(Bayesian) Statistics with Rankings

Marina Meilă
University of Washington
with Chris Meek, Raman Arora, Alnur Ali, Brendan Murphy, Harr Chen, Bhushan Mandhani, Le Bao, Kapil Phadnis, Artur Patterson, Jeff Bilmes

Columbia University 4/11/16
Permutations (rankings) data represents preferences

Burger preferences $n = 6$ options, $N = 600$ “voters”
med-rare med rare ...
done med-done med ...
med-rare rare med ...

Presidential Election Ireland, 2000 $n = 5$ candidates, $N = 1100$ voters
Roch Scal McAl Bano Nall
Scal McAl Nall Bano Roch
Roch McAl

College programs admissions, Ireland $n = 533$ degree programs, $N = 53737$ high-school graduates, $t = 10$
DC116 DC114 DC111 DC148 DB512 DN021 LM054 WD048 LM020 LM050
WD028
DN008 TR071 DN012 DN052
FT491 FT353 FT471 FT541 FT402 FT404 TR004 FT351 FT110 FT352

Sushi preferences $n = 112$, $N = 5000$
sake | ebi | ika | uni | tamago | kappa-maki | tekka-maki | anago | toro | maguro
ebi | kappa-maki | tamago | ika | toro | maguro | tekka-maki | anago | sake | uni
toro | ebi | maguro | ika | tekka-maki | uni | sake | anago | kappa-maki | tamago
tekka-maki | tamago | sake | ebi | ika | kappa-maki | maguro | toro | uni | anago
uni | toro | ebi | anago | maguro | tekka-maki | ika | sake | kappa-maki | tamago

Ranking data
- discrete
- many valued
- combinatorial structure
An optimization problem: Consensus Ranking

Given a set of rankings \( \{\pi_1, \pi_2, \ldots, \pi_N\} \subset S_n \) find the consensus ranking (or central ranking) \( \pi_0 \) that best agrees with the data

**Presidential Election Ireland, 2000**

\( n = 5, N = 1100 \)

Roch Scal McAl Bano Nall
Scal McAl Nall Bano Roch
Roch McAl

Consensus = [ Roch Scal McAl Bano Nall ] ?
The Consensus Ranking problem

**Problem** (also called Preference Aggregation, Kemeny Ranking)
Given a set of rankings \( \{\pi_1, \pi_2, \ldots, \pi_N\} \subset S_n \) find the consensus ranking (or central ranking) \( \pi_0 \) such that

\[
\pi_0 = \arg\min_{\pi_0} \sum_{i=1}^{N} d(\pi_i, \pi_0)
\]

for \( d = \) inversion distance / Kendall \( \tau \)-distance / “bubble sort” distance

Relevance
- Voting in elections (APA, Ireland, Cambridge), panels of experts (admissions, hiring, grant funding)
- Aggregating user preferences (economics, marketing)
- Subproblem of other problems (building a good search engine: leaning to rank [Cohen, Schapire, Singer 99])
Consensus ranking problem

\[ \pi_0 = \arg\min_{\pi \in S_n} \sum_{i=1}^{N} d(\pi_i, \pi_0) \]

This talk

Will generalize the problem

- from finding \( \pi_0 \)
  to estimating statistical model (based on inversions)
  Max Likelihood or Bayesian framework

Will generalize the data

- from complete, finite permutations to
top-t rankings [MBao08]
countably many items \( (n \to \infty) \) [MBao08]
recursive inversion models[MeekM14]
signed permutations [MArora13]
Outline

Permutations and their representations
- Statistical models for permutations and the dependence of ranks
- Codes, inversion distance and the precedence matrix
- Mallows models over permutations

Complete rankings and Maximum Likelihood estimation
- GM as exponential family

Top-t rankings, infinite permutations, and Bayesian estimation
- Top-t rankings and infinite permutations
- Conjugate prior, Dirichlet process mixtures

Recursive inversion models and finding common structure in preferences

[Signed permutations and the reversal median problem]
Some notation

Base set \( \{ a, b, c, d \} \) contains \( n \) items (or alternatives)
E.g \( \{ \text{rare, med-rare, med, med-done, ...} \} \)

\( S_n = \) the symmetric group = the set of all permutations over \( n \) items

\( \pi = [c a b d] \in S_n \) a permutation/ranking

\( \pi = [c a] \) a top-t ranking (is a partial order)

\( t = |\pi| \leq n \) the length of \( \pi \)

We observe

data \( \pi_1, \pi_2, \ldots, \pi_N \sim \) sampled independently from distribution \( P \) over \( S_n \)

(where \( P \) is unknown)
Representations for permutations

- **Reference permutation**: \( id = [a, b, c, d] \)
- **Ranked list**:
  \[
  \pi = [c, a, b, d]
  \]
- **Cycle representation**:
  \[
  (2 \ 3 \ 1)
  \]
- **Function**: function on \([a, b, c, d]\)
- **Permutation matrix**:
  \[
  \Pi = \begin{bmatrix}
  0 & 1 & 0 & 0 \\ 
  0 & 0 & 1 & 0 \\ 
  1 & 0 & 0 & 0 \\ 
  0 & 0 & 0 & 1 
  \end{bmatrix}
  \]
- **Precedence matrix**:
  \[
  Q = \begin{bmatrix}
  - & 1 & 0 & 1 \\ 
  0 & - & 0 & 1 \\ 
  1 & 1 & - & 1 \\ 
  0 & 0 & 0 & - 
  \end{bmatrix}
  \]
  \(Q_{ij} = 1\) if \(i \prec_\pi j\),
- **Code**:
  \[
  (V_1, V_2, V_3) = (1, 1, 0) \\
  (s_1, s_2, s_3) = (2, 0, 0)
  \]
Representations for permutations

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

\[ \pi = [c \ a \ b \ d] \]  
(2 3 1)  
ranked list

\[ \begin{array}{cccc}
0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1
\end{array} \]  
permutation matrix

\[ Q = \begin{array}{cccc}
- & 1 & 0 & 1 \\
0 & - & 0 & 1 \\
1 & 1 & - & 1 \\
0 & 0 & 0 & -
\end{array} \]  
precedence matrix,  
\[ Q_{ij} = 1 \text{ if } i <_\pi j \]

\[ (V_1, V_2, V_3) = (1, 1, 0) \]  
code

\[ (S_1, S_2, S_3) = (2, 0, 0) \]  
code

reference permutation id = [a b c d]
Statistical models for permutations and the dependence of ranks

Several “natural” parametric distributions on $\mathbb{S}_n$ exist. Most suffer from dependencies between parameters.

- item $j$ has utility $\mu_j$
  - sample $u_j = \mu_j + \epsilon_j$, $j = 1: n$ independently
  - sort $(u_j)_{j=1:n} \Rightarrow \pi$

- item $j$ has weight $w_j > 0$
  - sample ranks 1, 2, ... sequentially $\propto$ remaining $w_j$’s

$$P([a, b, \ldots]) \propto \frac{w_a}{\sum_i w_i} \frac{w_b}{\sum_i w_i - w_a} \ldots$$

- inversion between $i$ and $j$ has cost $\alpha_{ij}$

$$P(\pi) \propto \exp \left( - \sum_{i<j} \alpha_{ij} Q_{ij}(\pi) \right)$$

interesting subclasses of the Bradley-Terry

(Generalized) Mallows models (coming next)

- are a subclass of Bradley-Terry models
- do not suffer from these dependencies
<table>
<thead>
<tr>
<th></th>
<th>GM</th>
<th>B-T</th>
<th>P-L</th>
<th>Thurstone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete parameter</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Tractable Z</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>“Easy” * parameter</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>Gauss</td>
</tr>
<tr>
<td>estimation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tractable marginals</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>Gauss**</td>
</tr>
<tr>
<td>Params “interpretable”</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>Gauss</td>
</tr>
</tbody>
</table>

* Refers to continuous parameters
** for top ranks

GM model
- computationally very appealing
- advantage comes from the code: the codes $(V_j), (S_j)$
- discrete parameter makes for challenging statistics
The precedence matrix $Q$

$$\pi = [c\ a\ b\ d]$$

$$Q(\pi) = \begin{array}{cccc}
a & b & c & d \\
-1 & 0 & 1 & a \\
0 & - & 0 & 1 & b \\
1 & 1 & - & 1 & c \\
0 & 0 & 0 & - & d \\
\end{array}$$

$Q_{ij}(\pi) = 1$ iff $i$ before $j$ in $\pi$

$Q_{ij} = 1 - Q_{ji}$

reference permutation $id = [a\ b\ c\ d]$: determines the order of rows, columns in $Q$
The number of inversions of $\pi$ and $Q(\pi)$

$$\pi = [c a b d]$$

$Q(\pi) = $

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>a</td>
</tr>
<tr>
<td>0</td>
<td>-</td>
<td>0</td>
<td>1</td>
<td>b</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>c</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>d</td>
</tr>
</tbody>
</table>

define $L(Q) = \text{sum( lower triangle (Q))}$
The number of inversions of $\pi$ and $Q(\pi)$

$$\pi = [c\ a\ b\ d]$$

$$Q(\pi) = \begin{bmatrix}
 a & b & c & d \\
 -1 & 0 & 1 & a \\
 0 & -1 & 0 & b \\
 1 & 1 & -1 & c \\
 0 & 0 & 0 & -d \\
\end{bmatrix}$$

define $L(Q) = \text{sum( lower triangle (Q))}$ then

$$\#\text{inversions (}\pi\text{)} = L(Q) = d(\pi, \text{id})$$
The inversion distance and $Q$

To obtain $d(\pi, \pi_0)$

1. Construct $Q(\pi)$
2. Sort rows and columns by $\pi_0$
3. Sum elements in lower triangle

\[ \pi = [c\ a\ b\ d], \quad \pi_0 = [b\ a\ d\ c] \]

