism (SNP, pronounced snip; plural snips) is a DNA sequence variation occurring commonly within a population (e.g. 1%) in which a single nucleotide — A, T, C or G — in the genome (or other shared sequence) differs between members of a biological species or paired chromosomes.**
Physiology Data

Physiology is the scientific study of function in living systems. A sub-discipline of biology, its focus is in how organisms, organ systems, organs, cells, and bio-molecules carry out the chemical or physical functions that exist in a living system.
Patient Survey

Church Health Survey™

SURVEY QUESTIONS continued

(Answer on separate response form. SA = Strongly Agree, A = Agree, U = Undecided, D = Disagree, SD = Strongly Disagree.)

73. Our worship services are prayerful experiences.
74. Our church does a very good job at discipling members.
75. I pray for my pastor / minister / clergy, their family(ies), and the staff of our church regularly.
76. I regularly attempt to establish relationships with people who do not have a church home.
77. I believe my voice is heard in this church.
78. Most of the ministry in our church is done by a small number of people.
79. We have good participation in serving in our worship services.
80. We have many people involved in one-on-one discipling in our church.

81. We have a place in our church where many people pray each week.
82. I believe you have to be called to be a lifetime missionary to do international missions.
83. I am excited to be a part of this church.
84. Our church is strong in benevolence ministry.
85. I have been to other churches, and I think our worship services are among the best.
86. The lifestyles of the members at our church are significantly different than the world’s lifestyle.
87. I believe the Bible teaches that prayer should be a daily part of every Christian’s life.
88. I believe that one of the purposes of Sunday school

Mental Health Survey

Patient Name: [Name]
Therapist Name: [Name]
Date Completed: [Date]
Date Received: [Date]

Mood and Behaviors Over the Past 2 Weeks

<table>
<thead>
<tr>
<th>Mood</th>
<th>Never</th>
<th>Rarely</th>
<th>Occasionally</th>
<th>Mostly</th>
<th>Always</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I feel sad, upset or depressed</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>2. I feel forgetful, forgetful, or as if I have no energy</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>3. I feel lonely, isolated or alone</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>4. I have trouble sleeping</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>5. I feel that I am running out of energy</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6. I feel sad, upset or depressed</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>7. I feel forgetful, forgetful, or as if I have no energy</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>8. I feel sad, upset or depressed</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>9. I feel forgetful, forgetful, or as if I have no energy</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

Please answer questions 6 to 19 as follows:
0 = never, 1 = hardly ever, 2 = occasionally, 3 = fairly often, 4 = very often

IN THE PAST 12 MONTHS HAVE YOU........

Question 4. Had trouble pronouncing any words because of problems with your teeth or dentures?

Question 5. Felt that your sense of taste has worsened because of problems with your teeth or dentures?

Question 6. Had any painful aching in your mouth?

Question 7. Found it uncomfortable to eat any foods because of problems with your teeth or dentures?

Question 8. Been self-conscious because of problems with your teeth or dentures?

Question 9. Felt tense because of problems with your teeth or dentures?

Question 10. Has your diet been unsatisfactory because of problems with your teeth, mouth or dentures? (Question relates to function - not about quality of food in prisons)

Question 11. Had to interrupt meals because of problems with your teeth or dentures?

Question 12. Found it difficult to relax because of problems with your teeth or dentures?

Question 13. Been a bit embarrassed because of problems with your teeth or dentures?

Question 14. Been a bit irritable with other people because of problems with your teeth or dentures?

Question 15. Had difficulty doing your usual activity because of problems with your teeth or dentures?

Question 16. Felt that life in general is less satisfying because of problems with your teeth or dentures?

Question 17. Been totally unable to function because of problems with your teeth or dentures?

If you have lost some, or all your natural teeth, we would like you to answer the following question.

Question 18. Have you ever had any kind of denture? (False teeth which you can take out) (Please tick one box)

IF YES, PLEASE ANSWER THE FOLLOWING, IF NO, PLEASE GO TO QUESTION 22.

Question 19. What type of denture do you have?

(If yes, please answer the following, if no, please go to question 22.)

- Full TOP denture.
- Full BOTTOM denture.
- TOP part denture.
- BOTTOM part denture.
- TOP and BOTTOM denture.

IF YES PLEASE ANSWER THE FOLLOWING, IF NO, PLEASE GO TO QUESTION 22.

Question 19b. What type of denture do you have?

