Latent Factor Models for Relational Data

Peter Hoff
Statistics, Biostatistics and
Center for Statistics and the Social Sciences
University of Washington
Part 1: Multiplicative factor models for network data
   1. Relational data
   2. Models via exchangeability
   3. Models via matrix representations
   4. Examples:
      (a) International conflict data
      (b) Protein-protein interaction data

Part 2: Rank selection with the SVD model
   1. A Gibbs sampling scheme for rank estimation
   2. A simulation study
Relational data: consist of

- a set of units or nodes $A$, and
- a set of measurements $Y \equiv \{y_{i,j}\}$ specific to pairs of nodes $(i,j) \in A \times A$.

Examples:

**International relations**

$A =$countries, $y_{i,j} =$ indicator of a dispute initiated by $i$ with target $j$.

**Needle-sharing network**

$A =$IV drug users, $y_{i,j} =$ needle-sharing activity between $i$ and $j$.

**Protein-protein interactions**

$A =$proteins, $y_{i,j} =$ the interaction between $i$ and $j$.

**Business locations**

$A_1 =$banks, $A_2 =$cities, $y_{i,j} =$ presence of an office of bank $i$ in city $j$. 
Inferential goals in the regression framework

\[ y_{i,j} \text{ measures } i \rightarrow j, \quad x_{i,j} \text{ is a vector of explanatory variables.} \]

\[
Y = \begin{pmatrix}
 y_{1,1} & y_{1,2} & y_{1,3} & \text{NA} & y_{1,5} & \cdots \\
 y_{2,1} & y_{2,2} & y_{2,3} & y_{2,4} & y_{2,5} & \cdots \\
 y_{3,1} & \text{NA} & y_{3,3} & y_{3,4} & \text{NA} & \cdots \\
 y_{4,1} & y_{4,2} & y_{4,3} & y_{4,4} & y_{4,5} & \cdots \\
 \vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{pmatrix}
\]

\[
X = \begin{pmatrix}
 x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} & x_{1,5} & \cdots \\
 x_{2,1} & x_{2,2} & x_{2,3} & x_{2,4} & x_{2,5} & \cdots \\
 x_{3,1} & x_{3,2} & x_{3,3} & x_{3,4} & x_{3,5} & \cdots \\
 x_{4,1} & x_{4,2} & x_{4,3} & x_{4,4} & x_{4,5} & \cdots \\
 \vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{pmatrix}
\]

Consider a basic (generalized) linear model

\[ y_{i,j} \sim \beta' x_{i,j} + \epsilon_{i,j} \]

A model can provide

- a measure of the association between \( X \) and \( Y \): \( \hat{\beta}, \text{se}(\hat{\beta}) \)
- predictions of missing or future observations: \( p(y_{1,4}|Y, X) \)
How might node heterogeneity affect inference?
Latent variable models

Deviations from ordinary regression models can be represented as

\[ y_{i,j} \sim \beta' x_{i,j} + z_{i,j} \]

A simple “latent variable” model might include row and column effects:

\[ z_{i,j} = u_i + v_j + \epsilon_{i,j} \Rightarrow y_{i,j} \sim \beta' x_{i,j} + u_i + v_j + \epsilon_{i,j} \]

\( u_i \) and \( v_j \) induce across-node heterogeneity that is additive on the scale of the regressors. Inclusion of these effects in the model can dramatically improve

- within-sample model fit (measured by \( R^2 \), likelihood ratio, BIC, etc.);
- out-of-sample predictive performance (measured by cross-validation).

But this model only captures heterogeneity of outdegree/indegree, and can’t represent more complicated structure, such as clustering, transitivity, etc.
Latent variable models via exchangeability

\(X\) represents known information about the nodes

\(Z\) represents any additional patterns

In this case we might be willing to use a model in which

\[
\{z_{i,j}\} \overset{d}{=} \{z_{gi,hj}\}
\]

for all permutations \(g\) and \(h\). This is a type of exchangeability for arrays, sometimes called weak exchangeability.

