Hierarchical eigenmodels for relational data

Peter Hoff

Statistics, Biostatistics and the CSSS
University of Washington

Outline

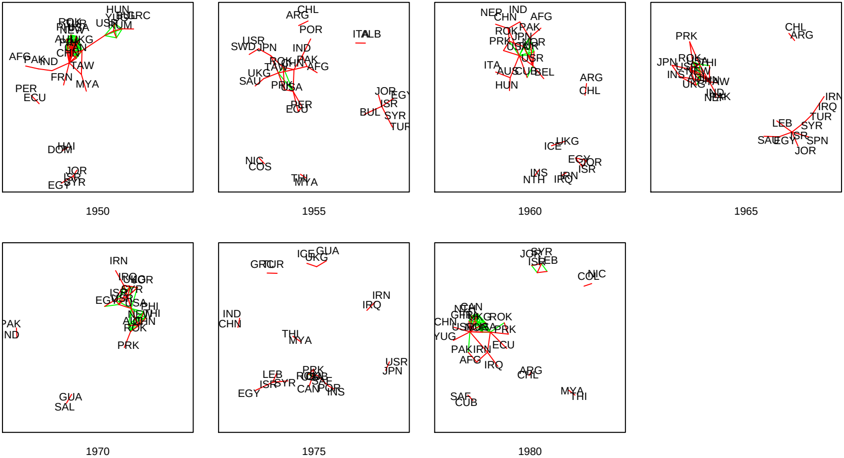
Multivariate matrix data

MMD refers to measurements of various types under various conditions:

- Longitudinal network data: $\mathbf{Y}$ an $n \times m \times T$ array
  - $y_{i,j,t}$ = friendship between people i and j at time t
  - $y_{i,j,t}$ = conflict between countries i and j at time t
- Multivariate relational data: $\mathbf{Y}$ and $n \times m \times p$ array
  - $y_{i,j,1}$ = friendship between people i and j, $y_{i,j,2}$=coworker status, ...
  - $y_{i,j,1}$ = conflict between countries i and j, $y_{i,j,2}$ =trade, ...
  - $y_{i,j,k}$ = measurement under factor 1=i, factor 2= j in block k
- Multigroup multivariate data: $\{Y_k \in \mathbb{R}^{n_k \times p}; k = 1, \ldots, K\}$
  - $y_{i,j,k}$ = expression data of gene j for person i in group k
  - $y_{i,j,k}$ = score of student i on question j in school k

Cold War data

Cooperation and conflict data collected on 85 countries every fifth year



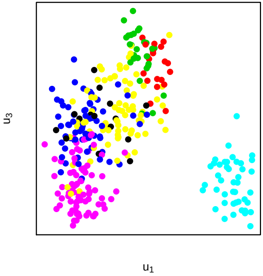
How can we numerically describe variability, similarity across $\mathbf{Y}_1, \ldots, \mathbf{Y}_7$?

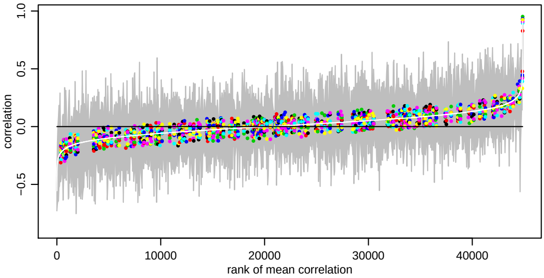
Leukemia data

Gene expression data on 327 cancer patients, each in one of seven groups:

group	BCR	E2A	Hyperdip50	MLL	T	TEL	other
sample size	15	27	64	20	43	79	79

We look at the 300 genes with highest rank variation across subjects.





$\mathbf{Y} = \mathbf{UDV}^T$ Left-singular vectors of $\mathbf{U}$ separate the groups.

$\mathbf{Y}_k = \mathbf{U}_k\mathbf{D}_k\mathbf{V}_k^T$ How do correlations $\mathbf{V}_k\mathbf{D}_k^2\mathbf{V}_k^T$ vary across groups?

