(Bayesian) Statistics with Rankings

Marina Meilă
University of Washington
www.stat.washington.edu/mmp

with Alnur Ali, Harr Chen, Bhushan Mandhani, Le Bao, Kapil Phadnis, Artur Patterson, Brendan Murphy, Jeff Bilmes

Permutations (rankings) data represents preferences

Burger preferences $n = 6, N = 600$

```
med-rare med rare ...
done med-done med ...
med-rare rare med ...
```

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
```

Ranking data

- discrete
- many valued
- combinatorial structure

The Consensus Ranking problem

Given a set of rankings $\{\pi_1, \pi_2, \ldots \pi_N\} \subset \mathbb{S}_n$ find the consensus ranking (or central ranking) π_0 that best agrees with the data

Elections Ireland, $n = 5$, $N = 1100$

`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 $\{\pi_1, \pi_2, \ldots \pi_N\} \subset \mathbb{S}_n$ find the consensus ranking (or central ranking) π_0 such that

$$\pi_0 \;=\; \operatorname*{argmin}_{\mathbb{S}_n} \sum_{i=1}^{N} d(\pi_i, \pi_0)$$

for $d =$ inversion distance / Kendall τ-distance / "bubble sort" distance

The Consensus Ranking problem

Problem (also called Preference Aggregation, Kemeny Ranking)
Given a set of rankings $\{\pi_1, \pi_2, \ldots \pi_N\} \subset \mathbb{S}_n$ find the consensus ranking (or central ranking) π_0 such that

$$\pi_0 \;=\; \operatorname*{argmin}_{\mathbb{S}_n} \sum_{i=1}^{N} d(\pi_i, \pi_0)$$

for d = inversion distance / Kendall τ-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])

Equivalent to finding the "mean" or "median" of a set of points

The Consensus Ranking problem

Problem (also called Preference Aggregation, Kemeny Ranking)
Given a set of rankings $\{\pi_1, \pi_2, \ldots \pi_N\} \subset \mathbb{S}_n$ find the consensus ranking (or central ranking) π_0 such that

$$\pi_0 \;=\; \operatorname*{argmin}_{\mathbb{S}_n} \sum_{i=1}^{N} d(\pi_i, \pi_0)$$

for d = inversion distance / Kendall τ-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])

Equivalent to finding the "mean" or "median" of a set of points

Fact: Consensus ranking for the inversion distance is NP hard

Consensus ranking problem

$$\pi_0 \;=\; \operatorname*{argmin}_{\mathbb{S}_n} \sum_{i=1}^{N} d(\pi_i, \pi_0)$$

This talk

Will generalize the problem

- from finding π_0
  to estimating statistical model

Will generalize the data

- From complete, finite permutations
  to top-t rankings, countably many items ($n \to \infty$)...

Outline

Some notation

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

$\mathbb{S}_n$ = the symmetric group = the set of all permutations over n items

$\pi = [\, c\, a\, b\, d \,] \in \mathbb{S}_n$ a permutation/ranking

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

$t = |\pi| \leq n$ the length of π

We observe
data $\pi_1, \pi_2, \ldots, \pi_N \sim$ sampled independently from distribution P over $\mathbb{S}_n$
(where P is unknown)

Representations for permutations

reference permutation $\mathrm{id} = [\, a\; b\; c\; d \,]$

$$\pi = [\, c\; a\; b\; d \,]$$ ranked list

$$(2\;3\;1)$$ cycle representation

$$\begin{array}{ccccc} & a & b & c & d \\ {[} & 2 & 3 & 1 & 4 \;] \end{array}$$ function on $\{a, b, c, d\}$

$$\Pi = \begin{array}{|c|c|c|c|} \hline 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ \hline 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 \\ \hline \end{array}$$ permutation matrix

$$Q = \begin{array}{|c|c|c|c|} \hline - & 1 & 0 & 1 \\ \hline 0 & - & 0 & 1 \\ \hline 1 & 1 & - & 1 \\ \hline 0 & 0 & 0 & - \\ \hline \end{array}$$ precedence matrix, $\quad Q_{ij} = 1$ if $i \prec_\pi j$,

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

$$(s_1, s_2, s_3) = (2, 0, 0)$$

Representations for permutations

reference permutation $\mathrm{id} = [\,a\ b\ c\ d\,]$

$$\pi = [\,c\ a\ b\ d\,] \quad \text{ranked list}$$

$$(2\ 3\ 1) \quad \text{cycle representation}$$

$$\begin{array}{cccccc} & a & b & c & d & \\ [& 2 & 3 & 1 & 4 &] \end{array} \quad \text{function on } \{a, b, c, d\}$$

$$\Pi = \begin{array}{|c|c|c|c|} \hline 0 & 0 & 1 & 0 \\ \hline 1 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 \\ \hline \end{array} \quad \text{permutation matrix}$$

$$Q = \begin{array}{|c|c|c|c|} \hline - & 1 & 0 & 1 \\ \hline 0 & - & 0 & 1 \\ \hline 1 & 1 & - & 1 \\ \hline 0 & 0 & 0 & - \\ \hline \end{array} \quad \text{precedence matrix,} \quad Q_{ij} = 1 \text{ if } i \prec_\pi j$$

$$(V_1, V_2, V_3) = (1, 1, 0) \quad \text{code}$$

$$(s_1, s_2, s_3) = (2, 0, 0)$$

Thurstone: Ranking by utility

The Thurstone Model

- item j has *expected utility* μ_j
- sample $u_j = \mu_j + \epsilon_j$, $j = 1:n$ (independently or not)
  u_j is the *actual utility* of item j
- sort $(u_j)_{j=1:n}$ to obtain a π

Thurstone: Ranking by utility

The Thurstone Model

- item j has *expected utility* μ_j
- sample $u_j = \mu_j + \epsilon_j$, $j = 1 : n$ (independently or not)
  u_j is the *actual utility* of item j
- sort $(u_j)_{j=1:n}$ to obtain a π

- rich model class
- typically $\epsilon_j \sim Normal(0, \sigma_j^2)$
- parameters interpretable
- some simple probability calculations are intractable
  - $P[a \prec b]]$ tractable, $P[i \,\text{in first place}]$ tractable
  - $P[i \,\text{in 85th place}]$ intractable

- each rank of π depends on all the ϵ_j

Plackett-Luce: Ranking as drawing without replacement

The Plackett-Luce model

- item j has *weight* $w_j > 0$

$$P([a,\, b,\, \ldots]) \;\propto\; \frac{w_a}{\sum_{i'} w_{i'}} \frac{w_b}{\sum_{i'} w_{i'} - w_a} \cdots$$

- items are drawn "without replacement" from distribution $(w_1,\, w_2 \ldots w_n)$ (Markov chain)
- normalization constant Z generally not known
- distribution of first ranks approximately independent

- item at rank j depends on all previous ranks

Bradley-Terry: penalizing inversions

The Bradley-Terry model

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

- exponential family model
- one parameter for every pair $)i, j)$
- α_{ij} is penalty for inverting i with j
  only qualitative interpretation
- normalization constant Z generally not known

- transitivity $i \prec j, j \prec k \implies i \prec k$
  therefore the sufficient statistics Q_{ij} are dependent

Bradley-Terry: penalizing inversions

The Bradley-Terry model

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

- exponential family model
- one parameter for every pair $)i,j)$
- α_{ij} is penalty for inverting i with j
  only qualitative interpretation
- normalization constant Z generally not known

- transitivity $i \prec j,\, j \prec k \implies i \prec k$
  therefore the sufficient statistics Q_{ij} are dependent
- Mallows models
  - are a subclass of Bradley-Terry models
  - do not suffer from this dependence
  - coming next...

Outline

The precedence matrix Q

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

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

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

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

reference permutation $\mathrm{id} = [\, a\, b\, c\, d \,]$: determines the order of rows, columns in Q

The number of inversions and Q

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

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

define

- $L(Q) = \sum_{i>j} Q_{ij} =$ sum(lower triangle (Q))

The number of inversions and Q

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

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

define

- $L(Q) = \sum_{i>j} Q_{ij}$ = sum(lower triangle (Q))

then

- $\#\text{inversions}(\pi) = L(Q) = d(\pi, \mathrm{id})$

