

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

- Models via exchangeability

- Models via matrix representations

- Examples: conflict data, protein interaction data

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 $\mathbf{Y} \equiv \{y_{i,j}\}$ 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 .

Inferential goals in the regression framework

$y_{i,j}$ measures $i \rightarrow j$, $\mathbf{x}_{i,j}$ is a vector of explanatory variables.

$$\mathbf{Y} = \begin{pmatrix} y_{1,1} & y_{1,2} & y_{1,3} & \text{NA} & y_{1,5} & \cdots \\ y_{2,1} & y_{2,2} & y_{2,3} & y_{2,4} & y_{2,5} & \cdots \\ y_{3,1} & \text{NA} & y_{3,3} & y_{3,4} & \text{NA} & \cdots \\ y_{4,1} & y_{4,2} & y_{4,3} & y_{4,4} & y_{4,5} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_{1,1} & \mathbf{x}_{1,2} & \mathbf{x}_{1,3} & \mathbf{x}_{1,4} & \mathbf{x}_{1,5} & \cdots \\ \mathbf{x}_{2,1} & \mathbf{x}_{2,2} & \mathbf{x}_{2,3} & \mathbf{x}_{2,4} & \mathbf{x}_{2,5} & \cdots \\ \mathbf{x}_{3,1} & \mathbf{x}_{3,2} & \mathbf{x}_{3,3} & \mathbf{x}_{3,4} & \mathbf{x}_{3,5} & \cdots \\ \mathbf{x}_{4,1} & \mathbf{x}_{4,2} & \mathbf{x}_{4,3} & \mathbf{x}_{4,4} & \mathbf{x}_{4,5} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

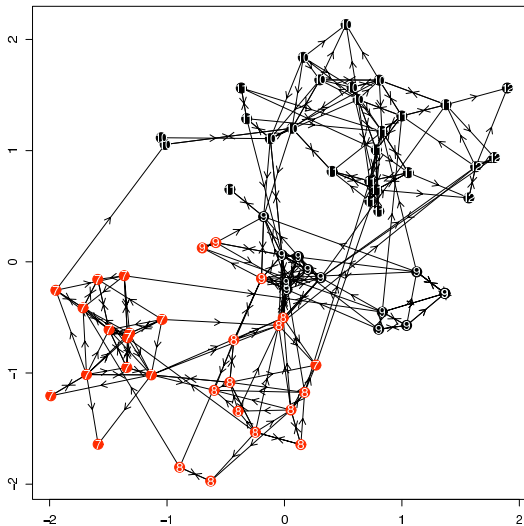
Consider a basic (generalized) linear model

$$y_{i,j} \sim \beta' \mathbf{x}_{i,j} + e_{i,j}$$

A model can provide

- ▶ a measure of the association between \mathbf{X} and \mathbf{Y} : $\hat{\beta}$, $\text{se}(\hat{\beta})$
- ▶ predictions of missing or future observations: $p(y_{1,4} | \mathbf{Y}, \mathbf{X})$

How might node heterogeneity affect network structure?



Latent variable models

Deviations from ordinary regression models can be represented as

$$y_{i,j} \sim \beta' \mathbf{x}_{i,j} + e_{i,j}$$

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

$$e_{i,j} = u_i + v_j + \epsilon_{i,j} \quad \Rightarrow \quad y_{i,j} \sim \beta' \mathbf{x}_{i,j} + u_i + v_j + \epsilon_{i,j}$$

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

E represents deviations from the regression model

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

$$\{e_{i,j}\} \stackrel{d}{=} \{e_{gi,hj}\}$$

for all permutations g and h . 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 $\{e_{i,j}\} \stackrel{d}{=} \{e_{i,j}^*\}$, where

$$e_{i,j}^* = f(\mu, u_i, v_j, \epsilon_{i,j})$$

and $\mu, \{u_i\}, \{v_j\}, \{\epsilon_{i,j}\}$ are all independent random variables.