<table>
<thead>
<tr>
<th></th>
<th>$b$</th>
<th>$a$</th>
<th>$d$</th>
<th>$c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>$\pi_0$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>$\pi_0$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$\pi_0$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

\[ d(\pi, \pi_0) = 4 \]
The inversion distance and $Q$

To obtain $d(\pi, \pi_0)$

1. Construct $Q(\pi)$
2. Sort rows and columns by $\pi_0$
3. Sum elements in lower triangle

$$\pi = [c\ a\ b\ d], \ \pi_0 = [b\ a\ d\ c]$$

\[
\begin{array}{cccc}
 b & a & d & c \\
- & 0 & 1 & 0 \\
1 & - & 1 & 0 \\
0 & 0 & - & 0 \\
1 & 1 & 1 & - \\
\end{array}
\]

$$d(\pi, \pi_0) = 4$$

To obtain $d(\pi_1, \pi_0) + d(\pi_2, \pi_0) + \ldots$

1. Construct $Q(\pi_1), Q(\pi_2), \ldots$
   $$Q = Q(\pi_1) + Q(\pi_2) + \ldots$$
2. Sort rows and columns of $Q$ by $\pi_0$
3. Sum elements in lower triangle of $Q$
The code of a permutation

Example $\pi = [c\ a\ b\ d]$, $\pi_0 = [b\ a\ d\ c]$

<table>
<thead>
<tr>
<th></th>
<th>$a$</th>
<th>$b$</th>
<th>$c$</th>
<th>$d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_2$</td>
<td>$-$</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$S_3$</td>
<td>0</td>
<td>$-$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$S_1$</td>
<td>1</td>
<td>1</td>
<td>$-$</td>
<td>1</td>
</tr>
<tr>
<td>$S_4$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$-$</td>
</tr>
</tbody>
</table>

code

$$(V_1, V_2, V_3) = (1, 1, 0)$$
The code of a permutation

Example $\pi = [c \ a \ b \ d], \quad \pi_0 = [b \ a \ d \ c]$

<table>
<thead>
<tr>
<th></th>
<th>$a$</th>
<th>$b$</th>
<th>$c$</th>
<th>$d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_2$</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$S_3$</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$S_1$</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>$S_4$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$V_1$</th>
<th>$V_2$</th>
<th>$V_3$</th>
<th>$V_4$</th>
</tr>
</thead>
</table>

code

$$(V_1, V_2, V_3) = (1, 1, 0)$$

or

$$(S_1, S_2, S_3) = (2, 0, 0)$$

d($\pi$, id) = 2
The code of a permutation

Example $\pi = [c \ a \ b \ d]$, $\pi_0 = [b \ a \ d \ c]$

<table>
<thead>
<tr>
<th></th>
<th>$a$</th>
<th>$b$</th>
<th>$c$</th>
<th>$d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_2$</td>
<td>$-$</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$S_3$</td>
<td>0</td>
<td>$-$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$S_1$</td>
<td>1</td>
<td>1</td>
<td>$-$</td>
<td>1</td>
</tr>
<tr>
<td>$S_4$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$-$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$V_1$</th>
<th>$V_2$</th>
<th>$V_3$</th>
<th>$V_4$</th>
</tr>
</thead>
</table>

code

$$(V_1, V_2, V_3) = (1, 1, 0)$$

or

$$(S_1, S_2, S_3) = (2, 0, 0)$$

$\text{d}(\pi, \text{id}) = 2$

Codes are defined w.r.t any $\pi_0$

<table>
<thead>
<tr>
<th></th>
<th>$b$</th>
<th>$a$</th>
<th>$d$</th>
<th>$c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_3$</td>
<td>$-$</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$S_2$</td>
<td>1</td>
<td>$-$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$S_4$</td>
<td>0</td>
<td>0</td>
<td>$-$</td>
<td>0</td>
</tr>
<tr>
<td>$S_1$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>$-$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$V_1$</th>
<th>$V_2$</th>
<th>$V_3$</th>
<th>$V_4$</th>
</tr>
</thead>
</table>

code $V_j(\pi|\pi_0)$, $S_j(\pi|\pi_0)$

$$(V_1, V_2, V_3) = (2, 1, 1)$$

or

$$(S_1, S_2, S_3) = (2, 0, 0)$$

$\text{d}(\pi, \text{id}) = 2$$
The code of a permutation

Example $\pi = [c\ a\ b\ d]$, $\pi_0 = [b\ a\ d\ c]$

<table>
<thead>
<tr>
<th></th>
<th>$a$</th>
<th>$b$</th>
<th>$c$</th>
<th>$d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_2$</td>
<td>$-$</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$S_3$</td>
<td>0</td>
<td>$-$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$S_1$</td>
<td>1</td>
<td>1</td>
<td>$-$</td>
<td>1</td>
</tr>
<tr>
<td>$S_4$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$-$</td>
</tr>
</tbody>
</table>

$$V_1 \ V_2 \ V_3 \ V_4$$

$$(V_1, V_2, V_3) = (1, 1, 0)$$

or

$$(S_1, S_2, S_3) = (2, 0, 0)$$

$$d(\pi, \text{id}) = 2$$

Codes are defined w.r.t any $\pi_0$

<table>
<thead>
<tr>
<th></th>
<th>$b$</th>
<th>$a$</th>
<th>$d$</th>
<th>$c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_3$</td>
<td>$-$</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$S_2$</td>
<td>1</td>
<td>$-$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$S_4$</td>
<td>0</td>
<td>0</td>
<td>$-$</td>
<td>0</td>
</tr>
<tr>
<td>$S_1$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>$-$</td>
</tr>
</tbody>
</table>

$$V_1 \ V_2 \ V_3 \ V_4$$

$$(V_1, V_2, V_3) = (2, 1, 1)$$

or

$$(S_1, S_2, S_3) = (3, 1, 0)$$

$$d(\pi, \pi_0) = 4$$
The code of a permutation

Example $\pi = [c \ a \ b \ d], \ \pi_0 = [b \ a \ d \ c]$

<table>
<thead>
<tr>
<th></th>
<th>$a$</th>
<th>$b$</th>
<th>$c$</th>
<th>$d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_2$</td>
<td>$-$</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$S_3$</td>
<td>0</td>
<td>$-$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$S_1$</td>
<td>1</td>
<td>1</td>
<td>$-$</td>
<td>1</td>
</tr>
<tr>
<td>$S_4$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$-$</td>
</tr>
</tbody>
</table>

$V_1 \ V_2 \ V_3 \ V_4$

code

$$(V_1, V_2, V_3) = (1, 1, 0)$$
or
$$(S_1, S_2, S_3) = (2, 0, 0)$$

d$(\pi, \text{id}) = 2$

Codes are defined w.r.t any $\pi_0$

<table>
<thead>
<tr>
<th></th>
<th>$b$</th>
<th>$a$</th>
<th>$d$</th>
<th>$c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_3$</td>
<td>$-$</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$S_2$</td>
<td>1</td>
<td>$-$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$S_4$</td>
<td>0</td>
<td>0</td>
<td>$-$</td>
<td>0</td>
</tr>
<tr>
<td>$S_1$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>$-$</td>
</tr>
</tbody>
</table>

$V_1 \ V_2 \ V_3 \ V_4$

code $V_j(\pi|\pi_0), \ S_j(\pi|\pi_0)$

$$(V_1, V_2, V_3) = (2, 1, 1)$$
or
$$(S_1, S_2, S_3) = (3, 1, 0)$$

d$(\pi, \pi_0) = 4$

- For any $\pi_0$, the code $(V_1(\pi|\pi_0) \ldots V_{n-1}(\pi|\pi_0))$ defines $\pi$ uniquely
The Generalized Mallows (GM) Model [Fligner, Verducci 86]

Generalized Mallows(GM) model

\[ P_{\pi_0, \theta}(\pi) = \frac{1}{Z(\theta)} \prod_{j=1}^{n-1} \exp \left[ -\theta_j V_j(\pi | \pi_0) \right] \quad \text{with} \quad Z(\theta) = \prod_{j=1}^{n-1} Z_j(\theta_j) \]
The Generalized Mallows (GM) Model [Fligner, Verducci 86]

Generalized Mallows (GM) model

\[ P_{\pi_0, \theta}(\pi) = \frac{1}{Z(\theta)} \prod_{j=1}^{n-1} \exp[-\theta_j V_j(\pi | \pi_0)] \]  

with  
\[ Z(\theta) = \prod_{j=1}^{n-1} Z_j(\theta_j) \]

- \( \pi_0 \) is the central permutation
  - \( \pi_0 \) mode of \( P_{\pi_0, \theta} \), unique if \( \theta > 0 \)
- \( \theta_j \geq 0 \) are dispersion parameters
  - for \( \theta = 0 \), \( P_{\pi_0, 0} \) is uniform over \( S_n \)
- \( Z_j(\theta_j) \) is tractable
The Generalized Mallows (GM) Model [Fligner, Verducci 86]

Generalized Mallows(GM) model

\[
P_{\pi_0, \theta}(\pi) = \frac{1}{Z(\theta)} \prod_{j=1}^{n-1} \exp[-\theta_j V_j(\pi | \pi_0)] \quad \text{with} \quad Z(\theta) = \prod_{j=1}^{n-1} Z_j(\theta_j)
\]

- \(\pi_0\) is the central permutation
  - \(\pi_0\) mode of \(P_{\pi_0, \theta}\), unique if \(\theta > 0\)
- \(\theta_j \geq 0\) are dispersion parameters
  - for \(\theta = 0\), \(P_{\pi_0, 0}\) is uniform over \(S_n\)
- \(Z_j(\theta_j)\) is tractable

Cost interpretation of the GM model

- \(GM^Y\): Cost = \(\sum_j \theta_j V_j\)
  - pay price \(\theta_j\) for every inversion w.r.t item \(j\)
- Assume stepwise construction of \(\pi\): \(\theta_j\) represents importance of step \(j\)
Outline

Permutations and their representations
- Statistical models for permutations and the dependence of ranks
- Codes, inversion distance and the precedence matrix
- Mallows models over permutations

Complete rankings and Maximum Likelihood estimation
- GM as exponential family

Top-t rankings, infinite permutations, and Bayesian estimation
- Top-t rankings and infinite permutations
- Conjugate prior, Dirichlet process mixtures

Recursive inversion models and finding common structure in preferences

[Signed permutations and the reversal median problem]
### ML Estimation of $\pi_0$: costs and main results

<table>
<thead>
<tr>
<th>Model</th>
<th>Data</th>
<th>Log-likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mallows</td>
<td>complete rankings</td>
<td>$\sum_{j=1}^{n-1} \tilde{V}_j(\pi_0)$</td>
</tr>
<tr>
<td>$GM^V$</td>
<td></td>
<td>$\sum_{j=1}^{n-1} \left[ \theta_j \tilde{V}_j(\pi_0) + \ln Z_j(\theta_j) \right]$</td>
</tr>
</tbody>
</table>

$\tilde{V}_j(\pi_0) = \frac{1}{\bar{V}_j(\pi_0)} \sum_{\pi \in \text{data}} V_j(\pi | \pi_0)$

$\pi_0^{ML}$ estimated exactly by B&B search.