(If yes, please answer the following, if no, please go to question 22.)

- Full TOP denture.
- Full BOTTOM denture.
- TOP part denture.
- BOTTOM part denture.
- TOP and BOTTOM denture.
Online Social Media
Environmental Data
1. Clinical questions
2. Insights from the data
3. Predictive models
Matrix?
Vector Collection

$\mathbf{x} = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & \cdots & x_N \end{bmatrix}$

Feature

$x_1$

$x_2$

$x_3$

$x_4$

\vdots

\vdots

\vdots

x_N$
Graph

Matrix Methods
Spectrum Analysis
Spectral Clustering

• Algorithms that cluster points using eigenvectors of matrices derived from the data

• Obtain data representation in the low-dimensional space that can be easily clustered

• Variety of methods that use the eigenvectors differently

• Difficult to understand….
Graph Theory Basics

• A graph $G = (V,E)$ consists of a vertex set $V$ and an edge set $E$.

• If $G$ is a directed graph, each edge is an ordered pair of vertices.

• A bipartite graph is one in which the vertices can be divided into two groups, so that all edges join vertices in different groups.
Similarity Graph

- Distance decrease similarity increase
- Represent dataset as a weighted graph $G(V,E)$
- $W_{ij}$ represent similarity between vertex
- If $W_{ij}=0$ where isn’t similarity; $W_{ii}=0$

$V=\{x_i\}$: Set of $n$ vertices representing data points

$E=\{W_{ij}\}$: Set of weighted edges indicating pair-wise similarity between points
Graph Partitioning

• Clustering can be viewed as partitioning a similarity graph

• Bi-partitioning task:
  • Divide vertices into two disjoint groups (A,B)
Partitioning Criterion

• Traditional definition of a “good” clustering:
  • Points assigned to same cluster should be highly similar.
  • Points assigned to different clusters should be highly dissimilar.
Graph Cut

• Express partitioning objectives as a function of the “edge cut” of the partition.
• Cut: Set of edges with only one vertex in a group. We want to find the minimal cut between groups. The groups that have the minimal cut would be the partition
Min-Cut

Minimize weight of connections between groups $\min_{A,B} \text{Cut}(A,B)$
Normalized Cut

- Consider the connectivity between groups relative to the density of each group.
- Normalize the association between groups by volume.
  - \( \text{Vol}(A) \): The total weight of the edges originating from group A.
- Why use this criterion?
  - Minimizing the normalized cut is equivalent to maximizing normalized association.
  - Produces more balanced partitions.

\[
\min_{A,B} NCut(A, B) = \frac{Cut(A, B)}{\text{vol}(A)} + \frac{Cut(A, B)}{\text{vol}(B)}
\]
Spectral Graph Theory

• Possible approach
  • Represent a similarity graph as a matrix
  • Apply knowledge from Linear Algebra...

• The eigenvalues and eigenvectors of a matrix provide global information about its structure.

• Spectral Graph Theory
  • Analyze the “spectrum” of matrix representing a graph.
  • Spectrum : The eigenvectors of a graph, ordered by the magnitude(strength) of their corresponding eigenvalues
Matrix Representation

\[
\begin{array}{cccccc}
 x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\
 x_1 & 0 & 0.8 & 0.6 & 0 & 0.1 & 0 \\
 x_2 & 0.8 & 0 & 0.8 & 0 & 0 & 0 \\
 x_3 & 0.6 & 0.8 & 0 & 0.2 & 0 & 0 \\
 x_4 & 0 & 0 & 0.2 & 0 & 0.8 & 0.7 \\
 x_5 & 0.1 & 0 & 0 & 0.8 & 0 & 0.8 \\
 x_6 & 0 & 0 & 0 & 0.7 & 0.8 & 0 \\
\end{array}
\]

\[
\begin{array}{cccccc}
 x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\
 x_1 & 1.5 & 0 & 0 & 0 & 0 & 0 \\
 x_2 & 0 & 1.6 & 0 & 0 & 0 & 0 \\
 x_3 & 0 & 0 & 1.6 & 0 & 0 & 0 \\
 x_4 & 0 & 0 & 0 & 1.7 & 0 & 0 \\
 x_5 & 0 & 0 & 0 & 0 & 1.7 & 0 \\
 x_6 & 0 & 0 & 0 & 0 & 0 & 1.5 \\
\end{array}
\]
Laplacian Matrix

