Latent Factor Models for Relational Data

Peter Hoff
Statistics, Biostatistics and
Center for Statistics and the Social Sciences
University of Washington
Outline

Part 1: Multiplicative factor models for network data
  Relational data
  Eigenvalue decomposition model
  Example: Adolescent social network
  Singular value decomposition model
  Example: International conflicts

Part 2: Extension to multiway data
  Multiway data
  Multiway latent factor models
  Example: Cold war cooperation and conflict

Summary and future work
Relational data

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 \( j \)

Document analysis
\( A_1 = \) words, \( A_2 = \) documents, \( y_{i,j} = \) wordcount of \( i \) in document \( j \)
Data on 82 12th graders from a single high school:
52 boys, 28 girls
\( \hat{Pr}(y_{i,j} = 1|\text{same sex}) = 0.077 \)
\( \hat{Pr}(y_{i,j} = 1|\text{opposite sex}) = 0.056 \)
### Inferential goals in the regression framework

$y_{i,j}$ measures $i \rightarrow j$, $x_{i,j}$ 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

$$
y_{i,j} \sim \beta' x_{i,j} + e_{i,j}
$$

A model can provide

- a measure of the association between $X$ and $Y$: $\hat{\beta}$, $\text{se}(\hat{\beta})$
- imputations of missing observations: $p(y_{1,4}|Y, X)$
- a probabilistic description of network features: $g(\tilde{Y})$, $\tilde{Y} \sim p(\tilde{Y}|Y, X)$
Nodal heterogeneity and independence assumptions
Model fit

glm(formula = y ~ x, family = binomial(link = "logit"))

Coefficients:

|              | Estimate | Std. Error | z value | Pr(>|z|) |
|--------------|----------|------------|---------|----------|
| (Intercept)  | -2.8332  | 0.1123     | -25.24  | <2e-16 *** |
| x            | 0.3471   | 0.1428     | 2.43    | 0.0151 * |

This result says that a model with preferential association is a better description of the data than an i.i.d. binary model.
Model lack of fit

Neither of these models do well in terms of representing other features of the data - for example, transitivity:

\[ t(\mathbf{Y}) = \sum_{i<j<k} y_{i,j}y_{j,k}y_{k,i} \]
Latent variable models

Deviations from ordinary regression models can be represented as

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

A simple “latent variable” model might include additive node effects:

\[ e_{i,j} = u_i + u_j + \epsilon_{i,j} \quad \Rightarrow \quad y_{i,j} \sim \beta' x_{i,j} + u_i + u_j + \epsilon_{i,j} \]

\{u_1, \ldots, u_n\} represent 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.
Fit of additive effects model
Latent variable models via exchangeability

\( \mathbf{X} \) represents known information about the nodes
\( \mathbf{E} \) represents deviations from the regression model
In this case we might be willing to use a model for \( \mathbf{E} \) in which

\[
\{ e_{i,j} \} \overset{d}{=} \{ e_{g_i,g_j} \}
\]

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

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

\[
e^{*}_{i,j} = f(u_i, u_j, \epsilon_{i,j})
\]

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

\[ E = M + \mathcal{E} \]

**M** represents “systematic” patterns and \( \mathcal{E} \) represents “noise”. Every symmetric **M** has a representation of the form \( M = U\Lambda U' \) where

- **U** is an \( m \times m \) matrix with orthonormal columns
- \( \Lambda \) is an \( n \times n \) diagonal matrix, with diagonal elements \( \{\lambda_1, \ldots, \lambda_n\} \)

Many data analysis procedures for symmetric matrix-valued data **Y** are related to this decomposition. Given a model of the form

\[ Y = M + \mathcal{E} \]

where \( \mathcal{E} \) is independent noise, the ED provides

**Interpretation:** \( y_{i,j} = u_i'\Lambda u_j + \epsilon_{i,j}, \) \( u_i \) and \( u_j \) are the \( i \)th, \( j \)th rows of **U**

**Estimation:** \( \hat{M}_K = \hat{U}[1:K]\hat{\Lambda}[1:K,1:K]\hat{U}'[1:K] \) if **M** is assumed to be of rank \( K \).
Generalized bilinear regression

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

Interpretation:
Think of \( \{u_1, \ldots, u_n\} \) as vectors of **latent nodal attributes**:

\[ u_i' \Lambda u_j = \sum_{k=1}^{K} \lambda_k u_{i,k} u_{j,k} \]

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

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

and the parameters can be estimated using a rank-likelihood or multinomial probit. Alternatively, parametric models include

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

**Estimation:** Given \( \Lambda \), the predictor is linear in \( U \). This bilinear structure can be exploited (EM, Gibbs sampling).
Eigenmodel fit

Parameters this model can be fit with the eigenmodel package in R:

eigenmodel_mcmc(Y,X,R=3)

The latent factors are able to represent the network transitivity.
Underlying structure
Missing variables
The eigenmodel, without having explicit race information, captures a large degree of the racial homophily in friendship:
The SVD model for asymmetric data

\[ E = M + \mathcal{E} \]

\( M \) represents “systematic” patterns and \( \mathcal{E} \) 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 diagonal 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 \( M'M \) 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 \( \{u_1, \ldots, u_m\}, \{v_1, \ldots, v_n\} \), \( m_{i,j} = u'_i D v_j \). 