B&B = Branch-and-Bound
ML Estimation of $\pi_0$: costs and main results

<table>
<thead>
<tr>
<th>Model</th>
<th>Data</th>
<th>Log-likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mallows</td>
<td>complete rankings</td>
<td>$\sum_{j=1}^{n-1} \bar{V}_j(\pi_0)$</td>
</tr>
<tr>
<td>$GM^V$</td>
<td>$\sum_{j=1}^{n-1} [\theta_j \bar{V}_j(\pi_0) + \ln Z_j(\theta_j)]$</td>
<td>[M&amp;al07] $\pi_0^{ML}$ estimated exactly by B&amp;B search. B&amp;B=Branch-and-Bound</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\bar{V}<em>j(\pi_0) = \frac{1}{N} \sum</em>{\pi \in \text{data}} V_j(\pi</td>
</tr>
<tr>
<td>$GM^S$</td>
<td>complete rankings</td>
<td>$\sum_{j=1}^{n-1} [\theta_j \bar{S}_j(\pi_0) + \ln Z_j(\theta_j)]$</td>
</tr>
<tr>
<td></td>
<td>top-t rankings</td>
<td>$\sum_{j=1}^{t} [\theta_j \bar{S}_j(\pi_0) + \ln Z_j(\theta_j)]$</td>
</tr>
<tr>
<td></td>
<td>top-t rankings, $n = \infty$</td>
<td>$\bar{S}<em>j(\pi_0) = \frac{1}{N} \sum</em>{\pi \in \text{data}} s_j(\pi</td>
</tr>
</tbody>
</table>
Sufficient statistics [M&al07]

- Define \( Q \equiv Q(\pi_{1:N}) = \frac{1}{N} \sum_{i=1}^{N} Q(\pi_i) \)
- Sufficient statistics are sum of preference matrices for data

\[
Q(\pi) = 
\begin{pmatrix}
- & 0 & 1 & 0 \\
1 & - & 1 & 0 \\
0 & 0 & - & 0 \\
1 & 1 & 1 & - \\
\end{pmatrix}
\]

\( Q \) for large samples from Mallows models

\( \theta = 1 \)  
\( \theta = 0.3 \)  
\( \theta = 0.03 \)
Sufficient statistics [M&al07]

- Define \( Q \equiv Q(\pi_{1:N}) = \frac{1}{N} \sum_{i=1}^{N} Q(\pi_i) \)
- Sufficient statistics are sum of preference matrices for data

\[
Q(\pi) = \begin{bmatrix}
- & 0 & 1 & 0 \\
1 & - & 1 & 0 \\
0 & 0 & - & 0 \\
1 & 1 & 1 & -
\end{bmatrix}
\]

\( Q \) for large samples from Mallows models

\[
\theta = 1 \\
\theta = 0.3 \\
\theta = 0.03
\]

Consensus ranking

\[
= \arg\min_{\pi_0} L(\Pi^T_0 Q \Pi_0) = \arg\min_{\pi_0} L_{\pi_0}(Q)
\]

= argmin lower triangle of \( Q \) over all row and column permutations \( \pi_0 \)
Search Algorithm Idea

Wanted: \( \arg\min_{\pi_0} L(\Pi_0^T Q \Pi_0) = \arg\min_{\pi_0} L_{\pi_0}(Q) = \arg\min \) lower triangle of \( Q \) over all row and column permutations
Search Algorithm Idea

Wanted: $\arg\min_{\pi_0} L(\Pi_0^T Q \Pi_0) = \arg\min_{\pi_0} L_{\pi_0}(Q) = \arg\min$ lower triangle of $Q$ over all row and column permutations
Search Algorithm Idea

Wanted: \( \arg\min_{\pi_0} L(\Pi_0^T Q \Pi_0) = \arg\min_{\pi_0} L_{\pi_0}(Q) = \arg\min \) lower triangle of \( Q \) over all row and column permutations.
Search Algorithm Idea

Wanted: \( \arg\min_{\pi_0} L(\Pi_0^T Q \Pi_0) = \arg\min_{\pi_0} L_{\pi_0}(Q) = \arg\min \) lower triangle of \( Q \) over all row and column permutations
Search Algorithm Idea

Wanted: $\arg\min_{\pi_0} L(\Pi_0^T Q \Pi_0) = \arg\min_{\pi_0} L_{\pi_0}(Q) = \arg\min$ lower triangle of $Q$ over all row and column permutations
Parameter spaces and sufficient statistics spaces

Parameters

- GM model is \textit{curved} exponential family
  - $n - 1$ discrete and $n - 1$ continuous parameters
- Full exponential family = \textit{inversions (Bradley-Terry) model}

\[
P(\pi) \propto \exp \left( - \sum_{i < j} \alpha_{ij} Q_{ij}(\pi) \right)
\]

- not tractable [Diaconis87]
Parameter spaces and sufficient statistics spaces

Parameters

- GM model is *curved* exponential family
  - $n - 1$ discrete and $n - 1$ continuous parameters
- Full exponential family = *inversions* (Bradley-Terry) model

\[
P(\pi) \propto \exp \left( - \sum_{i<j} \alpha_{ij} Q_{ij}(\pi) \right)
\]

- not tractable [Diaconis87]

Sufficient statistics

- space of “skew-symmetric” matrices with $[0, 1]$ elements
  \[
  \mathcal{A} = \{ Q \mid Q_{ik} + Q_{ki} = 1, Q_{ik} > 0 \}
  \]
Parameter spaces and sufficient statistics spaces

Parameters

- GM model is *curved* exponential family
  - $n - 1$ discrete and $n - 1$ continuous parameters
- Full exponential family = *inversions* (Bradley-Terry) model

$$P(\pi) \propto \exp \left( - \sum_{i<j} \alpha_{ij} Q_{ij}(\pi) \right)$$

- not tractable [Diaconis87]

Sufficient statistics

- space of “skew-symmetric” matrices with $[0, 1]$ elements
  $$A = \{ Q \mid Q_{ik} + Q_{ki} = 1, Q_{ik} > 0 \}$$
- space of sufficient statistics = *linear orderings polytope* (difficult to describe [SturmfelsWelker11, Grötschel85])
  $$Q = \{ Q = \frac{1}{N} \sum_{1=1}^{N} Q(\pi_i) \}$$
Parameter spaces and sufficient statistics spaces

Parameters

- GM model is \textit{curved} exponential family
  - \( n - 1 \) discrete and \( n - 1 \) continuous parameters
- Full exponential family = inversions (Bradley-Terry) model

\[
P(\pi) \propto \exp \left( - \sum_{i<j} \alpha_{ij} Q_{ij}(\pi) \right)
\]

- not tractable [Diaconis87]

Sufficient statistics

- space of “skew-symmetric” matrices with \([0, 1]\) elements
  \[
  \mathcal{A} = \{ Q \mid Q_{ik} + Q_{ki} = 1, \ Q_{ik} > 0 \}
  \]
- space of sufficient statistics = linear orderings polytope (difficult to describe [SturmfelsWelker11, Grötschel85])
  \[
  Q = \{ Q = \frac{1}{N} \sum_{i=1}^{N} Q(\pi_i) \}
  \]
- space of means of GM model \( \mathcal{M} = \{ E_{\pi_{0}, \theta}[Q] \} \)
  - not a polytope
  - characterized algorithmically by [Mallows 57] for Mallows, [M&al07] for GMM
Consistency and rates of ML estimates

- \( \frac{Q_{ij}}{N} \rightarrow P[ \text{item } i <_{\pi_0} \text{item } j ] \) as \( N \rightarrow \infty \) [FlignerVerducci86]
- Therefore
  - for any \( \pi_0 \) fixed, \( \hat{\theta}^{ML} \) is consistent [FlignerVerducci86]
  - the discrete parameter \( \pi_0^{ML} \) consistent when \( \theta_j \) non-increasing [FlignerVerducci86, M–in prep]
  - is it “unbiased”?

**Theorem 1** [M–in prep] For any \( N \) finite

\[
E[\theta^{ML}] > \theta \quad \text{Bias!}
\]

and the order of magnitude of \( \theta^{ML} - \theta \) is \( \frac{1}{\sqrt{N}} \) w.h.p.
The Bias of $\theta^{ML}$

$\theta_j$ estimates for $j = 1, 8$
Outline

Permutations and their representations
  Statistical models for permutations and the dependence of ranks
  Codes, inversion distance and the precedence matrix
  Mallows models over permutations

Complete rankings and Maximum Likelihood estimation
  GM as exponential family

Top-t rankings, infinite permutations, and Bayesian estimation
  Top-t rankings and infinite permutations
  Conjugate prior, Dirichlet process mixtures

Recursive inversion models and finding common structure in preferences

[Signed permutations and the reversal median problem]
Top-t rankings and very many items

2000 Presidential Elections Ireland, $n = 5$, $N = 1100$
Roch Scal McAl Bano Nall
Scal McAl Nall Bano Roch
Roch McAl

College programs $n = 533$, $N = 53737$, $t = 10$
DC116 DC114 DC111 DC148 DB512 DN021 LM054 WD048 LM020 LM050
WD028
DN008 TR071 DN012 DN052
FT491 FT353 FT471 FT541 FT402 FT404 TR004 FT351 FT110 FT352

Google search: [Columbia Statistics](http://stat.columbia.edu)
stat.columbia.edu
gsas.columbia.edu
[www.gocolumbialions.com/SportSelect.db...](http://www.gocolumbialions.com/SportSelect.db...)
[www.grad-schools.usnews.rankingsandreviews.com](http://www.grad-schools.usnews.rankingsandreviews.com)