The inversion distance and Q

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

Refence permutation

$\mathrm{id} = [\, a\; b\; c\; d \,]$

$Q(\pi)$

	a	b	c	d	
	–	1	0	1	a
	0	–	0	1	b
	1	1	–	1	c
	0	0	0	–	d

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

Reference permutation

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

$\Pi_0^T Q(\pi) \Pi_0$

	b	a	d	c	
	–	0	1	0	b
	1	–	1	0	a
	0	0	–	0	d
	1	1	1	–	c

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

The inversion distance and Q

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

1. Construct $Q(\pi)$
2. Sort rows and columns by π_0
3. Sum elements in lower triangle

The inversion distance and Q

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

1. Construct $Q(\pi)$
2. Sort rows and columns by π_0
3. Sum elements in lower triangle

Note also that
To obtain
$d(\pi_1, \pi_0) + d(\pi_2, \pi_0) + \dots$

1. Construct $Q(\pi_1), Q(\pi_2), \dots$
2. Sum
   $Q = Q(\pi_1) + Q(\pi_2) + \dots$
3. Sort rows and columns of Q by π_0
4. Sum elements in lower triangle of Q

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

b	a	d	c	
–	0	1	0	b
1	–	1	0	a
0	0	–	0	d
1	1	1	–	c

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

A decomposition for the inversion distance

$d(\pi, \pi_0) = \#$ inversions between π and π_0

$$\begin{aligned} d([\,c\,a\,b\,d\,],[\,b\,a\,d\,c\,]) &= \#\underbrace{(\text{inversions w.r.t } b)}_{V_1} \\ &+ \#\underbrace{(\text{inversions w.r.t } a)}_{V_2} \\ &+ \#\underbrace{(\text{ inversions w.r.t } d)}_{V_3} \\ &+ \ldots \end{aligned}$$

$V_j = \#$ inversions where $\pi_0(j)$ is disfavored

The code of a permutation

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

	a	b	c	d	
S_2	–	1	0	1	a
S_3	0	–	0	1	b
S_1	1	1	–	1	c
S_4	0	0	0	–	d
	V_1	V_2	V_3	V_4	

code

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

The code of a permutation

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

	a	b	c	d	
S_2	–	1	0	1	a
S_3	0	–	0	1	b
S_1	1	1	–	1	c
S_4	0	0	0	–	d
	V_1	V_2	V_3	V_4	

code

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

or

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

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

The code of a permutation

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

	a	b	c	d	
S_2	–	1	0	1	a
S_3	0	–	0	1	b
S_1	1	1	–	1	c
S_4	0	0	0	–	d
	V_1	V_2	V_3	V_4	

code

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

or

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

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

Codes are defined w.r.t any π_0

	b	a	d	c	
S_3	–	0	1	0	b
S_2	1	–	1	0	a
S_4	0	0	–	0	d
S_1	1	1	1	–	c
	V_1	V_2	V_3	V_4	

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

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

The code of a permutation

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

Codes are defined w.r.t any π_0

	a	b	c	d	
S_2	–	1	0	1	a
S_3	0	–	0	1	b
S_1	1	1	–	1	c
S_4	0	0	0	–	d
	V_1	V_2	V_3	V_4	

	b	a	d	c	
S_3	–	0	1	0	b
S_2	1	–	1	0	a
S_4	0	0	–	0	d
S_1	1	1	1	–	c
	V_1	V_2	V_3	V_4	

code

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

or

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

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

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

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

or

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

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

Codes and inversion distance summary

Inversion distance facts

- $d(\pi, \pi_0) = \sum_j V_j(\pi|\pi_0) = \sum_j S_j(\pi|\pi_0)$

Codes and inversion distance summary

Inversion distance facts

- $d(\pi, \pi_0) = \sum_j V_j(\pi|\pi_0) = \sum_j S_j(\pi|\pi_0)$
- $d(\pi, \pi_0) = L(\Pi_0^T Q(\pi)\Pi_0) \stackrel{def}{=} L_{\pi_0}(Q(\pi))$

Codes facts

- $(V_{1:n-1})$ or $(S_{1:n-1})$ defined w.r.t any reference permutation
  - we denote them $V_j(\pi|\pi_0)$ or $S_j(\pi|\pi_0)$

Codes and inversion distance summary

Inversion distance facts

- $d(\pi, \pi_0) = \sum_j V_j(\pi|\pi_0) = \sum_j S_j(\pi|\pi_0)$
- $d(\pi, \pi_0) = L(\Pi_0^T Q(\pi)\Pi_0) \stackrel{def}{=} L_{\pi_0}(Q(\pi))$

Codes facts

- $(V_{1:n-1})$ or $(S_{1:n-1})$ defined w.r.t any reference permutation
  - we denote them $V_j(\pi|\pi_0)$ or $S_j(\pi|\pi_0)$
- $(V_{1:n-1})$ or $(S_{1:n-1})$ uniquely represent π
  - with $n-1$ independent parameters

	b	a	d	c	
S_3	–	0	1	0	b
S_2	1	–	1	0	a
S_4	0	0	–	0	d
S_1	1	1	1	–	c
	V_1	V_2	V_3	V_4	

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

The Mallows Model

- **The Mallows model** is a distribution over $\mathbb{S}_n$ defined by

$$P_{\pi_0,\theta}(\pi) \;=\; \frac{1}{Z(\theta)} e^{-\theta d(\pi,\pi_0)}$$

- π_0 is the central permutation
  - π_0 mode of $P_{\pi_0,\theta}$, unique if $\theta > 0$
- $\theta \geq 0$ is a dispersion parameter
  - for $\theta = 0$, $P_{\pi_0,0}$ is uniform over $\mathbb{S}_n$

The Mallows Model

- **The Mallows model** is a distribution over $\mathbb{S}_n$ defined by

$$P_{\pi_0,\theta}(\pi) \;=\; \frac{1}{Z(\theta)} e^{-\theta d(\pi,\pi_0)}$$

- π_0 is the central permutation
  - π_0 mode of $P_{\pi_0,\theta}$, unique if $\theta > 0$
- $\theta \geq 0$ is a dispersion parameter
  - for $\theta = 0$, $P_{\pi_0,0}$ is uniform over $\mathbb{S}_n$

- $d(\pi,\pi_0) = \sum_j V_j(\pi|\pi_0)$ therefore $P_{\pi_0,\theta}$ is product of $P_\theta(V_j(\pi|\pi_0)$

$$P_{\pi_0,\theta}(\pi) \;=\; \frac{1}{Z(\theta)} \prod_{j=1}^{n-1} e^{-\theta V_j(\pi|\pi_0)} \;\; \text{and} \;\; Z(\theta) \;=\; \prod_{j=1}^{n-1} \underbrace{\frac{1-e^{-\theta(n-j+1)}}{1-e^{-\theta}}}_{Z_j(\theta)}$$

The Generalized Mallows (GM) Model [Fligner, Verducci 86]

Mallows model $P_{\pi_0,\theta}(\pi) = \frac{1}{Z_\theta} \exp\left(-\theta \sum_{j=1}^{n-1} V_j(\pi|\pi_0)\right)$

Idea: $\theta \rightarrow \vec{\theta} = (\theta_1, \theta_2, \ldots \theta_{n-1})$

Generalized Mallows(GM) model

$$P_{\pi_0,\vec{\theta}}(\pi) = \frac{1}{Z(\vec{\theta})} \prod_{j=1}^{n-1} e^{-\theta_j V_j(\pi|\pi_0)} \qquad \text{with} \qquad Z(\vec{\theta}) = \prod_{j=1}^{n-1} Z_j(\theta_j)$$

The Generalized Mallows (GM) Model [Fligner, Verducci 86]

Mallows model $P_{\pi_0,\theta}(\pi) = \frac{1}{Z_\theta} \exp\left(-\theta \sum_{j=1}^{n-1} V_j(\pi|\pi_0)\right)$
Idea: $\theta \to \vec{\theta} = (\theta_1, \theta_2, \ldots \theta_{n-1})$
Generalized Mallows(GM) model

$$P_{\pi_0,\vec{\theta}}(\pi) = \frac{1}{Z(\vec{\theta})} \prod_{j=1}^{n-1} e^{-\theta_j V_j(\pi|\pi_0)} \qquad \text{with} \qquad Z(\vec{\theta}) = \prod_{j=1}^{n-1} Z_j(\theta_j)$$

Similar definitions with S_j instead of V_j: models denoted GM^V, GM^S

The Generalized Mallows (GM) Model [Fligner, Verducci 86]

Mallows model $P_{\pi_0,\theta}(\pi) = \frac{1}{Z_\theta} \exp\left(-\theta \sum_{j=1}^{n-1} V_j(\pi|\pi_0)\right)$

Idea: $\theta \ \rightarrow \ \vec{\theta} = (\theta_1, \theta_2, \ldots \theta_{n-1})$

Generalized Mallows(GM) model

$$P_{\pi_0,\vec{\theta}}(\pi) \ = \ \frac{1}{Z(\vec{\theta})} \prod_{j=1}^{n-1} e^{-\theta_j V_j(\pi|\pi_0)} \qquad \text{with} \quad Z(\vec{\theta}) \ = \ \prod_{j=1}^{n-1} Z_j(\theta_j)$$

Similar definitions with S_j instead of V_j: models denoted GM^V, GM^S **Cost interpretation of the GM models**

- GM^V: Cost $= \sum_j \theta_j V_j$

  pay price θ_j for every inversion w.r.t item j
- GM^S: Cost $= \sum_j \theta j S_j$

  pay price θ_j for every inversion in picking rank j
- Assume stepwise construction of π: θ_j represents importance of step j

Outline

The (Max Likelihood) estimation problem

Burger preferences $n = 6$, $N = 600$

