Hierarchical models for multiway data

Peter Hoff

Statistics, Biostatistics and the CSSS
University of Washington
Array-valued data

$y_{i,j,k} =$

- $j$th variable of $i$th subject under condition $k$ (psychometrics).
- type-$k$ relationship between $i$ and $j$ (relational data/network).
- sample mean of variable $i$ for group $j$ in state $k$ (cross-classified data)
Cold war cooperation and conflict

- 66 countries
- 8 years (1950, 1955, ..., 1980, 1985)
- $y_{i,j,t} =$ relation between $i,j$ in year $t$
- also have data on gdp, polity
\{y_i : x_i = x\} \overset{\text{iid}}{\sim} \text{multivariate normal}(\mu_x, \Sigma)

- $n = 1116$ survey participants
- $4 \times 4 \times 2 \times 4 = 128$ levels of $x$
- $\{\beta_x\}$ a $4 \times 4 \times 2 \times 4 \times 2$ array
- $> 1/2$ levels have $\leq 5$ samples


**Reduced rank models**

\[ Y = \Theta + E \]

- \( \Theta \) contains the “main features” we hope to recover,
- \( E \) is “patternless.”

**Matrix decomposition:** If \( \Theta \) is a rank-\( R \) matrix, then

\[ \theta_{i,j} = \langle u_i, v_j \rangle = \sum_{r=1}^{R} u_{i,r} v_{j,r} \quad \Theta = \sum_{r=1}^{R} u_r v_r^T = \sum_{r=1}^{R} u_r \otimes v_r \]

**Array decomposition:** If \( \Theta \) is a rank-\( R \) array, then

\[ \theta_{i,j,k} = \langle u_i, v_j, w_k \rangle = \sum_{r=1}^{R} u_{i,r} v_{j,r} w_{k,r} \quad \Theta = \sum_{r=1}^{R} u_r \otimes v_r \otimes w_r \]

(Harshman[1970], Kruskal[1976,1977], Harshman and Lundy[1984], Kruskal[1989])
Some things you should know

1. Computing the rank
   - matrix: easy to do
   - array: no known algorithm

2. Possible rank
   - matrix: $R_{\text{max}} = \min(m_1, m_2)$
   - array: $\max(m_1, m_2, m_3) \leq R_{\text{max}} \leq \min(m_1 m_2, m_1 m_3, m_2 m_3)$

3. Probable rank
   - matrix: “almost all” matrices have full rank.
   - array: a nonzero fraction (w.r.t. Lebesgue measure) have less than full rank.

4. Least squares approximation
   - matrix: SVD of $Y$ provides the rank $R$ least-squares approximation to $\Theta$.
   - array: iterative “least squares” methods, but solution may not exist (de Silva and Lim[2008])

5. Uniqueness
   - matrix: The representation $\Theta = \langle U, V \rangle = UV^T$ is not unique.
   - array: The representation $\Theta = \langle U, V, W \rangle$ is essentially unique.
A model-based approach

For a $K$-way array $Y$,

$$
Y = \Theta + E
$$

$$
\Theta = \sum_{r=1}^{R} u^{(1)}_r \otimes \ldots \otimes u^{(K)}_r \equiv \langle U^{(1)}, \ldots, U^{(K)} \rangle
$$

$$
u^{(k)}_1, \ldots, u^{(k)}_{m_k} \overset{iid}{\sim} \text{multivariate normal}(\mu_k, \Psi_k),
$$

with $\{\mu_k, \Psi_k, k = 1, \ldots, K\}$ to be estimated.

Some motivation:

- shrinkage: $\Theta$ contains lots of parameters.
- hierarchical: covariance among columns of $U^{(k)}$ is identifiable.
- estimation: $p(Y|U^{(1)}, \ldots, U^{(K)})$ multimodal, MCMC “stochastic search”
- adaptability: incorporate reduced rank arrays as a model component
  - multilinear predictor in a GLM
  - multilinear effects for regression parameters
$K = 3 \quad , \quad R = 4 \quad , \quad (m_1, m_2, m_3) = (10, 8, 6)$

1. Generate $M$, a random array of roughly full rank
2. Set $\Theta = \text{ALS}_4(M)$
3. Set $Y = \Theta + E$, $\{e_{i,j,k}\} \sim \text{iid normal}(0, v(\Theta)/4)$.