The singular value decomposition model

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

\mathbf{M} represents “systematic” patterns and \mathcal{E} represents “noise”. Every \mathbf{M} has a representation of the form $\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{V}'$ where, in the case $m \geq n$,

\mathbf{U} is an $m \times n$ matrix with orthonormal columns;

\mathbf{V} is an $n \times n$ matrix with orthonormal columns;

\mathbf{D} is an $n \times n$ diagonal matrix, with diagonal elements $\{d_1, \dots, d_n\}$ typically taken to be a decreasing sequence of non-negative numbers.

Recall,

- ▶ The squared elements of the diagonal of \mathbf{D} are the eigenvalues of $\mathbf{M}'\mathbf{M}$ and the columns of \mathbf{V} are the corresponding eigenvectors.
- ▶ The matrix \mathbf{U} can be obtained from the first n eigenvectors of $\mathbf{M}\mathbf{M}'$. The number of non-zero elements of \mathbf{D} is the rank of \mathbf{M} .
- ▶ Writing the row vectors as $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$, $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$, $m_{i,j} = \mathbf{u}_i' \mathbf{D} \mathbf{v}_j$.

Data analysis with the singular value decomposition

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

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

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

Interpretation: $y_{i,j} = \mathbf{u}_i' \mathbf{D} \mathbf{v}_j + \epsilon_{i,j}$, \mathbf{u}_i and \mathbf{v}_j are the i th, j th rows of \mathbf{U} ,

Estimation: $\hat{\mathbf{M}}_K = \hat{\mathbf{U}}_{[1:K]} \hat{\mathbf{D}}_{[1:K,1:K]} \hat{\mathbf{V}}'_{[1:K]}$ if \mathbf{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 $\hat{\mathbf{M}}_K$ a good estimator?
($\mathbf{E}[\mathbf{Y}'\mathbf{Y}] = \mathbf{M}'\mathbf{M} + m\sigma^2\mathbf{I}$)
- ▶ Work on model-based factor analysis: Rajan and Rayner (1997), Minka (2000), Lopes and West (2004).

Generalized bilinear regression

$$y_{i,j} \sim \beta' \mathbf{x}_{i,j} + \mathbf{u}'_i \mathbf{D} \mathbf{v}_j + \epsilon_{i,j}$$

Interpretation:

Think of $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$, $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ as vectors of **latent nodal attributes**:

$$\mathbf{u}'_i \mathbf{D} \mathbf{v}_j = \sum_{k=1}^K d_k u_{i,k} v_{j,k}$$

In general, a latent variable model relating \mathbf{X} to \mathbf{Y} is

$$g(y_{i,j}) = \beta' \mathbf{x}_{i,j} + \mathbf{u}'_i \mathbf{D} \mathbf{v}_j + \epsilon_{i,j}$$

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

- ▶ If $y_{i,j}$ is binary, $\log \text{odds}(y_{i,j} = 1) = \beta' \mathbf{x}_{i,j} + \mathbf{u}'_i \mathbf{D} \mathbf{v}_j + \epsilon_{i,j}$
- ▶ If $y_{i,j}$ is count data, $\log E[y_{i,j}] = \beta' \mathbf{x}_{i,j} + \mathbf{u}'_i \mathbf{D} \mathbf{v}_j + \epsilon_{i,j}$
- ▶ If $y_{i,j}$ is continuous, $E[y_{i,j}] = \beta' \mathbf{x}_{i,j} + \mathbf{u}'_i \mathbf{D} \mathbf{v}_j + \epsilon_{i,j}$

Estimation: Given \mathbf{D} , \mathbf{V} , the predictor is linear in \mathbf{U} . This bilinear structure can be exploited (EM, Gibbs sampling, variational methods).