```
med-rare med rare ...
done med-done med ...
med-rare rare med ...
```

- Data $\{\pi_i\}_{i=1:N}$ i.i.d. sample from $\mathbb{S}_n$
- Model Mallows $P_{\pi_0,\theta}$ or GM $P_{\pi_0,\vec{\theta}}$

- **Parameter estimation:** π_0 known,estimate θ or $\vec{\theta}$.
  This problem is easy (convex, univariate)

- **Central permutation estimation:** $\vec{\theta}$ known, estimate π_0
  Known as **Consensus ranking** if $\theta = 1$ ($\approx$**MinFAS**)
  This problem is NP hard. (many heuristic/approx. algorithms exist)

- **General estimation:** estimate both π_0 and θ or $\vec{\theta}$.
  ...at least as hard as consensus ranking. Will show it's no harder.

The likelihood

- **Likelihood** of π_0, θ = P[data $| \pi_0, \theta$]
- **Max Likelihood estimation** ${\pi_0}^*, \theta^*$ = argmax P[data $| \pi_0, \theta$]

Mallows

$$logl(\theta, \pi_0) = \frac{1}{N} \ln P(\pi_{1:N}; \theta, \pi_0) = -\theta \sum_{j=1}^{n-1} \frac{\sum_{i=1}^{N} V_j(\pi_| \pi_0)}{N} + \sum_{j=1}^{n-1} \ln Z_j(\theta)$$

Generalized Mallows

$$logl(\theta, \pi_0) = \frac{1}{N} \ln P(\pi_{1:N}; \theta, \pi_0) = -\sum_{j=1}^{n-1} [\theta_j \overbrace{\frac{\sum_{i=1}^{N} V_j(\pi_i | \pi_0)}{N}}^{\bar{V}_j} + \ln Z_j(\theta_j)]$$

The likelihood

- **Likelihood** of π_0, θ = P[data $| \pi_0, \theta$]
- **Max Likelihood estimation** $\pi_0{}^*, \theta^*$ = argmax P[data $| \pi_0, \theta$]

Mallows

$$logl(\theta, \pi_0) = \frac{1}{N} \ln P(\pi_{1:N}; \theta, \pi_0) = -\theta \sum_{j=1}^{n-1} \frac{\sum_{i=1}^{N} V_j(\pi_{|}\pi_0)}{N} + \sum_{j=1}^{n-1} \ln Z_j(\theta)$$

Generalized Mallows

$$logl(\theta, \pi_0) = \frac{1}{N} \ln P(\pi_{1:N}; \theta, \pi_0) = -\sum_{j=1}^{n-1} [\theta_j \overbrace{\frac{\sum_{i=1}^{N} V_j(\pi_i|\pi_0)}{N}}^{\bar{V}_j} + \ln Z_j(\theta_j)]$$

- Likelihood is separable and concave in each $\theta_j \Longrightarrow$ estimation of θ_j is straightforward
  - by convex minimization of $\theta_j \bar{V}_j + \ln Z_j(\theta_j)$ (numerical)
- Dependence on π_0 complicated

ML Estimation of π_0: costs and main results

	$\pi_{1:N}$ complete rankings (GM^s, GM^V)	$\pi_{1:t}$ top-t rankings, $N \leq \infty$ (only GM^s)
Mallows	$\sum_{j=1}^{n-1} \frac{\sum_i V_j(\pi\|\pi_0)}{N}$	$\sum_{j=1}^{t} \frac{\sum_i S_j(\pi\|\pi_0)}{N}$
GM	$\sum_{j=1}^{n-1} \left[\theta_j \frac{\sum_i V_j(\pi_i\|\pi_0)}{N} + \ln Z_j(\theta_j)\right]$	$\sum_{j=1}^{t} \left[\theta_j \frac{\sum_i S_j(\pi_i\|\pi_0)}{N} + \ln Z_j(\theta_j)\right]$
Mallows	[M&al07] ${\pi_0}^{ML}$ can be found exactly by B&B search on matrix $Q(\pi_{1:N})$.	[MBao08] ${\pi_0}^{ML}$ can be found exactly by B&B search on matrix $R(\pi_{1:N})$.
GM	[M&al07] ${\pi_0}^{ML}, \vec{\theta}^{ML}$ can be found exactly by B&B search on matrix $Q(\pi_{1:N})$.	[MBao08] A local maximum for $\pi_0, \vec{\theta}$ can be found by alternate maximization: $\pi_0\|\vec{\theta}$ by B&B, $\vec{\theta}\|\pi_0$ by convex unidimensional.

$Q(\pi_{1:N}) = \sum_{i=1:N} Q(\pi_i)$ $\quad$ $R(\pi_{1:N}) = \sum_{i=1:N} R(\pi_i)$ (defined next

B&B = branch-and-bound

- the search may not be tractable

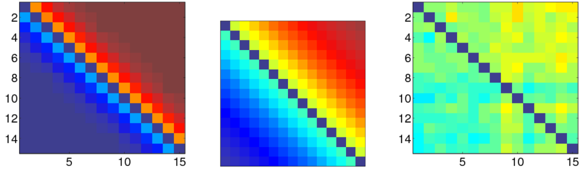
Sufficient statistics (complete permutations) [M&al07]

$Q(\pi)$

—	0	1	0
1	—	1	0
0	0	—	0
1	1	1	—

Q for large samples from Mallows models

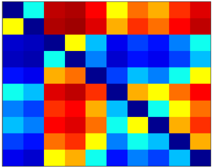
$\theta = 1$ $\theta = 0.3$ $\theta = 0.03$



- 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

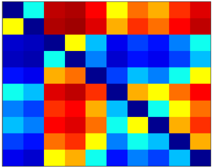
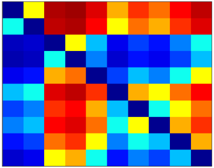
Search Algorithm Idea

Wanted: $\mathrm{argmin}_{\pi_0} \mathrm{L}(\Pi_0^{\mathrm{T}} \mathrm{Q} \Pi_0) = \mathrm{argmin}_{\pi_0} \mathrm{L}_{\pi_0}(\mathrm{Q})$ = argmin lower triangle of Q over all row and column permutations



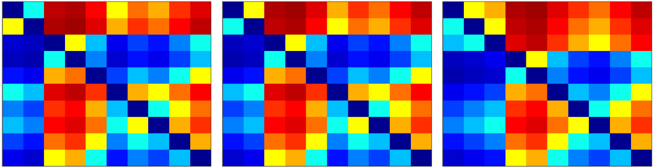
Search Algorithm Idea

Wanted: $\text{argmin}_{\pi_0} \mathrm{L}(\Pi_0^T Q \Pi_0) = \text{argmin}_{\pi_0} \mathrm{L}_{\pi_0}(Q) =$ argmin lower triangle of Q over all row and column permutations

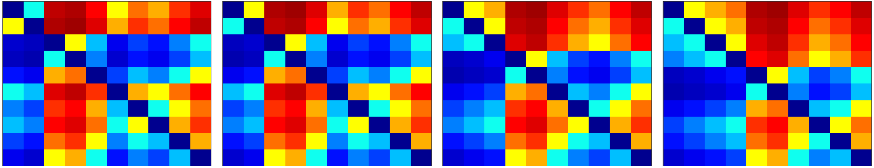
Search Algorithm Idea