\[ \begin{align*}
\text{D} & = \begin{bmatrix}
x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\
x_1 & 1.5 & 0 & 0 & 0 & 0 & 0 \\
x_2 & 0 & 1.6 & 0 & 0 & 0 & 0 \\
x_3 & 0 & 0 & 1.6 & 0 & 0 & 0 \\
x_4 & 0 & 0 & 0 & 1.7 & 0 & 0 \\
x_5 & 0 & 0 & 0 & 0 & 1.7 & 0 \\
x_6 & 0 & 0 & 0 & 0 & 0 & 1.5 \\
\end{bmatrix} \\
\text{W} & = \begin{bmatrix}
x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\
x_1 & 0 & 0.8 & 0.6 & 0 & 0.1 & 0 \\
x_2 & 0.8 & 0 & 0.8 & 0 & 0 & 0 \\
x_3 & 0.6 & 0.8 & 0 & 0.2 & 0 & 0 \\
x_4 & 0 & 0 & 0 & 0.8 & 0 & 0.8 \\
x_5 & 0.1 & 0 & 0 & 0 & 0.8 & 0 \\
x_6 & 0 & 0 & 0 & 0.7 & 0.8 & 0 \\
\end{bmatrix}
\end{align*} \]

\[ \begin{align*}
\text{D} - \text{W} & = \begin{bmatrix}
x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\
x_1 & 1.5 & -0.8 & -0.6 & 0 & -0.1 & 0 \\
x_2 & -0.8 & 1.6 & -0.8 & 0 & 0 & 0 \\
x_3 & -0.6 & -0.8 & 1.6 & -0.2 & 0 & 0 \\
x_4 & 0 & 0 & -0.2 & 1.7 & -0.8 & -0.7 \\
x_5 & -0.1 & 0 & 0 & 0.8 & 1.7 & -0.8 \\
x_6 & 0 & 0 & 0 & -0.7 & -0.8 & 1.5 \\
\end{bmatrix}
\end{align*} \]

\[ \text{L} = \text{D} - \text{W} \]

Positive Semi-Definite!!
Normalized Laplacian

\[ \tilde{L} = D^{-1/2} (D - W) D^{-1/2} \]
Spectral Clustering

• Three basic stages:

  • Pre-processing
    • Construct a matrix representation of the dataset.

  • Decomposition
    • Compute eigenvalues and eigenvectors of the matrix.
    • Map each point to a lower-dimensional representation based on one or more eigenvectors.

  • Grouping
    • Assign points to two or more clusters, based on the new representation.
When the data incorporates multiple scales standard spectral clustering fails. Note, that spectral clustering without local scaling was set manually. The data points were normalized to occupy the high impact of the matrix $X$. As was suggested by [6] the scaling parameter is some measure of when two points are considered similar. This provides an intuitive way for selecting possible values for the scaling parameter. As was suggested by [6] the scaling parameter is some measure of when two points are considered similar. This provides an intuitive way for selecting possible values for the scaling parameter. When the data contains multiple scales, even using the high impact of the matrix $X$ fails to provide good clustering (see examples at the right of top row). When the data contains multiple scales, even using the high impact of the matrix $X$ fails to provide good clustering (see examples at the right of top row).
Local Scaling

Instead of selecting a single scaling parameter $\sigma$, calculate a local scaling parameter $\sigma_i$ for each data point.

\[ w_{ij} = \exp \left( -\frac{\left\| x_i - x_j \right\|^2}{2 \sigma_i \sigma_j} \right) \]

Can be changed to other distances

\[ 1 - \frac{x_i^\top x_j}{\left\| x_i \right\|_2 \left\| x_j \right\|_2} \]

Wine Spill

Spill a drop of wine on the cloth

Spread/diffuse to the neighborhood
Label Propagation

\[ y_1 = 1 \]
\[ y_i = 0 \ (i \neq 1) \]

\[ f_i^{(t+1)} = \alpha \sum_{j \sim i} w_{ij} f_j^{(t)} + (1 - \alpha) y_i \]

\[ f^{(t+1)} = \alpha W f^{(t)} + (1 - \alpha) y \]