▶
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

$$Y = M + \mathcal{E}$$

where $\mathcal{E}$ is independent noise, the SVD provides

**Interpretation:** $y_{i,j} = u_i' D v_j + \epsilon_{i,j}$, $u_i$ 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, ...

**Notes:**
- How to select $K$? Given $K$, is $\hat{M}_K$ a good estimator? ($E[Y'Y] = M'M + m\sigma^2 I$)
International conflict network, 1990-2000

11 years of international relations data (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

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

log distance
polity initiator
intergov org
polity target
polity interaction
log pop target
log pop initiator

regression coefficient

-3 -2 -1 0 1
International conflict network: description of latent variation
International conflict network: prediction experiment

The diagram illustrates the number of pairs checked against the number of links found, with different lines representing different values of K (K=0, K=1, K=2, K=3). The graph shows a linear relationship between the number of pairs checked and the number of links found for each value of K.
Multiway data

Data on pairs is sometimes called two-way data.

More generally, data on triples, quadruples, etc. is called multi-way data.

$A_1, A_2, \ldots, A_p$ represent classes of objects.

$y_{i_1, i_2, \ldots, i_p}$ is the measurement specific to $i_1 \in A_1, \ldots, i_p \in A_p$.

**Examples:**

**International relations**

$A_1 = A_2 = \text{countries}$, $A_3 = \text{time}$,

$y_{i,j,t} = \text{indicator of a dispute between } i \text{ and } j \text{ in year } t$.

**Social networks**

$A_1 = A_2 = \text{individuals}$, $A_3 = \text{information sources}$,

$y_{i,j,k} = \text{source } k\text{'s report of the relationship between } i \text{ and } j$.

**Document analysis**

$A_1 = A_2 = \text{words}$, $A_3 = \text{documents}$,

$y_{i,j,k} = \text{co-occurrence of words } i \text{ and } j \text{ in document } k$. 
Factor models for multiway data

Recall the decomposition of a two-way array of rank $R$:

$$m_{i,j} = u'_i D v_j = \sum_{r=1}^{R} u_{i,r} v_{j,r} d_r$$

Now generalize to a three-way array:

$$m_{i,j,k} = \sum_{r=1}^{R} u_{i,r} v_{j,r} w_{k,r} d_r$$

▶ $\{u_1, \ldots, u_{n_1}\}$ represents variation along the 1st dimension
▶ $\{v_1, \ldots, v_{n_2}\}$ represents variation along the 2nd dimension
▶ $\{w_1, \ldots, w_{n_3}\}$ represents variation along the 3rd dimension

Consider the $k$th “slab” of $M$, which is an $n_1 \times n_2$ matrix:

$$m_{i,j,k} = \sum_{r=1}^{R} u_{i,r} v_{j,r} w_{k,r} d_r$$

$$= \sum_{r=1}^{R} u_{i,r} v_{j,r} d_{k,r} = u'_i D_k v_j \quad \text{where } d_{k,r} = w_{k,r} d_k$$
Cold war cooperation and conflict data

\begin{align*}
\beta_{\text{trade}} & \quad 1950 & 1960 & 1970 & 1980 \\
\beta_{\text{gdp}} & \quad 1950 & 1960 & 1970 & 1980 \\
\beta_{\text{distance}} & \quad 1950 & 1960 & 1970 & 1980 \\
\lambda_1 & \quad 1950 & 1960 & 1970 & 1980 \\
\lambda_2 & \quad 1950 & 1960 & 1970 & 1980 \\
\lambda_3 & \quad 1950 & 1960 & 1970 & 1980 
\end{align*}
Cold war cooperation and conflict data
Summary and future directions

Summary:

- Latent factor models are a natural way to represent patterns in relational or array-structured data.
- The latent factor structure can be incorporated into a variety of model forms.
- Model-based methods
  - give parameter estimates;
  - accommodate missing data;
  - provide predictions;
  - are easy to extend.

Future Directions:

- Dynamic network inference
- Generalization to multi-level models