Reduced rank matrix approximation

Low rank approximations are useful for describing row/column variability:

Symmetric matrices: $\mathbf{Y} = \mathbf{U}\Lambda\mathbf{U}^T + \mathbf{E}$, $y_{i,j} = \mathbf{u}_i^T\Lambda\mathbf{u}_j + e_{i,j}$

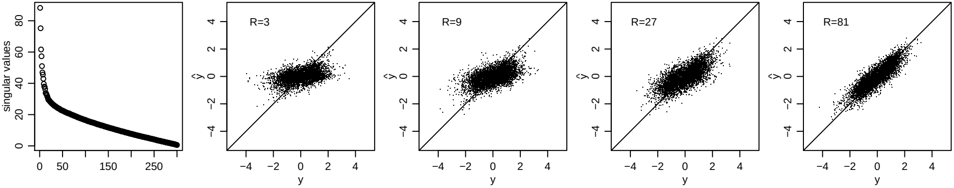
Rectangular matrices: $\mathbf{Y} = \mathbf{U}\mathbf{D}\mathbf{V}^T + \mathbf{E}$, $y_{i,j} = \mathbf{u}_i^T\mathbf{D}\mathbf{v}_j + e_{i,j}$

The column dimension R of $\mathbf{U}$ is generally much smaller than that of $\mathbf{Y}$,

$$R << \min(m, n)$$

so that $\mathbf{U}\Lambda\mathbf{U}^T$, $\mathbf{U}\mathbf{D}\mathbf{V}^T$ provide low-rank approximations to $\mathbf{Y}$.

$$\min_{\mathbf{M}:\mathrm{rank}(\mathbf{M})=R} ||\mathbf{Y} - \mathbf{M}||^2 = ||\mathbf{Y} - \hat{\mathbf{U}}_{[,1:R]}\hat{\mathbf{D}}_{[1:R,1:R]}\hat{\mathbf{V}}_{[,1:R]}^T||^2$$



Model-based estimation

$$\mathbf{Y}_{m\times n} = \mathbf{UDV}^T + \mathbf{E}$$

- $\mathbf{U}$ and $\mathbf{V}$ are $m \times R$ and $n \times R$ orthonormal matrices ;
- $\mathbf{D}$ is a diagonal matrix of positive numbers;
- $\mathbf{E}$ is a matrix of i.i.d. Gaussian noise.

Parameters to estimate include $\mathbf{U}, \mathbf{D}, \mathbf{V}$ and the error variance. Why not just use the SVD?

- Estimation: MSE of LS estimate can be very high.
- Missing data and prediction.
- A model accommodates regression, non-normal and hierarchical data.

Pooling information

Consider p variables measured on individuals in K groups, and let $\mathbf{Y}$ be the $n_k \times p$ data matrix for group k.

$$\begin{aligned} \mathbf{Y}_1 &= \mathbf{U}_1\mathbf{D}_1\mathbf{V}_1^T + \mathbf{E}_1 \\ \vdots \quad &\vdots \quad \vdots \\ \mathbf{Y}_K &= \mathbf{U}_K\mathbf{D}_K\mathbf{V}_K^T + \mathbf{E}_K \end{aligned}$$

Recall, $E[\mathbf{Y}_k^T\mathbf{Y}] = \mathbf{V}_k\mathbf{D}_k^2\mathbf{V}_k^T$, so $\mathbf{V}_k$ represents the covariance/principle components of the observations in group k. Should we

- assume $\mathbf{V}_1 = \mathbf{V}_2 = \cdots = \mathbf{V}_K$?
- estimate each $\mathbf{V}_k$ separately (perhaps using SVD)?
- do something in-between?

$$\hat{\mathbf{V}}_k = w_k\tilde{\mathbf{V}}_k + (1 - w_k)\sum_{j \neq k} \theta_k \tilde{\mathbf{V}}_j$$

A model for heterogeneity among $\{\mathbf{V}_1, \ldots, \mathbf{V}_K\}$ would help determine the right balance.

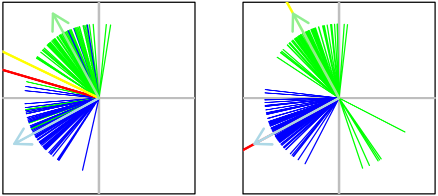
The matrix Langevin distribution

$$\mathbf{V}_1, \ldots, \mathbf{V}_K \sim \text{i.i.d. } p(\mathbf{V}) \propto \operatorname{etr}(\mathbf{M}^T\mathbf{V})$$

where $\mathbf{M}$ is any $p \times R$ matrix. It is convenient to reparameterize:

$$\begin{aligned} \mathbf{M} &= \mathbf{A}\mathbf{B}\mathbf{C}^T \\ &= \mathbf{A}\mathbf{C}^T\mathbf{C}\mathbf{B}\mathbf{C}^T \\ &= \mathbf{H}[\mathbf{C}\mathbf{B}\mathbf{C}^T] \end{aligned}$$

- $\mathbf{H} \in \mathcal{V}_{p,R}$ and is the mode of $\mathbf{V}$.
- $\mathbf{C}\mathbf{B}\mathbf{C}^T$ is positive definite and describes covariation.
- If $\mathbf{M}$ is orthogonal then $\mathbf{C} = \mathbf{I}$ and $\operatorname{tr}(\mathbf{M}^T\mathbf{V}) = \sum_{r=1}^R b_{r,r}\mathbf{h}_r^T\mathbf{v}_r$.