**Theorem (Aldous, Hoover):** Let \(\{z_{i,j}\}\) be a weakly exchangeable array. Then \(\{z_{i,j}\} \overset{d}{=} \{z^*_{i,j}\}\), where

\[
z^*_{i,j} = f(\mu, u_i, v_j, \epsilon_{i,j})
\]

and \(\mu, \{u_i\}, \{v_j\}, \{\epsilon_{i,j}\}\) are all independent random variables.
The singular value decomposition model

\[ Z = M + E \]

M represents “systematic” patterns in the effects and E represents “noise”.

Every M has a representation of the form \( M = UDV' \) where, in the case \( m \geq n \),

- \( U \) is an \( m \times n \) matrix with orthonormal columns;
- \( V \) is an \( n \times n \) matrix with orthonormal columns;
- \( D \) is an \( n \times n \) diagonal matrix, with elements \( \{d_1, \ldots, d_n\} \) typically taken to be a decreasing sequence of non-negative numbers.

Recall,

- The squared elements of the diagonal of \( D \) are the eigenvalues of \( MM' \) and the columns of \( V \) are the corresponding eigenvectors.
- The matrix \( U \) can be obtained from the first \( n \) eigenvectors of \( MM' \). The number of non-zero elements of \( D \) is the rank of \( M \).
- Writing the row vectors as \( \{u_1, \ldots, u_m\}, \{v_1, \ldots, v_n\} \), \( m_{i,j} = u'_iDv_j \).
Multiplicative effects

\[ y_{i,j} \sim \beta' x_{i,j} + u_i' D v_j + \epsilon_{i,j} \]

**Interpretation:**

Think of \{u_1, \ldots, u_m\}, \{v_1, \ldots, v_n\} as vectors of \textit{latent nodal attributes}:

\[ u_i' D v_j = \sum_{k=1}^{K} d_{k} u_{i,k} v_{j,k} \]

In general, a latent variable model relating \( X \) to \( Y \) is

\[ g(\mathbb{E}[y_{i,j}]) = \beta' x_{i,j} + u_i' D v_j + \epsilon_{i,j} \]

For example, some potential models are

- If \( y_{i,j} \) is binary, \( \log \text{odds} (y_{i,j} = 1) = \beta' x_{i,j} + u_i' D v_j + \epsilon_{i,j} \)
- If \( y_{i,j} \) is count data, \( \log \mathbb{E}[y_{i,j}] = \beta' x_{i,j} + u_i' D v_j + \epsilon_{i,j} \)
- If \( y_{i,j} \) is continuous, \( \mathbb{E}[y_{i,j}] = \beta' x_{i,j} + u_i' D v_j + \epsilon_{i,j} \)

**Estimation:** Given \( D, V \), the predictor is linear in \( U \). This bilinear structure can be exploited (EM, Gibbs sampling, variational methods).
International conflict network, 1990-2000

11 years of international relations data (thanks to Mike Ward and Xun Cao)

- $y_{i,j} =$ indicator of a militarized disputes initiated by $i$ with target $j$;
- $x_{i,j}$ an 8-dimensional covariate vector containing an intercept and
  1. population initiator
  2. population target
  3. polity score initiator
  4. polity score target
  5. polity score initiator $\times$ polity score target
  6. log distance
  7. number of shared intergovernmental organizations

Model: $y_{i,j}$ are independent binary random variables with log-odds

$$\log \text{odds}(y_{i,j} = 1) = \beta' x_{i,j} + u'_i D v_j + \epsilon_{i,j}$$
International conflict network: parameter estimates

![Graph showing parameter estimates for various factors like log distance, polity initiator, intergov org, polity target, polity interaction, log pop target, log pop initiator.](image)
International conflict network: description of latent variation
International conflict network: prediction experiment

The graph shows the relationship between the number of pairs checked and the number of links found, with different values of K (K=0, K=1, K=2, K=3). The lines indicate the trend for each K value, with each line representing a different level of network complexity.