International conflict network, 1990-2000

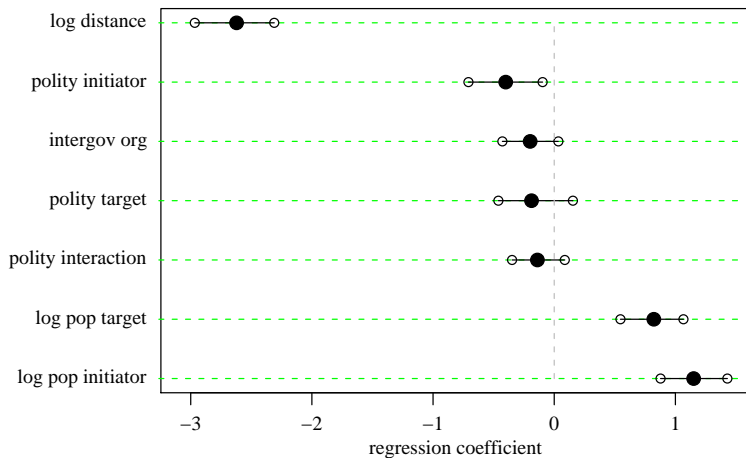
11 years of international relations data (Mike Ward and Xun Cao)

- ▶ $y_{i,j}$ = indicator of a militarized disputes initiated by i with target j ;
- ▶ $\mathbf{x}_{i,j}$ 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

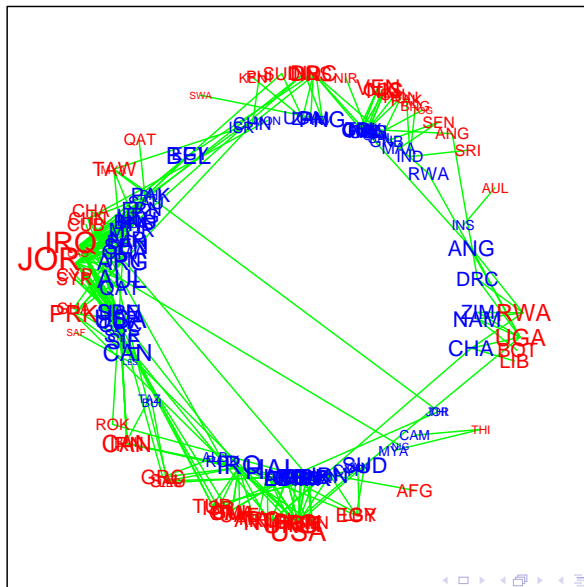
Model: $y_{i,j}$ are independent binary random variables with log-odds

$$\log \text{odds}(y_{i,j} = 1) = \beta' \mathbf{x}_{i,j} + \mathbf{u}'_j \mathbf{D} \mathbf{v}_j + \epsilon_{i,j}$$

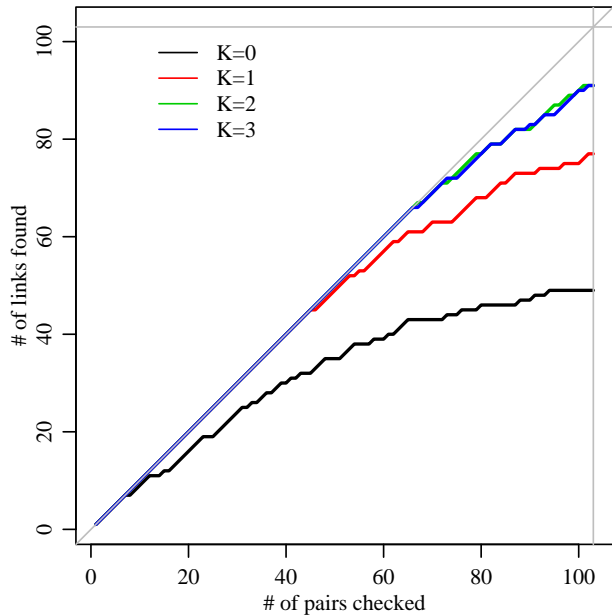
International conflict network: parameter estimates



International conflict network: description of latent variation



International conflict network: prediction experiment



Analyzing undirected data

In many applications $y_{i,j} = y_{j,i}$ by design, and so

$$(y_{i,j} = y_{j,i}) \sim \beta' \mathbf{x}_{i,j} + e_{i,j}$$

and $\mathbf{E} = \{e_{i,j}\}$ is a symmetric array. How should \mathbf{E} be modeled?