A hierarchical eigenmodel

$$
\begin{aligned}
\mathbf{Y}_1 &= \mathbf{U}_1 \mathbf{D}_1 \mathbf{V}_1^T + \mathbf{E}_1 \\
\vdots \quad & \vdots \quad \vdots \\
\mathbf{Y}_K &= \mathbf{U}_K \mathbf{D}_K \mathbf{V}_K^T + \mathbf{E}_K
\end{aligned}
$$

$$
\begin{array}{ccc}
\mathbf{U}_1 \sim \text{uniform}(\mathcal{V}_{n_1,R}) & \text{diag}(\mathbf{D}_1) \sim \text{normal}(\mathbf{0}, \tau^2 I) & \mathbf{V}_1 \sim \text{Langevin}(\mathbf{M}) \\
 & \vdots & \\
\mathbf{U}_K \sim \text{uniform}(\mathcal{V}_{n_K,R}) & \text{diag}(\mathbf{D}_K) \sim \text{normal}(\mathbf{0}, \tau^2 I) & \mathbf{V}_K \sim \text{Langevin}(\mathbf{M})
\end{array}
$$

$$
\begin{aligned}
\mathbf{M} &= \mathbf{ABC}^T \\
\mathbf{A} &\sim \text{uniform}(\mathcal{V}_{p,R}) \\
\text{diag}(\mathbf{B}) &\sim \text{normal}^+(\mathbf{0}, \eta^2 I) \\
\mathbf{C} &\sim \text{uniform}(\mathcal{V}_{R,R})
\end{aligned}
$$

Full conditional distributions

$$
\begin{aligned}
p(\mathbf{V}_1, \ldots, \mathbf{V}_K | \mathbf{A}, \mathbf{B}, \mathbf{C}^T) &= \prod_{k=1}^{K} c(\mathbf{B}) \mathrm{etr}(\mathbf{C}\mathbf{B}\mathbf{A}^T \mathbf{V}_k) \\
&= c(\mathbf{B})^K \mathrm{etr}(K\mathbf{C}\mathbf{B}\mathbf{A}^T \bar{\mathbf{V}})
\end{aligned}
$$

Evidently,

$$
\begin{aligned}
p(\mathbf{A} | \mathbf{V}_1, \ldots, \mathbf{V}_K, \mathbf{B}, \mathbf{C}) &\propto \mathrm{etr}([K\bar{\mathbf{V}}\mathbf{C}\mathbf{B}]^T \mathbf{A}) \\
p(\mathbf{C} | \mathbf{V}_1, \ldots, \mathbf{V}_K, \mathbf{A}, \mathbf{B}) &\propto \mathrm{etr}([K\bar{\mathbf{V}}^T \mathbf{A}\mathbf{B}]^T \mathbf{C})
\end{aligned}
$$

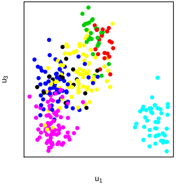
Additionally,

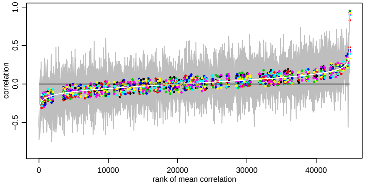
- The full conditional of $\mathbf{B}$ is nonstandard but low-dimensional.
- The full conditional distributions of $\{\mathbf{U}_k\}$ and $\{\mathbf{V}_k\}$ are Langevin.
- Full conditional distributions of $\{\mathbf{D}_k, \sigma_k\}$ are standard.

Gibbs sampling can be implemented with the aid of a rejection sampler for the matrix Langevin distribution.

Leukemia data analysis

group	BCR	E2A	Hyperdip50	MLL	T	TEL	other
sample size	15	27	64	20	43	79	79





Data $\mathbf{Y}$ is 327×300, or can be broken into $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_7\}$ of variable row dimension but common column dimension.