Wanted: $\mathrm{argmin}_{\pi_0} \mathrm{L}(\Pi_0^{\mathrm{T}} \mathrm{Q} \Pi_0) = \mathrm{argmin}_{\pi_0} \mathrm{L}_{\pi_0}(\mathrm{Q})$ = argmin lower triangle of Q over all row and column permutations



Search Algorithm Idea

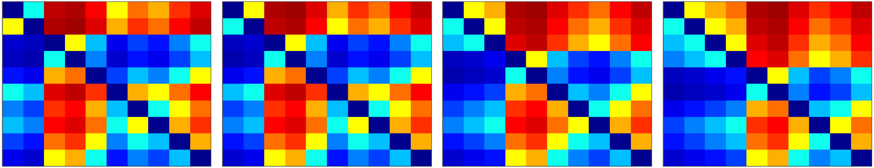
Wanted: $\mathrm{argmin}_{\pi_0} \mathrm{L}(\Pi_0^{\mathrm{T}} \mathrm{Q} \Pi_0) = \mathrm{argmin}_{\pi_0} \mathrm{L}_{\pi_0}(\mathrm{Q}) =$ argmin lower triangle of Q over all row and column permutations



...

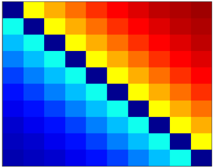
Search Algorithm Idea

Wanted: $\mathrm{argmin}_{\pi_0} \mathrm{L}(\Pi_0^{\mathrm{T}} \mathrm{Q} \Pi_0) = \mathrm{argmin}_{\pi_0} \mathrm{L}_{\pi_0}(\mathrm{Q}) =$ argmin lower triangle of Q over all row and column permutations



. . .

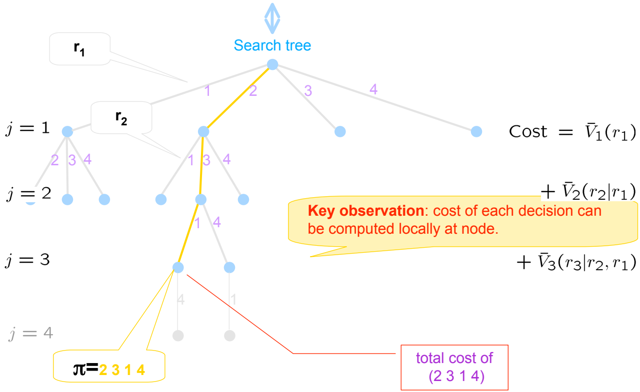
. . .



The Branch-and-Bound Algorithm

$$\pi_0^{-1} = \operatorname*{argmin}_{r_1, r_2, \ldots r_{n-1}} \sum_{j=1}^{n-1} \bar{V}_j(r_j)$$

Total cost of a permutation



Branch and Bound algorithm

Node ρ stores: $r_j, \text{parent}\,, j = |\rho|,\ V_j(\rho),\ \theta_j,\ C(\rho),\ L(\rho)$; $S =$ priority queue with nodes to be expanded.

Initialize: $S = \{\rho_\emptyset\},\ \rho_\emptyset =$the empty sequence, $j = 0,\ C(\rho_\emptyset) = V(\rho_\emptyset) = L(\rho_\emptyset) = 0$

Repeat

remove $\rho \in \operatorname*{argmin}_{\rho \in S} L(\rho)$ from S

if $|\rho| = n$ *(Return)* **Output** $\rho,\ L(\rho) = C(\rho)$ and **Stop**.

else *(Expand ρ)*

for $r_{j+1} \in [n] \setminus \rho$ create node $\rho' = \rho | r_{j+1},\ V_{j+1}(\rho') \ = \ V_j(r_{1:j-1}, r_{j+1}) - Q_{r_j r_{j+1}}$
compute $V^{min} = \min_{r_{j+1} \in [n] \setminus \rho} V_{j+1}(\rho | r_{j+1})$
calculate $A(\rho)$ admissible heuristic [MandhaniM09]
for $r_{j+1} \in [n] \setminus \rho$

calculate θ_{j+1} from $n - j - 1, V_{j+1}(\rho'))$
$C(\rho') \ = \ C(\rho) + \theta_{j+1} V_{j+1}(\rho'),\ L(\rho') \ = \ C(\rho') + A(\rho),$
store node $(\rho',\ j + 1,\ V_{j+1},\ \theta_{j+1},\ C(\rho'),\ L(\rho'))$ in S

ML Estimation of π_0: costs and main results

	$\pi_{1:N}$ complete rankings (GM^s, GM^V)	$\pi_{1:t}$ top-t rankings, $N \leq \infty$ (only GM^s)
Mallows	$\sum_{j=1}^{n-1} \frac{\sum_i V_j(\pi\|\pi_0)}{N}$	$\sum_{j=1}^{t} \frac{\sum_i S_j(\pi\|\pi_0)}{N}$
GM	$\sum_{j=1}^{n-1} \left[\theta_j \frac{\sum_i V_j(\pi_i\|\pi_0)}{N} + \ln Z_j(\theta_j)\right]$	$\sum_{j=1}^{t} \left[\theta_j \frac{\sum_i S_j(\pi_i\|\pi_0)}{N} + \ln Z_j(\theta_j)\right]$
Mallows	[M&al07] ${\pi_0}^{ML}$ can be found exactly by B&B search on matrix $Q(\pi_{1:N})$.	[MBao08] ${\pi_0}^{ML}$ can be found exactly by B&B search on matrix $R(\pi_{1:N})$.
GM	[M&al07] ${\pi_0}^{ML}, \vec{\theta}^{ML}$ can be found exactly by B&B search on matrix $Q(\pi_{1:N})$.	[MBao08] A local maximum for $\pi_0, \vec{\theta}$ can be found by alternate maximization: $\pi_0\|\vec{\theta}$ by B&B, $\vec{\theta}\|\pi_0$ by convex unidimensional.

$Q(\pi_{1:N}) = \sum_{i=1:N} Q(\pi_i)$ $\quad$ $R(\pi_{1:N}) = \sum_{i=1:N} R(\pi_i)$ (defined next

B&B = branch-and-bound

- the search may not be tractable

Algorithm summary

- Sufficient statistics $= Q(\pi_{1:N})$
- $\text{Cost}(\pi_0, \theta) = \theta L_{\pi_0}(Q(\pi_{1:N}))$ (lower triangle of Q after permuting rows and columns by π_0
- B&B Algorithm constructs π_0 one rank at a time
- Exact but not always tractable

- B&B Algorithms exist also for
  - GM^S
  - for multiple parameters $\vec{\theta}$
- Performance issues
  - Admissible heuristics help
  - Beam search and other approximations possible

What makes the search hard (or tractable)?

Running time $=$ time(compute Q) $+$ time(B&B)
$\mathcal{O}(n^2 N)$ independent of N

- Number nodes explored by B&B
  - independent of sample size N
  - independent of π_0
  - depends on dispersion $\vec{\theta}^{ML}$
- $\vec{\theta} = 0 \;\Rightarrow\;$ uniform distribution
  - all branches have equal cost
- $\theta^{ML}_{1:n-1}$ large $\Rightarrow$ likelihood decays fast around ${\pi_0}^{ML}$ $\Rightarrow$ pruning efficient
- Theoretical results
  - e.g if $\theta_j > T_j$, $j = 1 : n-1$, then B&B search defaults to greedy
- Practically
  - diagnoses possible during B&B run