Label Propagation

\[ f^{(t+1)} = \alpha W f^{(t)} + (1 - \alpha) y \]

\[ f^{(0)} = y \]

\[ f^{(1)} = \alpha W f^{(0)} + (1 - \alpha) y = \alpha W y + (1 - \alpha) y \]

\[ f^{(2)} = \alpha W f^{(1)} + (1 - \alpha) y = (\alpha W)^2 y + (1 - \alpha)(\alpha W + 1)y \]

\[ f^{(3)} = \alpha W f^{(2)} + (1 - \alpha) y = (\alpha W)^3 y + (1 - \alpha)((\alpha W)^2 + \alpha W + 1)y \]

\[ \vdots \]

\[ f^{(t)} = (\alpha W)^t y + (1 - \alpha) \left( \sum_{i=0}^{t-1} (\alpha W)^i \right) y \]

If \( W \) is a stochastic matrix, its spectral radius will be no larger than 1.

Will this procedure converge?

0 < \( \alpha < 1 \) \( \Rightarrow \) \( \lim_{t \to \infty} (\alpha W)^t = 0 \)

\( \rho(W) \leq 1 \) \( \Rightarrow \) \( \lim_{t \to \infty} \sum_{i=0}^{t-1} (\alpha W)^i = (1 - \alpha W)^{-1} \)

\[ f^* = \lim_{t \to \infty} f^{(t)} = (1 - \alpha)(1 - \alpha W)^{-1} \]
The similarities are no longer symmetric. We can use the row-normalized similarity matrix as the new similarity matrix:

\[
\tilde{W} = D^{-1}W \quad \tilde{w}_{ij} = \frac{w_{ij}}{d_i}
\]

\[
D = \text{diag}(d_1, d_2, \ldots, d_n)
\]

Row Normalization

\[
d_i = \sum_j w_{ij}
\]

\[
\begin{bmatrix}
\frac{w_{11}}{d_1} & \frac{w_{12}}{d_1} & \cdots & \frac{w_{1n}}{d_1} \\
\frac{w_{21}}{d_2} & \frac{w_{22}}{d_2} & \cdots & \frac{w_{2n}}{d_2} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{w_{n1}}{d_n} & \frac{w_{n2}}{d_n} & \cdots & \frac{w_{nn}}{d_n}
\end{bmatrix}
\]
**Bi-Stochastic Matrix**

We want a stochastic matrix which is symmetric

A **right stochastic** matrix is a real nonnegative square matrix, with each row summing to 1.

A **left stochastic** matrix is a real nonnegative square matrix, with each column summing to 1.

A **bi-(doubly) stochastic** matrix is a real nonnegative square matrix, with both row and column summing to 1.


Factorization
Matrix Factorization

\[ \mathcal{L} \left( X, \prod_{i=1}^{K} A_i \right) + \Omega \left( \bigcup_{i=1}^{K} A_i \right) \]

s.t. \[ C \left( \bigcup_{i=1}^{K} A_i \right) \]

- different types of product
- different loss function
- different regularizations
- different constraints
Matrix Factorization

\[ X \sim F \times G \]
Probabilistic Matrix Factorization

\[ p(V|\sigma_V^2) = \prod_{j=1}^{M} N(V_j|0, \sigma_V^2 I) \]

\[ p(U|\sigma_U^2) = \prod_{i=1}^{N} N(U_i|0, \sigma_U^2 I) \]

\[ p(R|U, V, \sigma^2) = \prod_{i=1}^{N} \prod_{j=1}^{M} \left[ N(R_{ij}|U_i^T V_j, \sigma^2) \right] \]

Nonnegative Matrix Factorization
Nonnegative Matrix Factorization

- Factorizing a nonnegative matrix to the product of two low-rank matrices

\[ F = (f_1, f_2, \ldots, f_k) \quad G = (g_1, g_2, \ldots, g_k) \]

\[ X = (x_1, x_2, \ldots, x_n) \]

\[ FG^T \approx X \]

\[
\mathcal{L}(F, G) = \left\| X - FG^T \right\|_F^2
\]

\[ \min_{F \in \mathbb{R}_{+}^{d \times r}, G \in \mathbb{R}_{+}^{n \times r}} \mathcal{L}(F, G) \]

Learning Medical Concepts

- Assume the **full** longitudinal patient matrix can be approximated by a low rank matrix
  - Macro-phenotype mapping matrix $U$: **Sparse, Non-negative**
  - Concept value evolution matrix $V$: **Temporal Smoothness**