- **K=0**: The line is the steepest, indicating a higher number of links found for each number of pairs checked.
- **K=1**: The line is slightly less steep than K=0, showing a moderate increase in links found.
- **K=2**: The line is the least steep, indicating the least increase in links found.
- **K=3**: The line is between K=0 and K=1, showing a trend somewhere between the two extremes.

The x-axis represents the number of pairs checked, while the y-axis represents the number of links found.
Parameters to estimate include the diagonal matrix $D$ and:

$$
U \in \mathcal{V}_{K,m} = \{m \times K \text{ matrices with orthonormal columns}\}
$$

$$
V \in \mathcal{V}_{K,n} = \{n \times K \text{ matrices with orthonormal columns}\}
$$

**A constructive definition for $p(U)$:**

1. $z_1 \sim \text{uniform}[m\text{-sphere}]$, set $U_{[,1]} = z_1$;
2. $z_2 \sim \text{uniform}[(m - 1)\text{-sphere}]$, set $U_{[,2]} = \text{Null}(U_{[,1]})z_2$;
3. $\vdots$
4. $K$. $z_K \sim \text{uniform}[(m - K + 1)\text{-sphere}]$, $U_{[,K]} = \text{Null}(U_{[,1]}, \ldots, U_{[,K-1]})z_K$.

The resulting distribution for $U$ is the uniform (invariant) distribution on the Steifel manifold, and is exchangeable under row and column permutations.

**Prior conditional distributions:**

$$
(U_{[,j]}|U_{[,\neg j]} \overset{d}{=} \text{Null}(U_{[,1]}, \ldots, U_{[,j-1]}, U_{[,j+1]}, \ldots, U_{[,K]})z_j; \\
\quad z_j \sim \text{uniform}[(m - K + 1)\text{-sphere}]
$$
Estimation details: posterior distribution for $U$

$$UDV' = d_1 U_{[1]} V_{[1]}' + \cdots + d_K U_{[K]} V_{[K]}' ,$$

so for any column $j$

$$Z - UDV' = (Z - U_{[-j]} D_{[-j,-j]} V_{[-j]}') - d_j U_{[j]} V'_{[j]}$$

$$\equiv R_j - d_j U_{[j]} V'_{[j]}$$

$$||Z - UDV'||^2 = ||R_j - d_j U_{[j]} V'_{[j]}||^2$$

$$= ||R_j||^2 - 2d_j U_{[j]}' R_j V_{[j]} + ||d_j U_{[j]} V'_{[j]}||^2$$

$$= ||R_j||^2 - 2d_j U_{[j]}' R_j V_{[j]} + d_j^2$$

Recall $Z = UDV + E$ where $E$ is normally distributed noise with variance $1/\phi$:

$$p(Z|U, V, D, \phi) = (2\pi\phi)^{-mn/2} \exp\{-\frac{1}{2}||R_j||^2 + \phi d_j U_{[j]}' R_j V_{[j]} - \frac{1}{2}\phi d_j^2\}$$

$$p(U_{[j]}|Z, U_{[-j]}, V, D) \propto p(U_{[j]}|U_{[-j]}) \times \exp\{ \phi d_j U_{[j]}' R_j V_{[j]} \}$$

$$(U_{[j]}|U_{[-j]}, Z, V, D) \overset{d}{=} \text{Null}(U_{[1]}, \ldots, U_{[j-1]}, U_{[j+1]}, \ldots, U_{[K]}) z_j \overset{\text{d}}{=} N_{-j} z_j ,$$

$$p(z_j) \propto \exp\{z_j' \mu\}, \quad \mu = \phi d_j N_{-j} R_j V_{[j]}$$
The von Mises-Fisher distribution

\[ p(z|\mu) = c(\kappa) \exp\{z'\mu\} , \text{ where } \kappa = ||\mu|| \]
Analyzing undirected data

In many applications $y_{i,j} = y_{j,i}$ by design, and so $(y_{i,j} = y_{j,i}) \sim \beta' \mathbf{x}_{i,j} + z_{i,j}$, and $\mathbf{Z} = \{z_{i,j}\}$ is a symmetric array. How should $\mathbf{Z}$ be modeled?