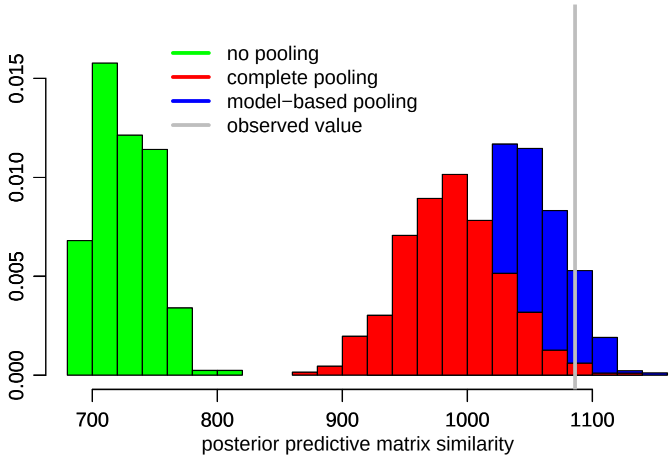
We'll fit the hierarchical eigenmodel and evaluate its goodness-of-fit using the matrix similarity statistic

$$t(\mathbf{Y}_1, \ldots, \mathbf{Y}_7) = \sum_{i<j} \mathrm{tr}(|\mathbf{A}_i^T \mathbf{A}_j|),$$

where $\mathbf{A}_k$ is $\mathbf{Y}_k$ with the "subject effects removed:"

$$\begin{aligned} \mathbf{Y}_k &= \hat{\mathbf{U}}\hat{\mathbf{D}}\hat{\mathbf{V}}^T \\ \mathbf{A}_k &= \hat{\mathbf{D}}\hat{\mathbf{V}}^T \end{aligned}$$

Goodness of fit



International relations data

Ordered probit model for discrete data:

$$y_{i,j,k} = \sum_{x \in \{-1,0,+1\}} x \delta_{I_x}(z_{i,j.k})$$

Eigenvalue decomposition model for latent $\mathbf{Z}$:

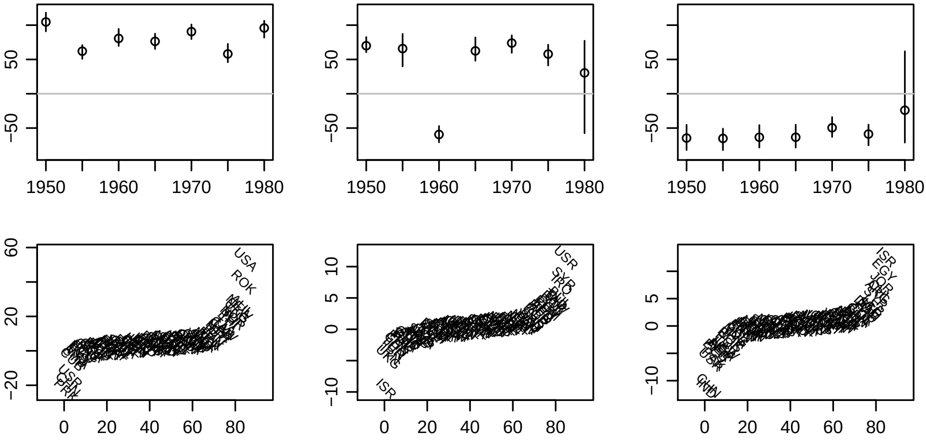
$$\begin{aligned} \mathbf{Z}_k &= \mathbf{U}_k \Lambda_k \mathbf{U}_k^T + \mathbf{E}_k \\ \mathbf{U}_1, \ldots, \mathbf{U}_K &\sim \text{i.i.d. Langevin}(\mathbf{M}) \\ \Lambda_1, \ldots, \Lambda_K &\sim \text{i.i.d. mvn}(\mathbf{0}, \tau^2 \mathbf{I}) \end{aligned}$$

Parameter estimation is similar to before: Letting $\mathbf{M} = \mathbf{ABC}^T$,