Admissible heuristics

To guarantee optimality we need lower bounds for the cost-to-go (admissible heuristics)

- admissible heuristic for Mallows Model [MPPB07]
- improved heuristic for Mallows model [Mandhani,M 09], first admissible heuristic for GMM model

- If data $\sim P_{\theta,\pi_0}$ with θ large, admissible heuristic A known $\Rightarrow$ number of expanded nodes is bounded above

Related work I

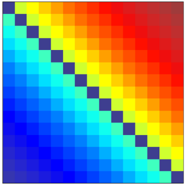
ML Estimation

[FV86] $\vec{\theta}$ estimation; heuristic for π_0

FV algorithm/Borda rule

1. Compute $\bar{q}_j, j = 1 : n$ column sums of Q
2. Sort $(\bar{q}_j)_{j=1}^n$ in increasing order; π_0 is sorting permutation

- $\bar{q}_j$ are Borda counts
- FV is consistent for infinite N



Related work II

Consensus Ranking $(\theta = 1)$
[CSS99] CSS ALGORITHM = greedy search on Q
improved by extracting strongly connected components
[Ailon,Newman,Charikar 05] Randomized algorithm guaranteed 11/7 factor approximation (ANC)
[Mohri, Ailon 08] linear program
[Mathieu, Schudy 07] $(1+\epsilon)$ approximation, time $\mathcal{O}(n^6/\epsilon + 2^{2^{O(1/\epsilon)}})$
[Davenport,Kalagnanan 03] Heuristics based on edge-disjoint cycles used by our B&B implementation
[Conitzer,D,K 05] Exact algorithm based on integer programming, better bounds for edge disjoint cycles (DK)
[Betzler,Brandt, 10] Exact problem reductions

- Most of this work based on the MinFAS view

$$Q_{ij} > .5 \quad \Leftrightarrow \quad i\bullet \xrightarrow{Q_{ij}-.5} \bullet j$$

Prune graph to a DAG removing minimum weight

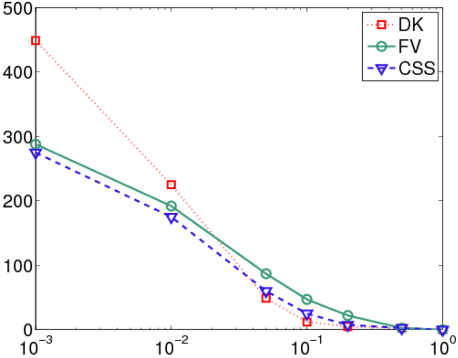
Related work III

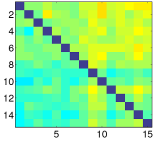
Extensions and applications to social choice

- Inferring rakings under partial and aggregated information [ShahJabatula08], [JabatulaFariasShah10]
- Vote elicitation under probabilistic models of choice [LuBoutillier11]
- Voting rules viewed as Maximum Likelihood [ConitzerSandholm08]
- ...

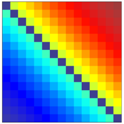
When is the B&B search tractable? I

Excess cost w.r.t B&B; data from Mallows model $n = 100$, $N = 100$

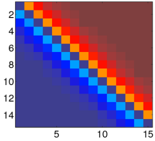




hard (uninteresting?)



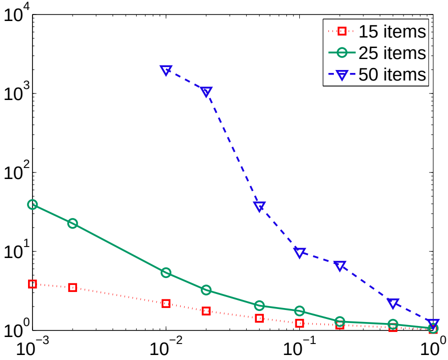
interesting



easy

Running time vs number items n

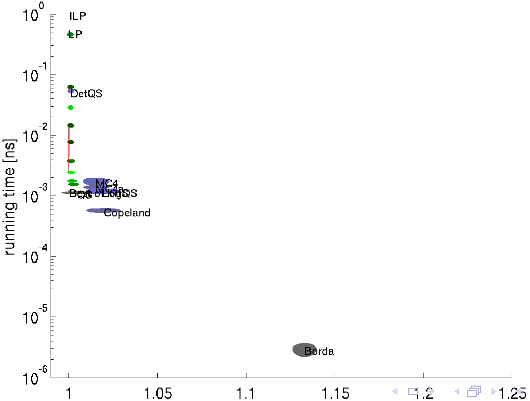
Data generated from Mallows(θ)



Extensive comparisons

- Experimental setup from [Coppersmith&al07]. Experiments by Alnur Ali [AliM11]
- Data: artificial (Mallows and Plackett-Luce), `Ski`, `Web-search`
  total 45 data sets, $n = 50 \ldots 350$, $N = 4 \ldots 100$ typically
- Algorithms `ILP`, `LP`, `B&B` (with limited queue), `Local Search (LS)`, `FV/Borda`, `QuickSort (QS)`, ... and combinations (total 104 algorithms)

Websearch data B&B is competitive (`Local Search`, `B&B`,`other`)



Other statistical models on rankings

Several "natural" parametric distributions on $\mathbb{S}_n$ exist.

- $P(\pi) \propto \exp\left(-\sum_{j=1}^{n-1} \theta_j V_j(\pi)\right)$ *Generalized Mallows*
- $P(\pi) \propto \exp\left(-\sum_{i<j} \alpha_{ij} Q_{ij}(\pi)\right)$ *Bradley-Terry*

$\boxed{\text{Mallows} \subset \text{GM} \subset \text{Bradley-Terry}}$

Other statistical models on rankings

Several "natural" parametric distributions on $\mathbb{S}_n$ exist.

- $P(\pi) \propto \exp\left(-\sum_{j=1}^{n-1} \theta_j V_j(\pi)\right)$ *Generalized Mallows*
- $P(\pi) \propto \exp\left(-\sum_{i<j} \alpha_{ij} Q_{ij}(\pi)\right)$ *Bradley-Terry*

$\boxed{\text{Mallows} \subset \text{GM} \subset \text{Bradley-Terry}}$

- item j has weight $w_j > 0$ *Plackett-Luce*

$$P([a, b, \ldots]) \propto \frac{w_a}{\sum_{i'} w_{i'}} \frac{w_b}{\sum_{i'} w_{i'} - w_a} \cdots$$

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

	GM	B-T	P-L	T
Discrete parameter	yes	no	no	no
Tractable Z	yes	no	no	no
"Easy"* param estimation	yes	no	no	Gauss
Tractable marginals	yes	no	no	Gauss**
Params "interpretable"	yes	no	no	Gauss

* 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

Outline

Top-t rankings and very many items

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
```