Learning Medical Concepts

- Assume that each patient has different medical concepts from other patients

- Formulation

\[
\begin{align*}
\min_{(s_i),(u_i),(v_i)} & \quad \frac{1}{2t_i} \| s_i - u_i v_i \|_F^2 + \lambda_1 \| u_i \|_1 + \lambda_2 \sum_{i=1}^{n} \frac{1}{2t_i} \| v_i \|_F^2 + \lambda_3 \sum_{i=1}^{n} \| v_i R_i \|_F^2 \\
\text{subject to} & \quad P_{\Omega_i}(s_i) = P_{\Omega_i}(x_i), \quad u_i \geq 0, \forall i
\end{align*}
\]

**Matrix Completion**
Completion via matrix factorization. Enforce a low rank factorization \( U_i, V_i \) and encourage the values of \( U_i V_i \) at the observed location to be close to the observed ones.

**Meaningful Medical Concepts**
Medical concepts involve non-negative components of medical features.

**Sparsity**
Controllable sparsity that encourages a few medical features in each concept.

**Temporal Smoothness**
Couple the columns of \( V_i \) and force them to close to each other.

**Overfitting Control**
Prevent \( V_i \) from overfitting.

---

Dictionary Learning

Seek for a set of basis which can sparsely represent the data set

\[
\sum_{i=1}^{n} \left\| x_i - \sum_{j=1}^{k} G_{ij} f_j \right\|^2
\]

\[
J_0 = \left\| X - FG^\top \right\|_F^2 + \lambda \sum_{i=1}^{n} \| G_i \|_1
\]

Data matrix
\[
X = [x_1, x_2, \cdots, x_n]
\]

Basis matrix
\[
F = [f_1, f_2, \cdots, f_k]
\]

Coding coefficient matrix, n by k

Enforce a sparse representation of the i-th data object

Non-negativity
Dictionary Learning

Seek for a sparse representation of the data in the c-th group

\[
\min \sum_{c=1}^{C} \left[ \|X_c - FG_c^T\|_F^2 + \lambda \sum_{i=1}^{n_c} \|G_{ci}\|_p \right] + \gamma \sum_{j=1}^{k} \|F_j\|_p
\]

s.t. \( F \geq 0, \ G_c \geq 0 \) (\( c = 1, 2, \cdots, C \))

A common dictionary shared over all C groups of data

The data groups are pre-defined
The dictionary is shared over all data groups

Dictionary Learning

Learn both shared and group-specific dictionaries

\[
\min \sum_c \left\| X_c - F^S_c G^S_c^T - F^I_c G^I_c^T \right\|_F^2 + \sum_c \left[ \gamma_I \phi(G^I_c) + \gamma_S \phi(G^S_c) \right]
\]

- **Shared dictionary**
- **Group specific dictionary for group \( c \)**

\[
\phi(G^I_c) = \sum_{i=1}^{n_c} \left\| G^I_c_{i} \right\|_1
\]

\[
\phi(G^S_c) = \sum_{i=1}^{n_c} \left\| G^S_c_{i} \right\|_1
\]

**References:**

Temporal Patterns

One-Side Convolution

Definition  (One-Side Convolution). The one-side convolution of $F \in \mathbb{R}^{n \times m}$ and $g \in \mathbb{R}^{t \times 1}$ is an $n \times t$ matrix with

$$(F * g)_{ij} = \sum_{k=1}^{t} g_{j-k+1}F_{ik}$$

Note that $g_j = 0$ if $j \leq 0$ or $j > t$, and $F_{ik} = 0$ if $k > m$. 
One-Side Convolutional NMF

\[
\min_{\mathcal{F}, \mathcal{G}} J \\
\text{s.t. } \forall r = 1, \cdots, R; c = 1, \cdots, C \\
\mathbf{F}^{(r)} \geq 0, \mathbf{g}_c^{(r)} \geq 0
\]

\[
F_{ik}^{(r)} \leftarrow F_{ik}^{(r)} \left( \frac{\sum_{c=1}^{C} \sum_{j=1}^{t} A_{cij}^{\beta-1} X_{cij} Y_{cij}^{\beta-2} g_{c_{j-k+1}}^{(r)}}{\sum_{c=1}^{C} \sum_{j=1}^{t} A_{cij}^{\beta-1} g_{c_{j-k+1}}^{(r)} + \lambda_1} \right)^{\eta(\beta)}
\]