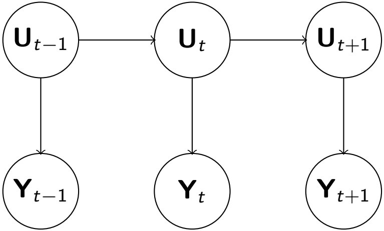
$$\begin{aligned} p(\mathbf{A}|\mathbf{U}_1, \ldots, \mathbf{U}_K, \mathbf{B}, \mathbf{C}) &\propto \mathrm{etr}([K\bar{\mathbf{U}}\mathbf{CB}]^T \mathbf{A}) \\ p(\mathbf{C}|\mathbf{U}_1, \ldots, \mathbf{U}_K, \mathbf{A}, \mathbf{B}) &\propto \mathrm{etr}([K\bar{\mathbf{U}}^T \mathbf{AB}]^T \mathbf{C}), \quad \text{although} \\ p(\mathbf{U}_k|\mathbf{Z}_k, \mathbf{M}) &\propto \mathrm{etr}(\mathbf{M}^T \mathbf{U}_k + \mathbf{U}_k \mathbf{Z}_k \mathbf{U}_k^T) \end{aligned}$$

This last distribution is a Bingham-Langevin distribution.

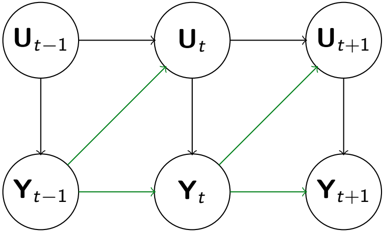
International relations data



Longitudinal social networks

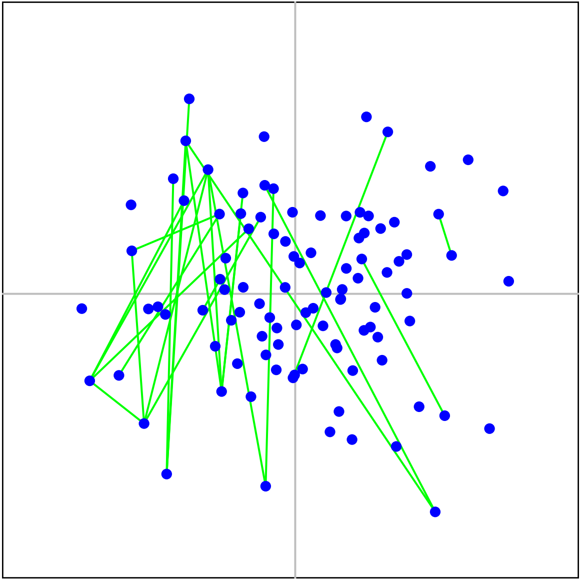


$$
\begin{aligned}
\mathbf{U}_t &\sim \text{Langevin}(\mathbf{U}_{t-1}\boldsymbol{\Sigma}) \\
\mathbf{Y}_t &\sim \text{probit}(\mathbf{U}_t \Lambda_t \mathbf{U}_t^T)
\end{aligned}
$$

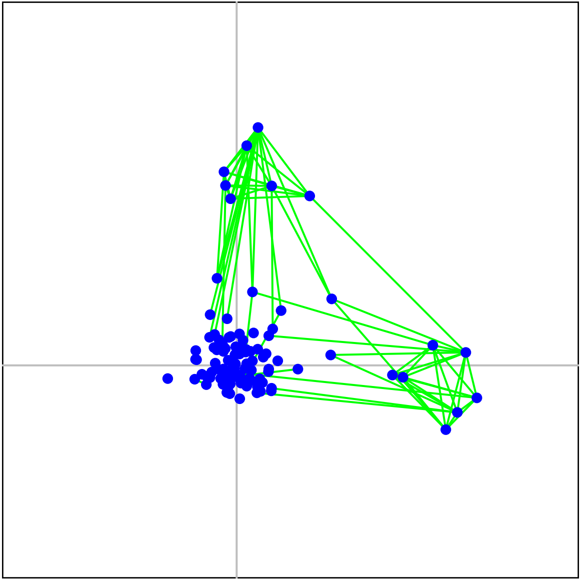


$$
\begin{aligned}
\mathbf{U}_t &\sim \text{Langevin}(\mathbf{U}_{t-1}\boldsymbol{\Sigma} + \alpha \mathbf{Y}_{t-1}\mathbf{U}_{t-1}) \\
\mathbf{Y}_t &\sim \text{probit}(\mathbf{U}_t \Lambda_t \mathbf{U}_t^T + \beta \mathbf{Y}_{t-1})
\end{aligned}
$$

A simulated network



Another simulated network



Discussion

Summary:

- SVD and EVD are natural ways to describe matrix patterns.
- Variability across matrices can be described by variability across decompositions.
- Modeling variability allows for information sharing across datasets.
- Parameter estimation can be done with Gibbs sampling.

Caveats:

- Interpretation of parameters is subtle.
- Models are more "statistical" than "generative."