Bing search: `UW Statistics` $n \to \infty$

```
www.stat.washington.edu/
www.stat.wisc.edu/
www.stat.washington.edu/courses
collegeprowler.com/university-of-washington/statistics
...
```

Models for Infinite permutations

- **Domain** of items to be ranked is countable, i.e $n \longrightarrow \infty$
- **Observed** the top t ranks of an infinite permutation
- Examples
  - Bing $\boxed{\texttt{UW Statistics}}$
    `www.stat.washington.edu/`
    `www.stat.wisc.edu/`
    `www.stat.washington.edu/courses`
    `collegeprowler.com/university-of-washington/statistics`
    ...
  - searches in data bases of biological sequences (by e.g Blast, Sequest, etc)
  - open-choice polling, "grassroots elections", college program applications
- 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

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

$$\begin{array}{llll} \pi(1) = c & S_1 & = & 2 \\ \pi(2) = a & S_2 & = & 0 \\ \pi(3) = ? & S_3 & = & ? \end{array}$$

$$\begin{array}{llll} \pi_0(1) = a & V_1 & = & 1 \\ \pi_0(2) = b & V_2 & \geq & 1 \\ \pi_0(3) = c & V_3 & = & 0 \end{array}$$

$$P_{\pi_0,\vec{\theta}}(\pi) = \prod_{j=1}^{t} e^{-\theta_j S_j}$$

$$P_{\pi_0,\theta}(\pi) = \prod_{j=1}^{n-1} \begin{cases} e^{-\theta V_j}, & \pi_0(j) \in \pi \\ P_\theta(V_j \geq v_j), & \pi_0(j) \notin \pi \end{cases}$$

sufficient statistics

no sufficient statistics

Example: $\pi = [\, c\, a \,]$

$$Q(\pi) = \begin{array}{c|c|c|c|c|c} & a & b & c & d & \\ \hline S_2 & - & 1 & 0 & 1 & a \\ & 0 & - & 0 & ? & b \\ S_1 & 1 & 1 & - & 1 & c \\ & 0 & ? & 0 & - & d \\ \hline & V_1 & V_2 & V_3 & V_4 & \end{array}$$

The Infinite Generalized Mallows Model (IGM) [MBao08]

$$P_{\pi_0,\vec{\theta}}(\pi) \;=\; \frac{1}{\prod_{j=1}^{t} Z(\theta_j)} \exp\left[-\sum_{j=1}^{t} \theta_j S_j(\pi \mid \pi_0)\right]$$

- distribution over top-t rankings
- π_0 is permutation of $\{1, 2, 3, \ldots\}$

  a discrete infinite "location" parameter
- $\theta_{1:t} > 0$ dispersion parameter

- product of t independent univariate distributions
- Normalization constant $Z(\theta_j) \;=\; 1/(1 - e^{-\theta_j})$
- $P_{\pi_0,\vec{\theta}}(\pi)$ is well defined marginal over the coset defined by π

IGM versus GM

$$P_{\pi_0,\vec{\theta}}(\pi) \;=\; \frac{1}{\prod_{j=1}^{t} Z(\theta_j)} \exp\left[-\sum_{j=1}^{t} \theta_j S_j(\pi \mid \pi_0)\right]$$

- all S_j have same range $\{0, 1, 2, \ldots\}$
- Z has simpler formula
- only top-t rankings observed

Sufficient statistics for top-t permutations [MBao09]

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

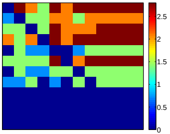
Lemma

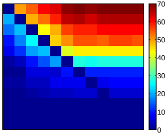
$$S_j(\pi|\pi_0) \;=\; L_{\pi_0}(R_j(\pi))$$

$$R_j(\pi) \;=\; \begin{array}{c|c|c|c|c|}
 & - & & & \\ \hline
 & & - & & \\ \hline
\pi(j) & 0 & 1 & - & 1 \\ \hline
 & & & & - \\ \hline
\end{array}$$

- $(R_j)_{kl} \;=\; 1$ iff item k at rank j and item l after k (observed or not)
- $(R_1, \dots R_t)$ sufficient statistics for multiple θ GM^s
- $R \;=\; \sum_{j=1}^{t} R_j$ sufficient statistics for single θ $Mallows^s$

$N = 2,\; n = 12$ $\qquad$ $N = 100,\; n = 12,\; t = 5$





Infinite Mallows Model: ML estimation

Theorem[M,Bao 08]

- Sufficient statistics

n	# distinct items observed in data
T	# total items observed in data
$Q = [Q_{kl}]_{k,l=1:n}$	frequency of $k \prec l$ in data
$q = [q_k]_{k=1:n}$	frequency of k in data
$\boxed{R = q\mathbf{1}^T - Q}$	sufficient statistics matrix

- log-likelihood$(\pi_0, \theta) = \theta L_{\pi_0}(R) = \theta$ Sum (Lower triangle (R permuted by π_0))
- The optimal ${\pi_0}^{ML}$ can be found exactly by a B&B algorithm searching on matrix R.
- The optimal θ^{ML} is given by

$$\theta \ = \ \log\left(1 + T/L_{\pi_0}(R)\right)$$

Infinite GMM: ML estimation

Theorem [M,Bao 08]

- Sufficient statistics

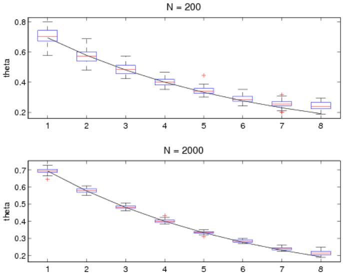
n	# distinct items observed in data
N_j	# total permutations with length $\geq j$
$Q^{(j)} = [Q^{(j)}_{kl}]_{k,l=1:n,\, j=1:t}$	frequency of $1_{[\pi(k)=j,\, \pi(l)<j]}$ in data
$q^{(j)} = [q^{(j)}_k]_{k=1:n}$	frequency of k in rank j in data
$R^{(j)} = q^{(j)}\mathbf{1}^T - Q^{(j)}$	sufficient statistics matrices

- For $\theta_{1:t}$ given, the optimal ${\pi_0}^{ML}$ can be found exactly by a B&B algorithm searching on matrix $R(\vec{\theta}) = \sum_j \theta_j R^{(j)}$.
- the cost is $L_{\pi_0}(R) = \text{Sum}(\text{Lower triangle}(R(\vec{\theta})\text{ permuted by }\pi_0))$
- The optimal ${\theta_j}^{ML}$ is given by $\theta_j \; = \; \log\left(1 + N_j / L_{\pi_0}(R^{(j)})\right)$

Hence, alternate maximization will converge to local optimum

ML Estimation: Remarks

- sufficient statistics Q, q, R finite for finite sample size N but don't compress the data
- data determine only a finite set of parameters
  - π_0 restricted to the observed items
  - θ restricted to the observed ranks



- Similar result holds for finite domains

Outline

GM are exponential family models I

GM^V for complete rankings
GM^S for top-t rankings, n finite or ∞

- have finite sufficient statistics
- are exponential family models in π_0, $\vec{\theta}$ '
- have conjugate priors