For each of 100 such simulated datasets, we obtain $\hat{\Theta}_{\text{LS}}$ and $\hat{\Theta}_{\text{HB}}$. 
Simulation study: known rank

- Bayesian MSE vs. least squares MSE
- Bayesian RSS vs. least squares RSS
- Black dots represent mode, gray dots represent mean
Simulation study: misspecified rank

least squares

log RSS

-2.5
-1.5
-0.5

assumed rank

1 2 3 4 5 6 7 8

log MSE

-3
-2
-1
0

hierarchical Bayes

-2.5
-1.5
-0.5

assumed rank

1 2 3 4 5 6 7 8
Simulation study: comments on rank selection

A hierarchical model - try DIC:

$R_{true} = 2 \quad \Pr(\hat{R} = r) = \{0.10, 0.74, 0.07, 0.05, 0.02, 0.01, 0.01, 0.00\}$

$R_{true} = 4 \quad \Pr(\hat{R} = r) = \{0.08, 0.15, 0.27, 0.28, 0.06, 0.07, 0.04, 0.05\}$

$R_{true} = 6 \quad \Pr(\hat{R} = r) = \{0.07, 0.18, 0.19, 0.17, 0.10, 0.08, 0.09, 0.12\}$

Keep in mind:

A rank $\hat{R} < R_{true}$ estimate might be better than the rank $R_{true}$ estimate.
The 2008 General Social Survey includes data on the following six variables:

- $y_1$ (words): number of correct answers out of 10 on a vocabulary test;
- $y_2$ (tv): hours of television watched in a typical day;
- $x_1$ (deg) highest degree obtained: none, high school, Bachelor’s, graduate;
- $x_2$ (age): 18-34, 35-47, 48-60, 61 and older;
- $x_3$ (sex): male or female;
- $x_4$ (child) number of children: 0, 1, 2, 3 or more.

**Nominal goal:** Estimate $E[y|x]$ for each of the 128 possible $x$-vectors.
Deep interaction example

Sampling model:

\[ \{ y_i : x_i = x \} \overset{\text{iid}}{\sim} \text{multivariate normal}(\mu_x, \Sigma) \]

Mean model:

\[
\begin{align*}
\mu_x &= \beta_x + \gamma_x \\
\{ \beta_x : x \in \mathcal{X} \} &= B \text{ is of reduced rank} \\
\{ \gamma_x : x \in \mathcal{X} \} \overset{\text{iid}}{\sim} \text{multivariate normal}(0, \Omega)
\end{align*}
\]

- This is a **full** model: \( \mu_x \) is unconstrained.
- This is a **hierarchical** model: \( \hat{\mu}_x \) borrows information from other \( x \)-groups.
Deep interaction example
Conflict and cooperation during the cold war

- $y_{i,j,t} \in \{-5, -4, \ldots, +1, +2\}$, the level of military conflict/cooperation
- $x_{i,j,t,1} = \log \text{gdp}_i + \log \text{gdp}_j$, the sum of the log gdps of the two countries;
- $x_{i,j,t,2} = (\log \text{gdp}_i) \times (\log \text{gdp}_j)$, the product of the log gdps;
- $x_{i,j,t,3} = \text{polity}_i \times \text{polity}_j$, where $\text{polity}_i \in \{-1, 0, +1\}$;
- $x_{i,j,t,4} = (\text{polity}_i > 0) \times (\text{polity}_j > 0)$.
Longitudinal network example
Longitudinal network example
Longitudinal network example
Longitudinal network example
Longitudinal network example
Longitudinal network example
Longitudinal network example
Multiway models provide a parsimonious representation of mean structure.
Bayesian estimation provides regularized estimates.
A Hierarchical model provides adaptive regularization.
Array structures can be incorporated into more complicated models.