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

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

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

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

$$\begin{aligned} \mathbf{M} &= \mathbf{U} \mathbf{\Lambda} \mathbf{U}' \\ m_{i,j} &= \mathbf{u}'_i \mathbf{\Lambda} \mathbf{u}_j \end{aligned}$$

This suggests a model of the form

$$(y_{i,j} = y_{j,i}) \sim \beta' \mathbf{x}_{i,j} + \mathbf{u}'_i \mathbf{\Lambda} \mathbf{u}_j + \epsilon_{i,j}$$

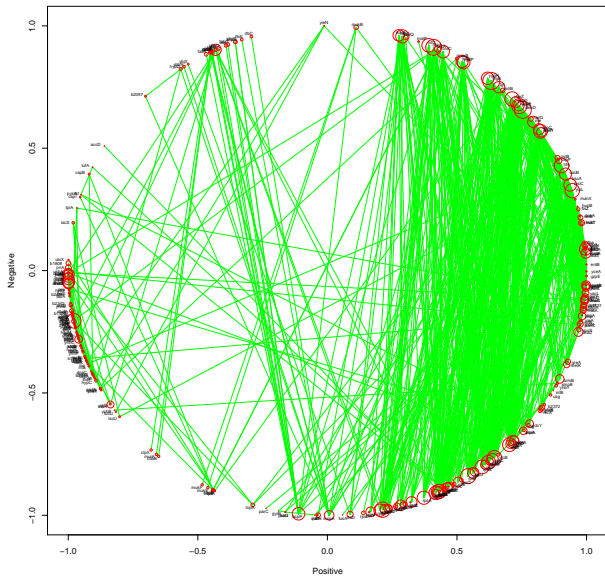
Protein-protein interaction network

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

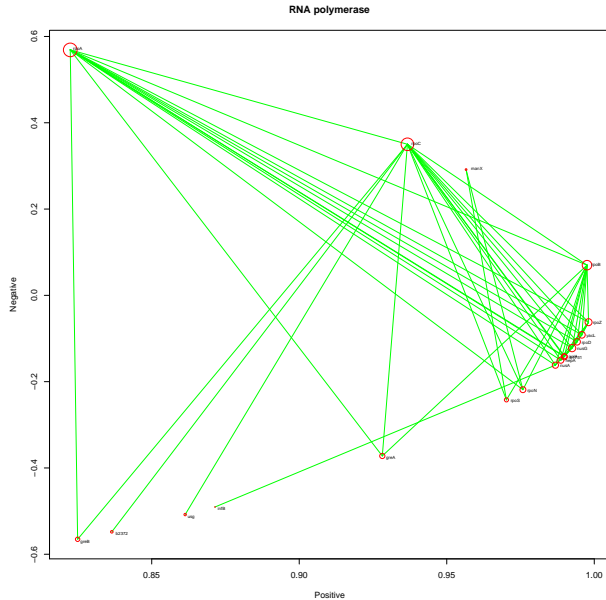
$$\text{Data: } \frac{1}{\binom{270}{2}} \sum_{i < j} y_{i,j} \approx 0.01$$