... searches in data bases of biological sequences (by e.g. Blast, Sequest, etc)
... open-choice polling, "grassroots elections", college program applications
Models for Infinite permutations

- **Domain** is countable, i.e. $n \rightarrow \infty$
- **Observe** the top $t$ ranks of an infinite permutation
Models for Infinite permutations

- **Domain** is countable, i.e. $n \rightarrow \infty$
- **Observe** the top $t$ ranks of an infinite permutation

**College programs** $n = 533, N = 53737, t = 10$

DC116 DC114 DC111 DC148 DB512 DN021 LM054 WD048 LM020 LM050 WD028
DN008 TR071 DN012 DN052 FT491 FT353 FT471 FT541 FT402 FT404 TR004 FT351 FT110 FT352

- **Mathematically more natural**
  - for large $n$, models should not depend on $n$
  - models can be simpler, more elegant than for finite $n$
Top-t rankings: $GM^S$, $GM^V$ are not equivalent

$p_0 = [a \ b \ c \ d]$
$p = [c \ a]$

\[
\begin{align*}
p(1) &= c & S_1 &= 2 \\
p(2) &= a & S_2 &= 0 \\
p(3) &= ? & S_3 &= ?
\end{align*}
\]

\[
P_{p_0, \theta}(p) = \prod_{j=1}^{t} e^{-\theta_j S_j}
\]

$sufficient statistics$

\[
P_{p_0, \theta}(p) = \prod_{j=1}^{n-1} \left\{ \begin{array}{l}
e^{-\theta_j V_j}, \, p_0(j) \in \pi \\
P_\theta(V_j \geq v_j), \, p_0(j) \notin \pi
\end{array} \right.
\]

$no sufficient statistics$

Example: $p = [c \ a]$

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_2$</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$S_1$</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>$V_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$V_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$V_3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$V_4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The Infinite (Generalized) Mallows model (IGM)

\[
P_{\pi_0, \theta}(\pi) = \exp \left[ - \sum_{j=1}^{t} (\theta_j S_j(\pi | \pi_0) - \ln Z(\theta_j)) \right]
\]

- \( \pi \) is observed top-\( t \) ranking
- \( \pi_0 \) is central permutation of \( \{1, 2, 3, \ldots\} \)
  - discrete infinite “location” parameter
- \( \theta_{1:t} > 0 \) dispersion parameter
  - dimension equal to \( t \)
- all \( S_j \) have same range \( \{0, 1, 2, \ldots\} \)
- Normalization constant \( Z(\theta_j) = 1/(1 - e^{-\theta_j}) \)
- \( P_{\pi_0, \theta}(\pi) \) is well defined marginal over the coset defined by \( \pi \)
Sufficient statistics for top-t permutations \cite{miao10}

Sufficient statistics are $t \times n$ precedence matrices $R_1, \ldots, R_t$

\begin{array}{c|cccc}
    \pi(j) & \_ & \_ & \_ & \_ \\
    \_ & 0 & 1 & \_ & 1 \\
\end{array}

\begin{align*}
S_j(\pi | \pi_0) &= L_{\pi_0}(R_j(\pi)) \quad \text{\cite{miao10}}
\end{align*}
Infinite GMM: ML estimation

**Theorem [M,Bao 10]**

- **Sufficient statistics**
  - \( n \): \# distinct items observed in data
  - \( N_j \): \# total permutations with length \( \geq j \)
  - \( R^{(j)} = [R_{kl}^{(j)}] \): frequency of rank \( k = j \), rank \( l > j \) in data

- log-likelihood \( l(\pi_0, \theta) = \text{Sum}( \text{Lower triangle} (\sum_j \theta_j R^{(j)}) \text{ permuted by } \pi_0 ) + \text{constant} \)
Infinite GMM: ML estimation

**Theorem [M,Bao 10]**

- Sufficient statistics
  - $n$  
  # distinct items observed in data
  - $N_j$  
  # total permutations with length $\geq j$
  - $R^{(j)} = [R_{kl}^{(j)}]$  
  frequency of rank($k$) = $j$, rank($l$) $> j$ in data

- log-likelihood $l(\pi_0, \bar{\theta}) = \text{Sum( Lower triangle}(\sum_j \theta_j R^{(j)}) \text{ permuted by } \pi_0) + \text{constant}$

- given $\pi_0$,

$$\theta_j^{ML} = \log \left( 1 + \frac{N_j}{L_{\pi_0}(R^{(j)})} \right)$$
Infinite GMM: ML estimation

**Theorem** [M,Bao 10]

- Sufficient statistics
  
  \[
  \begin{align*}
  n & \quad \text{# distinct items observed in data} \\
  N_j & \quad \text{# total permutations with length } \geq j \\
  R^{(j)} = [R_{kl}^{(j)}] & \quad \text{frequency of rank}(k) = j, \text{rank}(l) > j \text{ in data}
  \end{align*}
  \]

- log-likelihood
  
  \[
  l(\pi_0, \bar{\theta}) = \text{Sum} (\text{Lower triangle}(\sum_j \theta_j R^{(j)}) \text{ permuted by } \pi_0) + \text{constant}
  \]

- given \(\pi_0\),

  \[
  \theta_j^{ML} = \log \left(1 + \frac{N_j}{L \pi_0(R^{(j)})}\right)
  \]

- given \(\theta_{1:t}\), \(\pi_0^{ML}\) can be found exactly by a B&B algorithm searching on matrix \(\sum_j \theta_j R^{(j)}\).
ML Estimation: Remarks

- sufficient statistics $R_{1:t}$ finite for finite sample size $N$ but don’t compress the data
- data determine only a finite set of parameters
  - $\pi_0$ restricted to the observed items
  - $\theta$ restricted to the observed ranks

\[ N = 200 \]

\[ N = 2000 \]

rank $j$
GM are exponential family models

\[ GM^V \text{ for complete rankings} \]
\[ GM^S \text{ for top-t rankings, } n \text{ finite or } \infty \]

- have finite sufficient statistics
- are exponential family models in \( \pi_0, \tilde{\theta} \)
- have conjugate priors

Hyperparameters

- \( N_0 > 0 \) equivalent sample size
- \( R_j^0 \in \mathbb{R}^{n \times n} \) equivalent sufficient statistics
  - informative prior for \( \pi_0, \tilde{\theta} \)
Bayesian Inference: What operations are tractable?

Conjugate prior

\[ P_0(\pi_0, \theta) \propto \exp \left[ \sum_j (\theta_j (N_0 r_j) + N_0 \ln Z(\theta_j)) \right] \]

Posterior

\[ P(\pi_0, \theta) \propto \exp \left[ \sum_j (\theta_j (N_0 r_j + NL\pi_0(R_j)) + (N_0 + N) \ln Z(\theta_j)) \right] \]

- computing unnormalized prior, posterior \( \checkmark \)
- normalization constant, model averaging under prior, posterior \( \times \)
Bayesian Inference: What operations are tractable?

**Conjugate prior**

\[ P_0(\pi_0, \tilde{\theta}) \propto \exp \left[ \sum_j (\theta_j (N_0 r_j) + N_0 \ln Z(\theta_j)) \right] \]

**Posterior**

\[ P(\pi_0, \tilde{\theta}) \propto \exp \left[ \sum_j (\theta_j (N_0 r_j + NL\pi_0 (R_j)) + (N_0 + N) \ln Z(\theta_j)) \right] \]

- computing unnormalized prior, posterior \(\checkmark\)
- normalization constant, model averaging under prior, posterior \(\times\)

- “Toolbox” of tractable Bayesian operations \([M, Chen 10, 16]\)
  - integrating out \(\tilde{\theta}\) parameters
  - sampling \(\tilde{\theta} | \pi_0, \pi_0 | \tilde{\theta}\) from posterior
  - closed form posterior for \(N = 1\)
Bayesian Inference: What operations are tractable?

Conjugate prior

\[ P_0(\pi_0, \bar{\theta}) \propto \exp \left[ \sum_j (\theta_j (N_0 r_j) + N_0 \ln Z(\theta_j)) \right] \]

Posterior

\[ P(\pi_0, \bar{\theta}) \propto \exp \left[ \sum_j (\theta_j (N_0 r_j + NL_0 \pi_0 (R_j)) + (N_0 + N) \ln Z(\theta_j)) \right] \]

- computing unnormalized prior, posterior ✓
- normalization constant, model averaging under prior, posterior ❌

“Toolbox” of tractable Bayesian operations [M,Chen 10,16]

Lemma 1 [M,Bao 10] Posterior of \( \pi_0 \) and \( \theta_j | \pi_0 \)

\[ P(e^{-\theta_j} | \pi_0, N_0, r, \pi_{1:N}) = \text{Beta}(e^{-\theta_j} ; N_0 r_j + S_j, N_0 + N + 1) \]

\[ P(\pi_0 | N_0, r, \pi_{1:N}) \propto \prod_{j=1}^{t} \text{Beta}(N_0 r_j + S_j, N_0 + N + 1) \]
Bayesian Inference: What operations are tractable?

Conjugate prior

\[ P_0(\pi_0, \theta) \propto \exp \left[ \sum_j (\theta_j (N_0 r_j) + N_0 \ln Z(\theta_j)) \right] \]

Posterior

\[ P(\pi_0, \theta) \propto \exp \left[ \sum_j (\theta_j (N_0 r_j + NL_0 (R_j)) + (N_0 + N) \ln Z(\theta_j)) \right] \]

- computing unnormalized prior, posterior \( \checkmark \)
- normalization constant, model averaging under prior, posterior \( \times \)

“Toolbox” of tractable Bayesian operations \([M, Chen \ 10, 16]\)

Lemma 2\([M, Chen \ 10, 16]\) Bayesian averaging over \( \theta \)

\[ P(\pi | \pi_0, N_0, r, \pi_1:N) = \prod_{j=0}^{t} \frac{Beta(S_j(\pi | \pi_0) + N_0 r_j + S_j, N_0 + N + 2)}{Beta(N_0 r_j + S_j, N_0 + N + 1)} \]
Bayesian Inference: What operations are tractable?