**Modeling via exchangeability:** Let $\mathbf{Z}$ be a symmetric array (with an undefined diagonal), such that $\{z_{i,j}\} \overset{d}{=} \{z_{g_i,g_j}\}$. Then $\{z_{i,j}\} \overset{d}{=} \{z^*_{i,j}\}$, where

$$z^*_{i,j} = f(\mu, u_i, u_j, \epsilon_{i,j}),$$

$\mu, \{u_i\}, \{\epsilon_{i,j}\}$ are independent random variables, and $f(\cdot, u_i, u_j, \cdot) = f(\cdot, u_j, u_i, \cdot)$.

**Modeling via matrix decomposition:** Write $\mathbf{Z} = \mathbf{M} + \mathbf{E}$, with all matrices symmetric. All such $\mathbf{M}$ have an eigenvalue decomposition

$$\mathbf{M} = \mathbf{U} \Lambda \mathbf{U}'$$

$$m_{i,j} = u_i' \Lambda u_j$$

This suggests a model of the form

$$(y_{i,j} = y_{j,i}) \sim \beta' \mathbf{x}_{i,j} + u_i' \Lambda u_j + \epsilon_{i,j}$$
Protein-protein interaction network

Interactions among 270 proteins in E. coli (Butland, 2005).

Data: \( \frac{1}{\binom{270}{2}} \sum_{i<j} y_{i,j} \approx 0.01 \)

Model: \( \log \text{odds}(y_{i,j} = 1) = \mu + \mathbf{u}'_i \Lambda \mathbf{u}_j + \epsilon_{i,j} \)

(analysis by Ryan Bressler)
Protein-protein interaction network
Protein-protein interaction network

RNA polymerase

- hepA
- rpoZ
- yacL
- nusG
- rpoN
- rpoS
- manX
- b1731
- greA
- rpoB
- nusA
- infB
- rpoD
- rpoA
- rpoC
- greB
- polA
- usg
- b2372

Positive
Negative

0.85 0.90 0.95 1.00
−0.6 −0.4 −0.2 0.0 0.2 0.4 0.6

20
Protein-protein interaction network: prediction experiment

![Graph showing the relationship between the number of pairs checked and the number of links found.]
Data analysis with the singular value decomposition

Many data analysis procedures for matrix-valued data $Y$ are related to the SVD. Given a model of the form

$$Y = M + E$$

where $E$ is independent normal noise, the SVD provides

**Interpretation:** $y_{i,j} = u_i' D v_j + \epsilon_{i,j}$, \hspace{1cm} $u_i$ and $v_j$ are the $i$th, $j$th rows of $U$, $V$

**Estimation:** $\hat{M}_K = \hat{U}_{[1:K]} \hat{D}_{[1:K,1:K]} \hat{V}'_{[1:K]}$ if $M$ is assumed to be of rank $K$.

Applications:

- biplots (Gabriel 1971, Gower and Hand 1996)
- reduced-rank interaction models (Gabriel 1978, 1998)
- analysis of relational data (Harshman et al., 1982)
- Factor analysis, image processing, data reduction, . . .

But:

- How to select $K$? Given $K$, is $\hat{M}_K$ a good estimator? \hspace{1cm} $(E[Y'Y] = M'M + m\sigma^2 I)$
A toy example

\[ Y_{100 \times 100} = UDV' + E = \sum d_k U_{[,k]} V'_{[,k]} + E \]
A toy example

\[ Y_{100 \times 100} = UDV' + E = \sum d_k U_{[,k]} V'_{[,k]} + E \]
Evaluating the dimension

It can all be done by comparing

\[ M_1 : \ Y = duv' + E \]
\[ M_0 : \ Y = E \]