\[
g_{c_k}^{(r)} \leftarrow g_{c_k}^{(r)} \left( \frac{\sum_{i=1}^{n} \sum_{j=1}^{t} A_{cij}^{\beta-1} X_{cij} Y_{cij}^{\beta-2} F_{i_{j-k+1}}^{(r)}}{\sum_{i=1}^{n} \sum_{j=1}^{t} A_{cij}^{\beta-1} F_{i_{j-k+1}}^{(r)} + \lambda_2} \right)^{\eta(\beta)}
\]

\[
J = \sum_{c=1}^{C} d_{\beta} \left( \mathbf{A}_c \odot \mathbf{X}_c, \mathbf{A}_c \odot \left( \sum_{r=1}^{R} \mathbf{F}^{(r)} \ast \mathbf{g}_c^{(r)} \right) \right) + \lambda_1 \sum_{r=1}^{R} \| \mathbf{F}^{(r)} \|_1 + \lambda_2 \sum_{c=1}^{C} \sum_{r=1}^{R} \| \mathbf{g}_c^{(r)} \|_1
\]

\[
\eta(\beta) = \begin{cases} \\
\frac{1}{2-\beta}, & \beta < 1 \\
1, & 1 \leq \beta \leq 2 \\
\frac{1}{\beta-1}, & \beta > 2 \\
\end{cases}
\]

**Definition** (\(\beta\)-divergence) The \(\beta\)-divergence between two matrices \(\mathbf{A}\) and \(\mathbf{B}\) with the same size is

\[
d_{\beta}(\mathbf{A}, \mathbf{B}) = \frac{1}{\beta(\beta-1)} \sum_{ij} \left( A_{ij}^\beta + (\beta - 1) B_{ij}^\beta - \beta A_{ij} B_{ij}^{\beta-1} \right)
\]
The sequentiality of those events may indicate some impending disease conditions. How to interpret and make use of the sequentiality of the events?
A Graph Based Formulation

Definition (Temporal graph)

The temporal graph $G^n$ of sequence $s_n$ is a directed and weighted graph with our event set as its node set $\{1, \cdots, M\}$, where the weight of the edge from node $i$ to node $j$ is defined as

$$W_{ij}^n = \frac{1}{L_n} \sum_{1 \leq p \leq q \leq L_n} [x_{np} = i \land x_{nq} = j] \kappa_r (t_{nq} - t_{np}),$$

where $\kappa_r$ is a non-increasing function and $r > 0$ is a parameter.

$$\kappa(\delta) = \begin{cases} 
\exp(-\delta/r) & \delta \leq \Delta \\
0 & \delta > \Delta 
\end{cases}$$
An Example Graph

- Anti-inflammatories
- Glucocorticoids
- Pulmonary Disease
- Asthma
- Congestive Heart Failure
- ACE Inhibitors
- Diuretics
- Mineral Replacement
- Laxatives
- Antianginal Agents
Graph Basis

\[ W^n = \sum_{k=1}^{K} A_{nk} B^k \]

\[ \mathcal{J}(A, B) = \frac{1}{2} \sum_{n=1}^{N} \| W^n - \sum_{k=1}^{K} A_{nk} B^k \|_F^2 \]
Regularization

\[ J(A, B) = \frac{1}{2} \sum_{n=1}^{N} \| W^n - \sum_{k=1}^{K} A_{nk} B^k \|_F^2 + \lambda \Omega(A) \]

\[ \Omega(A) = \frac{1}{2} \sum_{n_1, n_2} \frac{1}{2} \| A_{n_1} - A_{n_2} \|_2^2 S_{n_1 n_2} \]

\[ \Omega(A) = \frac{1}{|\mathcal{L}|} \sum_{n \in \mathcal{L}} \text{loss}(A_n, Y_n | \mathcal{H}) \]

Similarity based regularizer: similar patients should have similar combination coefficients

Prediction based regularizer: the combination coefficients will be used as an alternative patient representation

Clustering

\[
\min_{F \geq 0, G \geq 0} \|X - FG^T\|^2, \quad \text{s.t. } G^T G = I.
\]

\[
\min_{F \geq 0, G \geq 0, S \geq 0} \|X - FS G^T\|^2, \quad \text{s.t. } F^T F = I, \quad G^T G = I.
\]

\[
\min_{H \geq 0, S \geq 0} \|X - HS H^T\|^2, \quad \text{s.t. } H^T H = I.
\]

Questions?

feiwang03@gmail.com