Conjugate prior

\[
P_0(\pi_0, \theta) \propto \exp \left[ \sum_j (\theta_j(N_0 r_j) + N_0 \ln Z(\theta_j)) \right]
\]

Posterior

\[
P(\pi_0, \theta) \propto \exp \left[ \sum_j (\theta_j(N_0 r_j + NL_p(R_j)) + (N_0 + N) \ln Z(\theta_j)) \right]
\]

- computing unnormalized prior, posterior \(\checkmark\)
- normalization constant, model averaging under prior, posterior \(\times\)

- "Toolbox" of tractable Bayesian operations [M, Chen 10, 16]

**Lemma 3** [M, Chen 10, 16] Normalized posterior for \(N = 1\)

\[
Z_1 = \frac{(n - t)!}{n!}
\]

- for \(N = 1\) sample, the posterior dispersion does not depend on the sample
- allows assigning to/sampling from the singleton clusters
Bayesian Inference: What operations are tractable?

**Conjugate prior**

$$P_0(\pi_0, \hat{\theta}) \propto \exp \left[ \sum_j (\theta_j (N_0 r_j) + N_0 \ln Z(\theta_j)) \right]$$

**Posterior**

$$P(\pi_0, \hat{\theta}) \propto \exp \left[ \sum_j (\theta_j (N_0 r_j + NL_0(R_j)) + (N_0 + N) \ln Z(\theta_j)) \right]$$

- computing unnormalized prior, posterior ✓
- normalization constant, model averaging under prior, posterior ❌

- "Toolbox" of tractable Bayesian operations [M, Chen 10,16]

**Lemma 4** [M, Chen 10, 16] Exact sampling of $\pi_0 | \hat{\theta}$ from the posterior possible by stagewise sampling.

$$P(\pi_0 | \hat{\theta}, N_0, r, \pi_1:N) \propto e^{-\sum_j \theta_j L_{\pi_0}(R_j)}$$
Bayesian Inference: What operations are tractable?

Conjugate prior

$$P_0(\pi_0, \theta) \propto \exp \left[ \sum_j (\theta_j (N_0 r_j) + N_0 \ln Z(\theta_j)) \right]$$

Posterior

$$P(\pi_0, \theta) \propto \exp \left[ \sum_j (\theta_j (N_0 r_j + NL\pi_0(R_j)) + (N_0 + N) \ln Z(\theta_j)) \right]$$

- computing unnormalized prior, posterior √
- normalization constant, model averaging under prior, posterior X

“Toolbox” of tractable Bayesian operations [M, Chen 10, 16]

Lemma 5 [M, Chen 10, 16]

$$P(\pi | \pi_0|_{\text{obs}}, \pi_{1:N}) = \prod_{j: \pi(j) \in \text{obs}} \frac{\Gamma(S_j + N_0 r_j + S_j + N_0 + N + 2)}{\Gamma(N_0 r_j + S_j + N_0 + N + 1)} \prod_{j: \pi(j) \notin \text{obs}} \frac{\Gamma(t_j + N_0 r_j + S_j + N_0 + N)}{\Gamma(N_0 + N + 1)}$$
Bayesian Inference: What operations are tractable?

Conjugate prior

\[ P_0(\pi_0, \theta) \propto \exp \left[ \sum_j (\theta_j(N_0 r_j) + N_0 \ln Z(\theta_j)) \right] \]

Posterior

\[ P(\pi_0, \theta) \propto \exp \left[ \sum_j (\theta_j(N_0 r_j + NL\pi_0(R_j)) + (N_0 + N) \ln Z(\theta_j)) \right] \]

- computing unnormalized prior, posterior \( \checkmark \)
- normalization constant, model averaging under prior, posterior \( \times \)

- “Toolbox” of tractable Bayesian operations [M,Chen 10,16]
  - exploited properties of sufficient statistics
  - power series manipulation
  - careful programming
  - approximating finite \( n \) with \( n = \infty \) speeds up computation
Clustering with Dirichlet mixtures via MCMC

General DPMM estimation algorithm \([\text{Escobar, West}95, \text{Neal}03]\)

**MCMC estimation for Dirichlet mixture**

**Input**  \(\alpha, g_0, \beta, \{f\}, \mathcal{D}\)

**State**  cluster assignments \(c(i), i = 1 : n\),

parameters \(\theta_k\) for all distinct \(k\)

**Iterate**  
1. for \(i = 1 : n\) (reassign data to clusters)
   1.1 if \(n_{c(i)} = 1\) delete this cluster and its \(\theta_{c(i)}\)
   1.2 resample \(c(i)\) by
      \[
      c(i) = \begin{cases} 
      \text{existing } k & \text{w.p } \propto \frac{n_k^{k-1}}{n-1+\alpha} f(x_i, \theta_k) \\
      \text{new cluster} & \text{w.p } \frac{\alpha}{n-1+\alpha} \int f(x_i, \theta) g_0(\theta) d\theta 
      \end{cases}
      \]
   1.3 if \(c(i)\) is new label, sample a new \(\theta_{c(i)}\) from \(g_0\)
2. (resample cluster parameters)
   for \(k \in \{c(1 : n)\}\)
   2.1 sample \(\theta_k\) from posterior \(g_k(\theta) \propto g_0(\theta, \beta) \prod_{i \in c_k} f(x_i, \theta)\)

\(g_k\) can be computed in closed form if \(g_0\) is conjugate prior

**Output**  a state with high posterior
College program admissions, Ireland

\[ n = 533 \text{ programs}, \ N = 53737 \text{ candidates}, \ t = 10 \text{ options} \]

DC116 DC114 DC111 DC148 DB512 DN021 LM054 WD048 LM020 LM050
WD028
DN008 TR071 DN012 DN052
FT491 FT353 FT471 FT541 FT402 FT404 TR004 FT351 FT110 FT352

- Data = all candidates’ rankings for college programs in 2000
  from [GormleyMurphy03] (they used EM for Mixture of Plackett-Luce models)
- [M, Chen 10, Ali, Murphy, M, Chen 10] used DPMM (parameters adjusted to get approx 20 clusters)
College program rankings: are there clusters?

- 33 clusters cover 99% of the data
- $\tilde{\theta}_c$ parameters large – cluster are concentrated
- number of significant ranks in $\sigma_c, \theta_c$ vary by cluster
College program rankings: are the clusters meaningful?

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Size</th>
<th>Description</th>
<th>Male (%)</th>
<th>Points avg(std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4536</td>
<td>CS &amp; Engineering</td>
<td>77.2</td>
<td>369 (41)</td>
</tr>
<tr>
<td>2</td>
<td>4340</td>
<td>Applied Business</td>
<td>48.5</td>
<td>366 (40)</td>
</tr>
<tr>
<td>3</td>
<td>4077</td>
<td>Arts &amp; Social Science</td>
<td>13.1</td>
<td>384 (42)</td>
</tr>
<tr>
<td>4</td>
<td>3898</td>
<td>Engineering (Ex-Dublin)</td>
<td>85.2</td>
<td>374 (39)</td>
</tr>
<tr>
<td>5</td>
<td>3814</td>
<td>Business (Ex-Dublin)</td>
<td>41.8</td>
<td>394 (32)</td>
</tr>
<tr>
<td>6</td>
<td>3106</td>
<td>Cork Based</td>
<td>48.9</td>
<td>397 (33)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>33</td>
<td>9</td>
<td>Teaching (Home Economics)</td>
<td>0.0</td>
<td>417 (4)</td>
</tr>
</tbody>
</table>

- Cluster differentiate by **subject area**
- ... also by **geography**
- ... show gender difference in preferences
College program rankings: the “prestige” question

- Question: are choices motivated by “prestige” (i.e., high entrance points scores)?
- If yes, then PR should be decreasing along the rankings points overall (quantiles).
- Unclustered data: PR decreases monotonically with rankings.
- Clustered data: PR not always monotonic
  - Simpson’s paradox!

![Graph showing PR decreases with rankings](image-url)
Outline

Permutations and their representations
  Statistical models for permutations and the dependence of ranks
  Codes, inversion distance and the precedence matrix
  Mallows models over permutations

Complete rankings and Maximum Likelihood estimation
  GM as exponential family

Top-t rankings, infinite permutations, and Bayesian estimation
  Top-t rankings and infinite permutations
  Conjugate prior, Dirichlet process mixtures

Recursive inversion models and finding common structure in preferences

[Signed permutations and the reversal median problem]
Recursive Inversion Models (RIM)

[Meek, M 14]

\[
\tau = \text{tree structure} \\
\pi_0(\tau) = \text{induced central ranking} \\
\theta_{1:n-1} = \text{parameters at nodes}
\]
Recursive Inversion Models (RIM)

[Meek, M 14]

\[ \tau = \text{tree structure} \]
\[ \pi_0(\tau) = \text{induced central ranking} \]
\[ \theta_{1:n-1} = \text{parameters at nodes} \]

Inversions are penalized by \( \theta_i \) parameters

Example: \( \bar{\theta} = (0.1, 1.2, 0.4) \)

\[ \text{Cost}(a|b|c|d) = 0 \]
Recursive Inversion Models (RIM)

[Meek, M 14]

\[ \tau = \text{tree structure} \]
\[ \pi_0(\tau) = \text{induced central ranking} \]
\[ \theta_{1:n-1} = \text{parameters at nodes} \]

Inversions are penalized by $\theta_i$ parameters

Example: $\bar{\theta} = (0.1, 1.2, 0.4)$

\[
\begin{align*}
\text{Cost}(a|b|c|d) &= 0 \\
\text{Cost}(b|a|c|d) &= 1.2
\end{align*}
\]
Recursive Inversion Models (RIM)

[Meek, M 14]

\[ \tau = \text{tree structure} \]
\[ \pi_0(\tau) = \text{induced central ranking} \]
\[ \theta_{1:n-1} = \text{parameters at nodes} \]

Inversions are penalized by \( \theta_i \) parameters

Example: \( \bar{\theta} = (0.1, 1.2, 0.4) \)

\[
\begin{align*}
\text{Cost}(a|b|c|d) & = 0 \\
\text{Cost}(b|a|c|d) & = 1.2 \\
\text{Cost}(c|b|a|d) & = 1.2 + 2 \times 0.1
\end{align*}
\]
Recursive Inversion Models (RIM)