To decide whether or not to add a dimension, we need to calculate

\[
\frac{p(M_1|\mathbf{Y})}{p(M_0|\mathbf{Y})} = \frac{p(M_1) p(\mathbf{Y}|M_1)}{p(M_0) p(\mathbf{Y}|M_0)}
\]

The numerator of the Bayes factor is the hard part. We need to integrate the following w.r.t. the prior distribution:

\[
p(\mathbf{Y}|M_1, \mathbf{u}, \mathbf{v}, d) = (2\pi)^{-nm/2} \exp\{-\frac{1}{2} ||\mathbf{Y}||^2/2 + \phi d\mathbf{u}'\mathbf{Y}\mathbf{v} - \phi d^2/2\}
\]
\[
= p(\mathbf{Y}|M_0) \times \exp\{-\phi d^2/2\} \times \exp\{\mathbf{u}'[d\phi\mathbf{Y}]\mathbf{v}\}
\]

Computing \(E[e^{\mathbf{u}'\mathbf{A}\mathbf{v}}]\) over \(\mathbf{u} \in S_m\) and \(\mathbf{v} \in S_n\) is difficult but doable.

- It involves the integral of a Bessel function and a weird hypergeometric function.
- These results provide a description of a joint distribution to represent “bivariate” dependence among points on spheres: \(p(\mathbf{u}, \mathbf{v}|\mathbf{A}) = c(\mathbf{A}) \exp\{\mathbf{u}'\mathbf{A}\mathbf{v}\}\)
100 datasets were generated for each of \((m, n) \in \{(10, 10), (100, 10), (100, 100)\}\):

- \(\mathbf{U} \sim \text{uniform}(\mathcal{V}_{5,m}), \mathbf{V} \sim \text{uniform}(\mathcal{V}_{5,n})\);  
- \(\mathbf{D} = \text{diag}\{d_1, \ldots, d_5\}, \{d_1, \ldots, d_5\} \sim \text{i.i.d. uniform}(\frac{1}{2}\mu_{mn}, \frac{3}{2}\mu_{mn})\), where 
  \[\mu_{mn}^2 = m + n + 2\sqrt{mn}.\]
- \(\mathbf{Y} = \mathbf{UDV}' + \mathbf{E}\), where \(\mathbf{E}\) is an \(m \times n\) matrix of standard normal noise.

The choice of \(\mu_{mn}\) is not arbitrary: The largest eigenvalue of \(\mathbf{E}\) is approximately 
\[\lambda_{\max} \approx (m + n + 2\sqrt{mn}).\]

(Edelman, 1988)
Simulation study

- $m=10$, $n=10$
- $m=100$, $n=10$
- $m=100$, $n=100$
Simulation study

<table>
<thead>
<tr>
<th>m=10   n=10</th>
<th>m=100  n=10</th>
<th>m=100  n=100</th>
</tr>
</thead>
<tbody>
<tr>
<td>posterior mode</td>
<td></td>
<td></td>
</tr>
<tr>
<td>max eigengap</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Simulation study

- Posterior mode
- Maximum eigengap
- Eigenvalue of correlation

- m=10  n=10
- m=100  n=10
- m=100  n=100
Simulation study

- $m=10, n=10$
- $m=100, n=10$
- $m=100, n=100$
Summary:

• Relevance disclaimer:
  – If your goal is to represent large components of variation in the data, then maybe you don’t need a model.
  – If you believe $Y \sim M + E$, and you want to represent large components of variation in $M$, then a model can be helpful.

• SVD models are a natural way to represent patterns in relational data;
• The SVD structure can be incorporated into a variety of model forms;
• BMA on model rank can be done.

Potentially straightforward extensions:

• Representation of higher dimensional arrays: $m_{i,j,k} = \sum_{l=1}^{L} l u_{i,l} v_{j,l} w_{l,k}$;
• Dynamic network inference (times series models for $Y, U, D, V$);
• Restrict latent characteristics to be orthogonal to design vectors.