Hyperparameters

- $N_0 > 0$ equivalent sample size
- Q^0 (or R_j^0)$\in \mathbb{R}^{n\times n}$ equivalent sufficient statistics

The conjugate prior I

Hyperparameters: $N_0 > 0$, Q^0 (or R_j^0)$\in \mathbb{R}^{n\times n}$
The conjugate prior (for GM^s, top-t, n finite or ∞)

- informative prior for both $\pi_0, \vec{\theta}$

$$
\begin{aligned}
P_0(\pi_0, \vec{\theta}) &\propto e^{-N_0 \sum_{j=1}^{t} (\theta_j L_{\pi_0}(R_j^0) + \ln Z_j(\theta_j))} \\
&\propto e^{-N_0 \sum_{j=1}^{t} (\text{sum of lower triangle}(\Pi_0 R_j^0 \Pi_0^T \Theta) + \ln Z_j(\theta_j))} \\
&\propto e^{-N_0 D(P_{{\pi_0}^0, \vec{\theta}^0} || P_{\pi_0, \vec{\theta}})}
\end{aligned}
$$

  with ${\pi_0}^0, \vec{\theta}^0$ ML estimates of sufficient statistics $R^0_{1:t}$, Π_0 the permutation matrix of π_0, Θ=diagonal matrix of $\vec{\theta}$

- non-informative for π_0

$$
P_0(\pi_0, \vec{\theta} | I_{1:t}, N_0) \propto e^{-N_0 \sum_{j=1}^{t} (\theta_j r_j + \ln Z_j(\theta_j))}
$$

Bayesian Inference: What operations are tractable?

Posterior $P_0(\pi_0, \vec{\theta}) \propto e^{\sum_j (\theta_j (N_0 r_j + N L_{\pi_0}(R_j)) + (N_0 + N) \ln Z(\theta_j))}$

- computing unnormalized prior, posterior ✓
- computing normalization constant of prior, posterior ?
- MAP estimation: produces ${\pi_0}^{Bayes}, \vec{\theta}^{Bayes}$ ✓ (by B&B)
- model averaging
  $P(\pi \mid N_0, r, \pi_{1:N}) = \sum_{\pi_0} \int_0^{\infty} GM^s(\pi|\pi_0, \theta) P(\pi_0, \theta | N_0, r, \pi_{1:N}) d\theta$?
- sample from $P(\pi_0, \theta | N_0, r, \pi_{1:N})$ Sometimes

- Bayesian Non-Parameteric Clustering (aka Dirichlet Process Mixture Models DPMM)
  - Is is efficient?

Clustering with Dirichlet mixtures via MCMC

General DPMM estimation algorithm [Neal03]

MCMC estimation for Dirichlet mixture

Input $\alpha, g_0, \beta, \{f\}, \mathcal{D}$

State cluster assignments $c(i)$, $i = 1:n$, parameters θ_k for all distinct k

terate
1. for $i = 1:n$(reassign data to clusters)
   1. if $n_{c(i)} = 1$ delete this cluster and its $\theta_{c(i)}$
   2. resample $c(i)$ by

$$c(i) = \begin{cases} \text{existing}\, k & \text{w.p} \propto \frac{n_k - 1}{n-1+\alpha} f(x_i, \theta_k) \\ \text{new cluster} & \text{w.p} \frac{\alpha}{n-1+\alpha} \int f(x_i, \theta) g_0(\theta) d\theta \end{cases} \quad (1)$$

   3. if $c(i)$ is new label, sample a new $\theta_{c(i)}$ from g_0
2. (resample cluster parameters)
   for $k \in \{c(1:n)\}$
   1. sample θ_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

utput a state with high posterior

Gibbs Sampling Algorithm for DPM of GM^s [M,Chen 10]

Input Parameters N_0, r, t, data $\pi_{1:n}$; initialization

Denote $c(i)$ = cluster label of π_i, π_{0c}, θ_c, N_c the parameters and sample size for cluster c, $N = \sum N_c$

- Repeat
  1. *Reassign points to clusters*For all points π_i resample c_i
     resample $c(i)$ by

$$c(i) = \begin{cases} \text{existing } c & \text{w.p } \propto \frac{n_k - 1}{n-1+N_0} P(\pi_i | \pi_{0c}, \ldots) \\ \text{new cluster} & \text{w.p } \frac{N_0}{n-1+N_0} Z_1 / n! \end{cases}$$

  2. *Resample cluster parameters*
     For all clusters c
     Sample $\pi_{0c} \sim P(\pi_0; N_0, I, \pi_{i \in c})$ directly for $N_c = 1$, Gibbs $\vec{\theta}|\pi_0$, $\pi_0|\vec{\theta}$ for $N_c > 1$

- We use Lemmas 1–5 (coming next)
  - to approximate the integrals
  - to sample
- Main Idea: replace GM^s with simpler Infinite GM

Integrating the posterior: some results I

Model GM^s, $n = \infty$
Prior uninformative $P_0(\pi_0, \vec{\theta}) \propto e^{-N_0 \sum_j (\theta_j r_j + \ln Z(\theta_j))}$ (improper for π_0!)
$Z(\theta) = \frac{1}{1-e^{-\theta}}$
Data $\pi_1, \ldots \pi_N$ top-t rankings, sufficient statistics $R_{1:t}$, total observed items $t \leq n_{obs} \leq Nt$
Posterior $P_0(\pi_0, \vec{\theta}) \propto e^{\sum_j (\theta_j (N_0 r_j + N L_{\pi_0}(R_j)) + (N_0+N) \ln Z(\theta_j))}$
Denote $S_j = L_{\pi_0}(R_j)$

- **Lemma 1**[MBao08] Posterior of π_0 and $\theta_j | \pi_0$

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

$$P(\pi_0 | N_0, r, \pi_{1:N}) \propto \prod_{j=1}^{t} Beta(N_0 r_j + S_j, N_0 + N + 1)$$

Integrating the posterior: some results II

- **Lemma 2**[MChen10] Normalized posterior for $N = 1$

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

- **Lemma 3** Bayesian averaging over $\vec{\theta}$

$$P(\pi|\pi_0, N_0, r, \pi_{1:N}) = \prod_{j=0}^{t} \frac{Beta(S_j(\pi|\pi_0) + N_0 r_j + S_j, N_0 + N + 2)}{Beta(N_0 r_j + S_j, N_0 + N + 1)}$$

- **Lemma 4** Exact sampling of $\pi_0 \,|\, \vec{\theta}$ from the posterior possible by stagewise sampling.

$$P(\pi_0|\vec{\theta}, N_0, r, \pi_{1:N}) \propto e^{-\sum_j \theta_j \overbrace{L_{\pi_0}(R_j)}^{\bar{V}_j(\pi_0)}}$$

Integrating the posterior: some results III

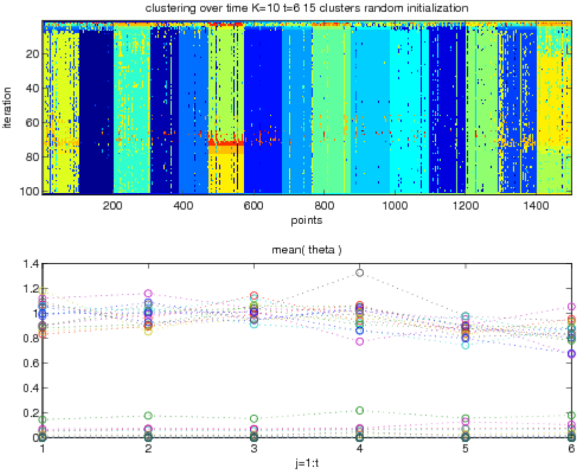
- Posterior of π_0 informative only for items observed in $\pi_{1:N}$, uniform over all other items.
  Wanted: to sum out the permutation of the unobserved items.
  Example: $\pi = [\, c\; a\; b\; d\,]$, data $\pi_{1:N}$ contain $\mathrm{obs} = \{a, c, d, e, \ldots\}$ but not b
- **Lemma 5**

$$
\begin{aligned}
P(\pi \mid \pi_0|_{\mathrm{obs}}) \quad = \quad & \prod_{j:\pi(j)\in \mathrm{obs}} Beta(S_j(\pi|\pi_0) + N_0 r_j + S_j, N_0 + N + 2) \\
& \prod_{j:\pi(j)\notin \mathrm{obs}} Beta(t_j + N_0 r_j + S_j, N_0 + N) \\
& / \prod_{j=0}^{t} Beta(N_0 r_j + S_j, N_0 + N + 1)
\end{aligned}
$$