[Meek, M 14]

$\tau = \text{tree structure}$

$\pi_0(\tau) = \text{induced central ranking}$

$\theta_{1:n-1} = \text{parameters at nodes}$

Inversions are penalized by $\theta_i$ parameters

Example: $\bar{\theta} = (0.1, 1.2, 0.4)$

\[
\begin{align*}
\text{Cost}(a|b|c|d) & = 0 \\
\text{Cost}(b|a|c|d) & = 1.2 \\
\text{Cost}(c|b|a|d) & = 1.2 + 2 \times 0.1
\end{align*}
\]
Recursive Inversion Models (RIM)

[Meek, M 14]

\(\tau = \) tree structure
\(\pi_0(\tau) = \) induced central ranking
\(\theta_{1:n-1} = \) parameters at nodes

Inversions are penalized by \(\theta_i\) parameters

Example: \(\bar{\theta} = (0.1, 1.2, 0.4)\)

\[
\begin{align*}
\text{Cost}(a|b|c|d) &= 0 \\
\text{Cost}(b|a|c|d) &= 1.2 \\
\text{Cost}(c|b|a|d) &= 1.2 + 2 \times 0.1
\end{align*}
\]

RIM distribution \(P_{\tau,\bar{\theta}}\)

Let \(v_i = \) number of inversions of \(\pi\) at node \(i\)

\[
P_{\tau,\bar{\theta}}(\pi) \propto \prod_{i \in \text{nodes}} \exp(-\theta_i v_i)
\]

\[
\begin{align*}
P(a|b|c|d) &\propto e^0 \\
P(b|a|c|d) &\propto e^{-1.2} \\
P(c|b|a|d) &\propto e^{-1.2 - 2 \times 0.1}
\end{align*}
\]
Recursive Inversion Models (RIM)

[Meek, M 14]

\[ \tau = \text{tree structure} \]
\[ \pi_0(\tau) = \text{induced central ranking} \]
\[ \theta_{1:n-1} = \text{parameters at nodes} \]

Inversions are penalized by \( \theta_i \) parameters

Example: \( \vec{\theta} = (0.1, 1.2, 0.4) \)

\[
\begin{align*}
\text{Cost}(a|b|c|d) &= 0 \\
\text{Cost}(b|a|c|d) &= 1.2 \\
\text{Cost}(c|b|a|d) &= 1.2 + 2 \times 0.1
\end{align*}
\]

RIM distribution \( P_{\tau, \vec{\theta}} \)

Let \( v_i = \text{number of inversions of } \pi \text{ at node } i \)

\[
P_{\tau, \vec{\theta}}(\pi) \propto \prod_{i \in \text{nodes}} \exp(-\theta_i v_i)
\]

Normalization constant

\[
Z(\tau, \theta) = \prod_{i \in \text{nodes}} G(L_i, R_i, \exp(-\theta_i))
\]

with \( G(L, R, q) = \frac{(q)_L R}{(q)_L (q)_L} \), \( (q)_n = \prod_{i=1}^{n} (1 - q^i) \).

Structure \( \tau \) known as Riffle Independence model [Huang, Guestrin 12]
The RIM is a general flexible model

- any tree structure
- any parameters (but $\theta_j \geq 0$ suffices)
- includes the Mallows and Generalized Mallows models
Max Likelihood Estimation for RIM

[M,Meek 14]

- **Problem** Given permutations $\pi_1, \ldots, \pi_N$, infer $\tau, \theta$
Max Likelihood Estimation for RIM

[M, Meek 14]

- **Problem** Given permutations $\pi_1, \ldots, \pi_N$, infer $\tau, \theta$
- **Identifiability and estimation of $\theta$**
  - reorder to obtain canonical representation, with $\theta_i \geq 0$ for all $i \in \text{nodes}$
  - given $\tau, \theta_i$ can be estimated by convex univariate minimization

```
0.1

1.2
apple banana

0.4
cherry durian
```
Max Likelihood Estimation for RIM

[M, Meek 14]

- **Problem**
  
  Given permutations $\pi_1, \ldots, \pi_N$, infer $\tau, \theta$

- **Identifiability and estimation of $\theta$**
  
  - reorder to obtain cannonical representation, with $\theta_i \geq 0$ for all $i \in \text{nodes}$
  
  - given $\tau$, $\theta_i$ can be estimated by convex univariate minimization

- **Identifiability of $\tau$**

  **Theorem** [M, Meek 14] A model $\tau, \theta$ is identifiable iff

  1. $\theta_i > 0$ for all $i \in \text{nodes}$
  2. $\theta_i \neq \theta_{pa(i)}$ for all $i \in \text{nodes}$ ($pa(i)$ is the parent of node $i$ in $\tau$)
Max Likelihood Estimation for RIM

[M, Meek 14]

- **Problem** Given permutations \( \pi_1, \ldots, \pi_N \), infer \( \tau, \theta \)
- **Identifiability and estimation of \( \theta \)**
  - reorder to obtain canonical representation, with \( \theta_i \geq 0 \) for all \( i \in \text{nodes} \)
  - given \( \tau \), \( \theta_i \) can be estimated by convex univariate minimization

\[
\begin{array}{c}
\text{0.1} \\
\hline
1.2 \\
\hline
\text{apple} & \text{banana} & \text{cherry} & \text{durian} \\
\end{array}
\]

- **Identifiability of \( \tau \)**
  - **Theorem** [M, Meek 14] A model \( \tau, \theta \) is identifiable iff
    1. \( \theta_i > 0 \) for all \( i \in \text{nodes} \)
    2. \( \theta_i \neq \theta_{pa(i)} \) for all \( i \in \text{nodes} \) (\( pa(i) \) is the parent of node \( i \) in \( \tau \))

- **Hardness of \( \tau \) estimation**
  - Estimating \( \pi_0 \) is NP-hard [Duchi, Mackey, Jordan 13]
  - Estimating \( \tau \) structure given \( \pi_0 \) is tractable
Sufficient statistics

\[ Q(d|a|b|c) = \begin{array}{cccc|c}
    a & b & c & d \\
    \hline
    - & 1 & 0 & 0 & a \\
    0 & - & 1 & 0 & b \\
    0 & 0 & - & 0 & c \\
    1 & 1 & 1 & - & d
\end{array} \]
Sufficient statistics

\[
Q(d|a|b|c) = \begin{bmatrix}
-1 & 1 & 1 & 0 \\
0 & - & 1 & 0 \\
0 & 0 & - & 0 \\
1 & 1 & 1 & -
\end{bmatrix}
\]

Cost\( (d|a|b|c) = 0.1 \times 2 + 1.2 \times 0 + 0.4 \times 1 \)
Max Likelihood Estimation algorithm(s)

- Estimating $\tau$ given $\pi_0$ is tractable
Max Likelihood Estimation algorithm(s)

- Estimating $\tau$ given $\pi_0$ is tractable
  - by Dynamic Programming (DP) algorithm, similar to Matrix Chain Multiplication, Inside(-Outside) algorithm $O(n^4)$
  - contains $\theta_j$ estimation at each DP “partial solution”

- Estimating $\pi_0$: Stochastic local search over $\pi_0$ space, similar to Simulated Annealing
  1. Sample $\pi_0^{new}$ from proposal distribution current $P_{\tau, \theta}$
  2. Given $\pi_0^{new}$, find $\tau^{opt}, \theta^{opt}$ by Dynamic Programming
  3. Bring to canonical form $\Rightarrow \tau^{new}, \theta^{new} \geq 0$
  4. Compute log-likelihood score, accept/reject like in Metropolis-Hastings, return to step 1
Experiments - Sushi preferences data

Data

$N = 5000$ permutations of $n = 10$ items

Compared with:

- **alph** $\pi_0$ fixed, $\tau, \theta | \pi_0$ optimize
- **GM** fixed $\tau$, optimize $\pi_0, \theta$
- **HG** fixed $\tau$ from [Huang, Guestrin, 12], optimize $\theta$
- **SA** Simulated Annealing

Test set log-likelihood w.r.t SA

$N_{test} = 300, N_{train} = 4700$, 30 replicates
Beyond sufficient statistics – handling partial rankings

“Sushi preference” data $n = 12$

- types of sushi
  - “My top 3 preferences are ika, maguro, tekka, in this order”
  - “I like uni least of all”
  - “I prefer fish to non-fish”

Three good things about the RIM

- RIM is a general model (includes Mallows, generalized Mallows)
- likelihood $P(\pi | \tau(\tilde{\theta}))$ factors according to tree (and partition function $Z$ tractable)
- RIM has sufficient statistics
Beyond sufficient statistics – handling partial rankings

“Sushi preference” data $n = 12$

Types of sushi

- ika | maguro | tekka | {all other types}
- {all but ebi} | ebi
- {sake, anago, ...} | {tamago, ika, ...}

Diagram showing partial rankings with probabilities:

- $E_1$: ika, maguro, tekka, {all other types}
- $E_2$: {all but ebi} | ebi

Probabilities:

- ika: 0.1
- maguro: 0.9
- tekka: 0.0
- {all other types}: 0.6
- {all but ebi}: 0.2
- ebi: 0.4
- sake, anago: 0.6
- tamago, ika: 0.1
- kappa, uni: 0.0
Beyond sufficient statistics – handling partial rankings

“Sushi preference” data \( n = 12 \) types of sushi

- ika | maguro | tekka | \{ all other types \}
- \{ all but ebi \} | ebi
- \{ sake, anago, . . . \} | \{ tamago, ika, . . . \}

Partial ranking \( \sigma \) [Huang & al, 10]

\[
\sigma = (E_1|E_2|\ldots|E_K)
\]

- \( E_1 \cup E_2 \cup \ldots \cup E_K = \) set of items
- \( \text{shape} \ (n_1, \ldots, n_K) \), \( n_k = |E_k|, \sum n_k = n \)
Beyond sufficient statistics – handling partial rankings

“Sushi preference” data $n = 12$

Partial ranking $\sigma$ [Huang & al, 10]

$\sigma = (E_1 | E_2 | \ldots | E_K)$ with

- $E_1 \cup E_2 \cup \ldots E_K =$ set of items
- shape $(n_1, \ldots n_K)$,
  $n_k = |E_k|$, $\sum n_k = n$

Three good things about the RIM

- RIM is a general model (includes Mallows, generalized Mallows)
- likelihood $P(\pi | \tau(\tilde{\theta}))$ factors according to tree ? YES [Huang et al, 10]
- RIM has sufficient statistics ? NO
Inferences with partial rankings in the RIM. Are they tractable?