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

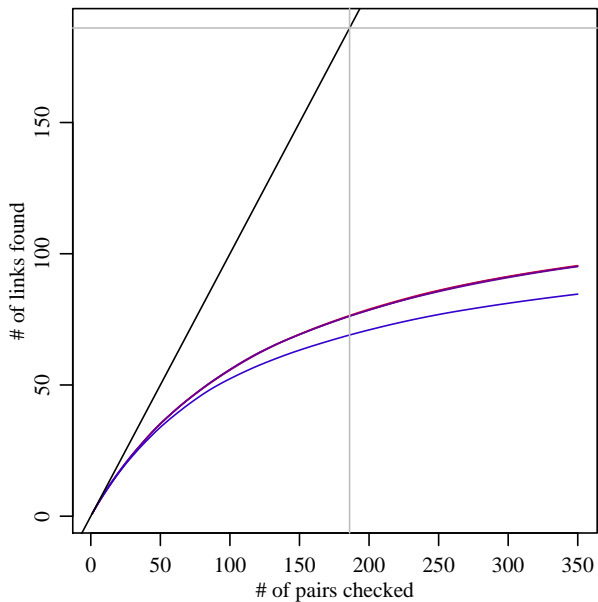
Protein-protein interaction network



Protein-protein interaction network



Protein-protein interaction network: prediction experiment



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, \dots, A_p represent classes of objects.

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

Examples:

International relations

$A_1 = A_2 =$ countries, $A_3 =$ time,

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

Social networks

$A_1 = A_2 =$ individuals, $A_3 =$ information sources,

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

Document analysis

$A_1 = A_2 =$ words, $A_3 =$ documents,

$y_{i,j,k}$ = co-occurrence of words i and j in document k .

Factor models for multiway data

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

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

Now generalize to a three-way array:

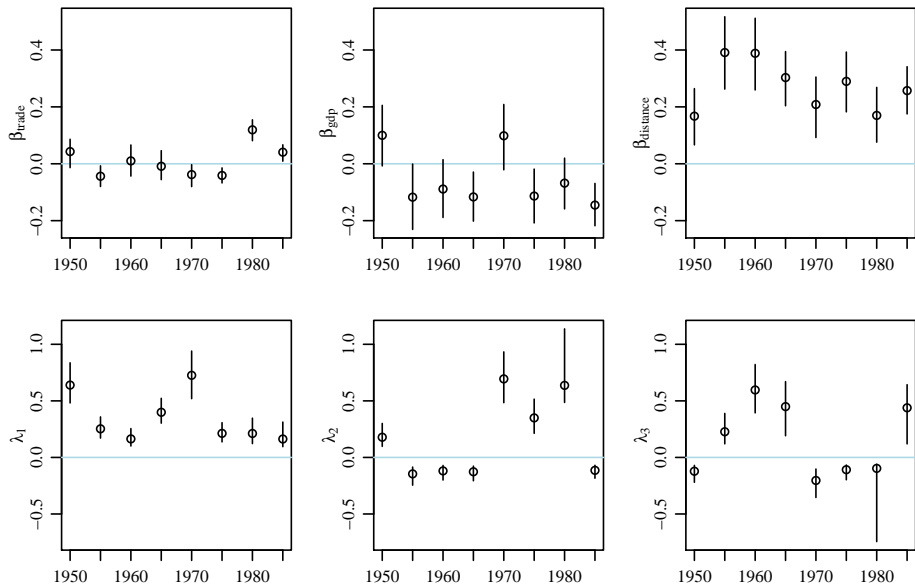
$$m_{i,j,k} = \sum_{r=1}^R u_{i,r} v_{j,r} w_{k,r} d_r$$

- ▶ $\{\mathbf{u}_1, \dots, \mathbf{u}_{n_1}\}$ represents variation along the 1st dimension
- ▶ $\{\mathbf{v}_1, \dots, \mathbf{v}_{n_2}\}$ represents variation along the 2nd dimension
- ▶ $\{\mathbf{w}_1, \dots, \mathbf{w}_{n_3}\}$ represents variation along the 3rd dimension

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

$$\begin{aligned} m_{i,j,k} &= \sum_{r=1}^R u_{i,r} v_{j,r} w_{k,r} d_r \\ &= \sum_{r=1}^R u_{i,r} v_{j,r} d_{k,r} = \mathbf{u}_i' \mathbf{D}_k \mathbf{v}_j \quad \text{where } d_{k,r} = w_{k,r} d_k \end{aligned}$$

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