The meaning of “tractable”

- Estimation of \( \pi_0 \) for RIM is intractable in the worst case
- We define tractable as \( \mathcal{O}(N \ poly(n)) \times \) time (memory) for complete data
Inferences with partial rankings in the RIM. Are they tractable?

The meaning of “tractable”
- Estimation of $\pi_0$ for RIM is intractable in the worst case
- We define tractable as $\mathcal{O}(N \text{poly}(n)) \times$ time (memory) for complete data

Main technical difficulty
- marginal probability of a partial ranking $\sigma$

$$P(\sigma|\tau(\tilde{\theta})) = \sum_{\pi \sim \sigma} P(\pi|\tau(\tilde{\theta}))$$

where linear extension $\{\pi \sim \sigma\}$ of $\sigma$ can have exponential size
Contributions

1. for marginal probability $P(\sigma|\tau(\hat{\theta}))$
   ▶ exact formula and polynomial algorithm
   ▶ proved algorithm no more than $2Nn$ more costly than for complete permutations (and sometimes much faster)

2. for pairwise marginals $E[Q_{ab}] = Pr[a \text{ precedes } b \mid \sigma, \tau(\hat{\theta})]$
   ▶ exact recursive (polynomial) algorithm
   ▶ proved algorithm no more costly than for complete permutations

3. for parameter $\hat{\theta}$ estimation (Maximum Likelihood)
   ▶ convex univariate minimization algorithm for each $\theta i$
   ▶ proved algorithm is $\mathcal{O}(Nn)$ more costly than for complete permutations

4. for structure search (Maximum Likelihood)
   previous work
   ▶ complete data: local (simulated annealing) search algorithm with exact, tractable steps [Meek M 14]
   ▶ partial rankings: EM algorithm with approximate (or exponential) E step [Huang & al 10]
   our contributions
   ▶ new “E step” based on completing the pairwise marginals $E[Q_{ab}]$
   ▶ algorithms above can use the completed pairwise marginals as if they were complete data
Computing the marginal probability $P(\sigma|\tau, \theta)$

RIM probability for complete data $P(\pi|\tau, \theta)$
(with $v_i =$ number of inversions of $\pi_0$ at node $i$)

$$P_{\tau, \theta}(\pi) = \prod_{i \in \text{nodes}} \frac{e^{-\theta_i v_i}}{G_{L_i, R_i}(\exp(-\theta_i))]$$

with $G_{L, R}(q) = \frac{(q)_{L+R}}{(q)_L(q)_R}$, $(q)_n = \prod_{i=1}^{n}(1 - q^i)$.

RIM probability for partial ranking $\sigma$
[M, Meek in prep]

$$P_{\tau, \theta}(\sigma) = \prod_{i \in \text{nodes}} \text{(factor at node } i)$$
Marginal $P(\pi|\tau, \tilde{\theta})$ for partial ranking $\sigma$

Sufficient to consider root node

Complete ranking $\pi = (c|a|b|d)$

\[
\text{factor} = \frac{e^{-2\theta}}{G_{2,2}(e^{-\theta})}
\]

Partial ranking $\sigma = (c|\{a, b, d\})$

\[
\text{factor} = \frac{e^{-2\theta}G_{0,1}(e^{-\theta})G_{2,1}(e^{-\theta})}{G_{2,2}(e^{-\theta})}
\]
Marginal $P(\pi|\tau, \tilde{\theta})$ for partial ranking $\sigma$

Sufficient to consider root node
Complete ranking $\pi = (c|a|b|d)$

Partial ranking $\sigma = (c|\{a, b, d\})$

factor $= \frac{e^{-2\theta}}{G_{2,2}(e^{-\theta})}$

In general, at some internal node where

- set $\mathcal{L}$ is merged with set $\mathcal{R}$
- partial ranking $\sigma$ restricted to $\mathcal{L} \cup \mathcal{R}$ is $E_1|E_2|\ldots|E_K$ with $E_k = L_k \cup R_k$, $L_k \subseteq \mathcal{L}$, $r_k \subseteq \mathcal{R}$
- factor of $P(\sigma|\tau(\tilde{\theta}))$ at this node is

$$g(l_1:k, r_1:k, \theta) = \frac{e^{-\theta v}G_{l_1,r_1}(e^{-\theta})G_{l_2,r_2}(e^{-\theta})\ldots G_{l_K,r_K}(e^{-\theta})}{G_{|\mathcal{L}|,|\mathcal{R}|}(e^{-\theta})}$$

where $v = \#$ inversions in $\sigma$ at node $\leq \#$ inversions in $\pi \sim \sigma$
Marginal $P(\pi|\tau, \theta)$ – how much extra computation?

How many additional factors?

Rem 1 $G_{0,r} = G_{l,0} = 1$
Marginal $P(\pi | \tau, \hat{\theta})$ – how much extra computation?

How many additional factors?

Rem 1 $G_{0,r} = G_{l,0} = 1$

Rem 2 at each node, at least one of $L_k, R_k$ decreases (and their initial sum is $n$)
  - Hence, no more than $n - 1$ extra factors (but sometimes much fewer)
Marginal $P(\pi | \tau, \tilde{\theta})$ – how much extra computation?

How many additional factors?

Rem 1 $G_{0,r} = G_{l,0} = 1$

Rem 2 at each node, at least one of $L_k, R_k$ decreases (and their initial sum is $n$)
  - Hence, no more than $n - 1$ extra factors (but sometimes much fewer)
  - Example top-$t$ rankings $\sigma = (ika | maguro | sake | \{everything else\}) P(\sigma | \tau, \tilde{\theta})$ has at most $t - 1$ non-trivial factors
Marginal $P(\pi|\tau, \tilde{\theta})$ — how much extra computation?

How many additional factors?

Rem 1 \( G_{0,r} = G_{l,0} = 1 \)

Rem 2 at each node, at least one of \( L_k, R_k \) decreases (and their initial sum is \( n \))

- Hence, no more than \( n - 1 \) extra factors (but sometimes much fewer)
- Example top-\( t \) rankings \( \sigma = (ika|maguro|sake|\{everything else\}) \) \( P(\sigma|\tau, \tilde{\theta}) \) has at most \( t - 1 \) non-trivial factors

How much additional computation?

- \( G_{L,R} \) is computed recursively over \( l = 0, \ldots L, r = 1, \ldots R \)
- Hence, all \( G_{l,r}(\theta) \) in numerator are cached while computing the denominator

- Overhead for whole sample of size \( N \) is no more than \( nN \) lookups+multiplications
- For comparison, for a complete whole sample
  - computation of sufficient statistics is \( \mathcal{O}(n^2N) \)
  - computation of \( Z \) given \( \tilde{\theta} \) is \( \mathcal{O}(n^2 \log n) \)
Independence properties

- define $Q_{ab} = 1$ iff $a$ precedes $b$
- $Q_{ab} \perp Q_{cd}$ whenever $\text{path}(a, b) \cap \text{path}(c, d) = \emptyset$
Independence properties

- define $Q_{ab} = 1$ iff $a$ precedes $b$
- $Q_{ab} \perp Q_{cd}$ whenever $\text{path}(a, b) \cap \text{path}(c, d) = \emptyset$
- Independence checking can reveal the “branching structure” (but not $\pi_0$)
- In progress: combine independence tests with local search to estimate $\pi$
Conclusion: No need to compromise!

Goals of inference in models on permutations
- Flexible w.r.t observation model (i.e. input data)
  - partial rankings, pairwise observations
- Flexible w.r.t generative model
  - RIMs are a class of flexible, identifiable, interpretable models
- Exact and tractable algorithms, closed form expression
Outline

Permutations and their representations
  Statistical models for permutations and the dependence of ranks
  Codes, inversion distance and the precedence matrix
  Mallows models over permutations

Complete rankings and Maximum Likelihood estimation
  GM as exponential family

Top-t rankings, infinite permutations, and Bayesian estimation
  Top-t rankings and infinite permutations
  Conjugate prior, Dirichlet process mixtures

Recursive inversion models and finding common structure in preferences

[Signed permutations and the reversal median problem]
Signed permutations in genetics

- DNA = ordered lists of genes
- Reversals (rearrangements) = a contiguous segment of the DNA is reversed in place, direction of the genes changes
Signed permutations in genetics

- DNA = ordered lists of genes
- Reversals (rearrangements) = a contiguous segment of the DNA is reversed in place, direction of the genes changes

Transforming Human into Mouse From P. Pevzner “Computational Molecular Biology”

6 reversals that involve 8 linkage groups

These transformations define $W_n$ the hyperoctahedral group on $\{1:n\}$
- The elements of $W_n$ are called signed permutations
Signed permutations. Three representations

- Signed permutation $\pi = [4 \ 2 \ 1 \ 3]$

- Reflected representation of $\pi$: $\pi^{\text{ref}} = [4 \ 2 \ 1 \ 3 \ 
3 \ 1 \ 2 \ 4]$

- Precedence matrix $C(\pi)$

  \[
  C_{ii'} = 1_{i \prec i'}
  \]

<table>
<thead>
<tr>
<th>e</th>
<th>1 2 3 4</th>
<th>4 3 2 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1 0 0 0</td>
<td>1 0 1 0</td>
</tr>
<tr>
<td>2</td>
<td>0 - 0 0</td>
<td>1 0 0 0</td>
</tr>
<tr>
<td>3</td>
<td>1 1 - 0</td>
<td>1 1 0 0</td>
</tr>
<tr>
<td>4</td>
<td>1 1 1 -</td>
<td>1 1 1 1</td>
</tr>
<tr>
<td>4</td>
<td>0 0 0 0</td>
<td>- 0 0 0</td>
</tr>
<tr>
<td>3</td>
<td>1 1 0 0</td>
<td>1 - 0 0</td>
</tr>
<tr>
<td>2</td>
<td>1 1 1 0</td>
<td>1 1 - 1</td>
</tr>
<tr>
<td>1</td>
<td>1 1 1 0</td>
<td>1 1 0 -</td>
</tr>
</tbody>
</table>

Group theory

hyperoctaedral group $W_n = \text{group of signed permutations of order } n$

Generators $\{\tau_1, \tau_2, \ldots, \tau_n, w_n\}$ with $w_n = \text{sign change at rank } n$

$\tau_j = \text{elementary transposition of ranks } j \text{ and } j + 1$

let $I = [1, 2, \ldots, n, n \ldots 2, 1]$

$\pi^{\text{ref}} = \text{permutation of } I \text{ such that } \pi^{\text{ref}}_j = \pi_j$ and $\pi^{\text{ref}}_{j+n} = \pi_j$ for $j \leq n$.

E.g. identity gives $\text{id}^{\text{ref}} = [1 \ldots n \ n \ldots 1]$
Inversion distance – algorithmic view

- Inversion distance \( d(\pi, \pi_0) = \# \text{ steps to bubble sort } \pi \text{ into } \pi_0 \)
- \( c_j(\pi|\pi_0) = \# \text{ steps to bring item } i = \pi_0(j) \text{ to } j^{th} \text{ position in } \pi^{ref} \)

\[
d(\pi, \pi_0) = c_1(\pi|\pi_0) + c_2(\pi|\pi_0) + \ldots + c_n(\pi|\pi_0)
\]

- Code of \( \pi \) w.r.t \( \pi_0 \) \( c(\pi|\pi_0) = (c_j(\pi|\pi_0))_{j=1:n} \)

Example \( \pi = [4 2 1 3], \pi_0 = [3 1 2 4] \)

<table>
<thead>
<tr>
<th>( j )</th>
<th>( \pi_0(j) )</th>
<th>action</th>
<th>current ( \pi^{ref} )</th>
<th>( c_j )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>move 3 left 3 steps, delete 3</td>
<td>([4 \ 2 \ 1 \ 3 1 \ 2 4])</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>move 1 left 3 steps, delete 1</td>
<td>([3 \ 1 \ 4 2 \ 1 2 \ 4])</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>move 2 left 1 step, delete 2</td>
<td>([3 \ 1 \ 2 4 \ 4])</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4 already in place, delete 4</td>
<td>([3 \ 1 \ 2 4])</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ 7 = d(\pi, \pi_0) \]

Algorithm **DISTANCE**(\( \pi, \pi_0 \))
- Represent \( \pi \) in reflected form \( \pi^{ref} \)
- For \( j = 1 : n \) ranks in \( \pi_0 \)
  1. let \( i = (\pi_0)_j \) the rank \( j \) element of \( \pi_0 \)
  2. move \( i \) left in \( \pi^{ref} \) to rank \( j \) by adjacent transpositions
  3. delete \( i \) from the list

Output: \( d(\pi, \pi_0) = \text{the total number of adjacent transpositions} \)
Consensus ranking for signed permutations [M, Arora 12]

- one can formulate consensus ranking w.r.t inversion distance on $\mathcal{W}_n$
- one can define Mallows, GM models, conjugate priors on $\mathcal{W}_n$
- sufficient statistics are (subtriangle) of precedence matrix
- estimation/consensus ranking by B&B algorithm

$$
\begin{array}{ccccccc}
& 3 & 1 & 2 & 4 & 4 & 2 & 1 & 3 \\
\pi_{0}^\text{ref} & - & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\
1 & 0 & - & 0 & 0 & 1 & 1 & 1 & 1 \\
2 & 1 & 1 & - & 0 & 1 & 1 & 1 & 1 \\
3 & 1 & 1 & 1 & - & 1 & 1 & 1 & 1 \\
4 & 0 & 0 & 0 & 0 & - & 0 & 0 & 0 \\
4 & 0 & 0 & 0 & - & 1 & - & 0 & 0 \\
3 & 1 & 1 & 0 & 0 & 1 & 1 & - & 1 \\
2 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & - \\
1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & - \\
\end{array}
$$
A surrogate for the reversal median

- Reversal distance $r(\pi, \pi_0) = \#$ reversals to turn $\pi$ into $\pi_0$
  (one reversal = several inversions)

- Reversal median problem: find $\pi_0$ minimizing

$$R(\pi_0) = \min_{\pi_0 \in \mathbb{W}_n} \sum_{k=1}^{m} r(\pi_k, \pi_0)$$

- Relevant in biology, known NP-hard, no practical algorithms in use

- Idea: Approximate reversal median by inversion median (a.k.a. consensus ranking)
When is this approximation good?

Assumptions

A1 \( \pi \) generated by \( r \) random reversals from \( \pi_0 \)
A2 sample size \( N \to \infty \) (asymptotic regime)
A3 each reversal independent of previous ones
A4 “number inversions/reversal not too large”

Theorem [M, in preparation] Under A1–4, we can show numerically that
\[
\arg\min_{\pi} \mathbb{E}[d(\pi, \pi_0)] = \arg\min_{\pi} \mathbb{E}[r(\pi, \pi_0)]
\]

Intuition

C matrices generated by random reversals

1 reversal 2 reversals 3 reversals 10 reversals
Does it work? Synthetic data

Sample size $N = 50, \ldots, 2000$ from $\mathbb{W}_n$, generated by $r = 1, 2, 3$ random reversals; results are averages over 10 runs.

<table>
<thead>
<tr>
<th>$r$</th>
<th>$N$</th>
<th>Objective $D(\hat{\pi}_0)$</th>
<th>Distance $d(\hat{\pi}_0, \pi^{true})$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$A_{\text{STAR}}$ $A_{\text{GREEDY}}$ $A_{\text{RAND}}$</td>
<td>$A_{\text{STAR}}$ $A_{\text{GREEDY}}$ $A_{\text{RAND}}$</td>
</tr>
<tr>
<td>1</td>
<td>50</td>
<td>125.0 125.6 370</td>
<td>0 1.2 135</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>120.8 129.0 370</td>
<td>0 16.5 134.7</td>
</tr>
<tr>
<td>1</td>
<td>1000</td>
<td>125.5 125.5 365</td>
<td>0 0 140.7</td>
</tr>
<tr>
<td>1</td>
<td>2000</td>
<td>119.1 129.9 362</td>
<td>0 25.2 136.9</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>168.8 170.1 338</td>
<td>0 4.4 139.3</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>175.4 186.1 336</td>
<td>0 43.3 153.4</td>
</tr>
<tr>
<td>2</td>
<td>1000</td>
<td>174.5 175.0 337</td>
<td>0 1.5 146.4</td>
</tr>
<tr>
<td>2</td>
<td>2000</td>
<td>171.4 182.5 340</td>
<td>9 47.3 149.4</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>203.0 205.6 325</td>
<td>0 15.3 143.2</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>198.1 206.4 330</td>
<td>21.1 57.1 135.7</td>
</tr>
<tr>
<td>3</td>
<td>1000</td>
<td>202.9 205.3 326</td>
<td>0 14.3 125.5</td>
</tr>
<tr>
<td>3</td>
<td>2000</td>
<td>201.1 210.7 324</td>
<td>49.4 94.5 132.6</td>
</tr>
<tr>
<td></td>
<td>$n = 50$</td>
<td>$n = 24$</td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>---------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AStar</td>
<td>Greedy</td>
<td>Rand</td>
</tr>
<tr>
<td>1</td>
<td>50</td>
<td>372.4</td>
<td>383.5</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>363.4</td>
<td>414.0</td>
</tr>
<tr>
<td>1</td>
<td>1000</td>
<td>370.3</td>
<td>370.3</td>
</tr>
<tr>
<td>1</td>
<td>2000</td>
<td>382.8</td>
<td>455.1</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>601.5</td>
<td>619.6</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>613.0</td>
<td>676.3</td>
</tr>
<tr>
<td>2</td>
<td>1000</td>
<td>601.5</td>
<td>613.5</td>
</tr>
<tr>
<td>2</td>
<td>2000</td>
<td>595.0</td>
<td>666.6</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>746.6</td>
<td>772.8</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>739.5</td>
<td>798.8</td>
</tr>
<tr>
<td>3</td>
<td>1000</td>
<td>748.2</td>
<td>768.8</td>
</tr>
<tr>
<td>3</td>
<td>2000</td>
<td>744.2</td>
<td>806.1</td>
</tr>
</tbody>
</table>

Median of (runtime $A_{\text{Star}}$ / runtime $G_{\text{reedy}}$) over 10 runs

<table>
<thead>
<tr>
<th>$N$</th>
<th>100</th>
<th>1000</th>
<th>100</th>
<th>1000</th>
<th>100</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r$</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

$n = 50$

| 3.5 | 3.5 | 3.4 | 3.4 | 3.4 |

$n = 24$

| 2.25 | 3   | 3   | 3   | 5   | 3   |
Results on Metazoan mtDNA data [Bourque & Pevzner 2002]

tree built using B&B

[Bourque & Pevzner 2002] binary tree
Conclusions

Why models based on inversions?
- Recognized as good/useful in applications
- Complementarity:
  - Utility based ranking models (Thurstone)
  - Stagewise ranking models (GM) – combinatorial
- Nice computational properties/Analyzable statistically
- The code grants GM its tractability
  - representation with independent parameters

The bigger picture
- Ranked data have rich structure
  - computationally incompletely exploited
  - structure of preferences incompletely modeled
- Statistical analysis of rankings combines
  - combinatorics, algebra
  - algorithms
  - statistical theory
- Modeling aspects
  - infinite number of items [MBao 08, 10]
  - top-t and other partial observations [MBao 08, MChen 10, MMeek–in prep]
  - flexible structure (RIM) [MeekM 14]
  - other finite groups (signed permutations/hyperoctahedral group) [MArora 13]
  - consistency, rates [MBa0 10]
  - conjugate prior [MBao]

- Algorithmic aspects
  - Maximum likelihood estimation algorithms and sufficient statistics [MPhadnisPattersonBilmes 04, 05, MandhaniM 08, MAli 10]
  - Bayesian inference and sampling [MChen 10, MChen 16]
Thank you