Useful? Good approximations for n finite

DPMM estimation artificial data

$K = 15$ clusters, $n = 10, t = 6$ $N = 30 \times K$, $\theta_j = 1$



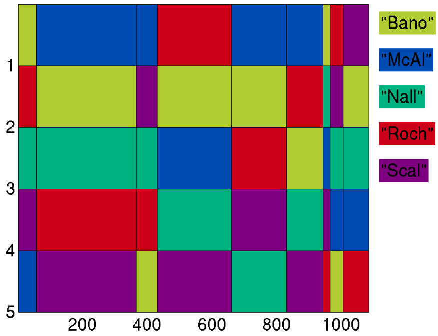
Ireland 2000 Presidential Election

- $n = 5$ candidates, votes=ranked lists of 5 or less
- individuals grouped by preferences

  multimodal distribution
- clustering problem
  - parametric, model based: EM algorithm [Busse07]
  - nonparametric: EBMS Exponential Blurring Mean Shift [MBao08]
  - nonparametric,model based: DPMM Dirichlet Process Mixtures [MChen10]

Ireland Presidential Election

$n = 5, t = 1:5 \ N = 1083$
found 12 clusters, sizes 236,...,1



- **Mary McAleese (Fianna Fail and Progressive Democrats)**
- Rosemary Scallon (Independent)
- Derek Nally (Independent)
- Mary Banotti (Fine Gael)
- Adi Roche (Labour)

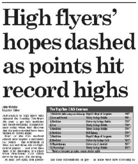
- Work in progress: this clustering different from [Murphy&Gormley]

College program admissions, Ireland

$n = 533$ programs, $N = 53737$ candidates, $t = 10$ 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
```

Students pay price of exam success as points jump



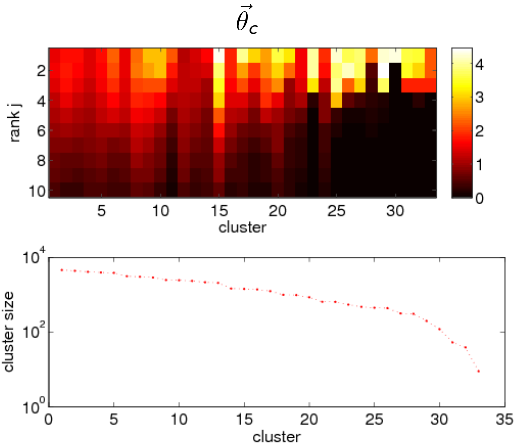
High flyers' hopes dashed as points hit record highs

Masterclass students set new record for grades

Minister insists school subjects are not being 'dumbed down'

- Data = all candidates' rankings for college programs in 2000 from [GormleyMurphy03] (they used EM for Mixture of Plackett-Luce models)
- we [MChen10, Ali Murphy M Chen 10] used DPMM (parameters adjusted to

College program rankings: are there clusters?



- 33 clusters cover 99% of the data
- $\vec{\theta}_c$ parameters large – cluster are concentrated
- number of significant ranks in σ_c, θ_c vary by cluster

College program rankings: are the clusters meaningful?

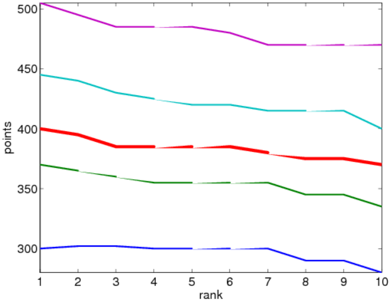
Cluster	Size	Description	Male (%)	Points avg(std)
1	4536	CS & Engineering	77.2	369 (41)
2	4340	Applied Business	48.5	366 (40)
3	4077	Arts & Social Science	13.1	384 (42)
4	3898	Engineering (Ex-Dublin)	85.2	374 (39)
5	3814	Business (Ex-Dublin)	41.8	394 (32)
6	3106	Cork Based	48.9	397 (33)
...	...	...	...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

- Cluster differentiate by **subject area**
- ... also by **geography**
- ... show gender difference in preferences

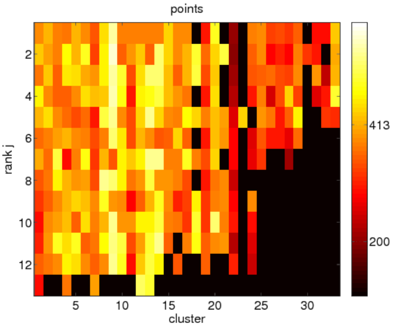
College program rankings: the "prestige" question

- Question: are choices motivated by "prestige" (i.e high **point requirements (PR)**)?
- If yes, then PR should be decreasing along the rankings

PR overall (quantiles)



PR for each cluster and rank



- Unclustered data: PR decreases monotonically with rankings
- Clustered data: PR not always monotonic
  - Simpson's paradox!

Summary: Contributions to the GM model

- For consensus ranking problem: New BB formulation
  - theoretical analysis tool:
    - intuition on problem hardness
    - admissible heuristics provide bounds on run time
  - competitive algorithm in practice
- For top-t rankings (single θ)
  - given correct sufficient statistics - all old algorithms can be used on it
  - BB algorithm (theoretical and practical tool)
- For infinite number of items (single or multiple θ)
  - introduced the Infinite GM model
  - given sufficient statistics, estimation algorithm
  - introduced conjugate prior, studied its properties
- Bayesian estimation/DPMM clustering (for finite top-t rankings)
  - efficient (approximate) Gibbs sampler for DPMM
- (not mentioned here)
  - confidence intervals, convergence rates
  - model selection (BIC for GMM)
  - EBMS non-parametric clustering
  - marginal calculation is polynomial

Conclusions

Why GM model?

- Recongnized 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 it's tractability
  - representation with independent parameters

The bigger picture

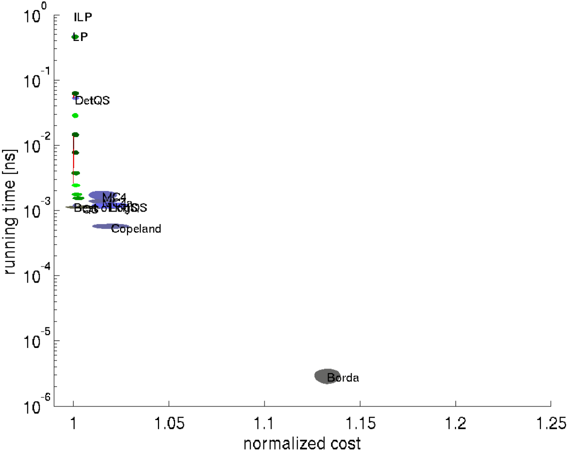
- Statistical analysis of ranking data combines
  - combinatorics, algebra
  - algorithms
  - statistical theory

Thank you

Extensive comparisons I

New experiment Websearch, all relevant algorithms

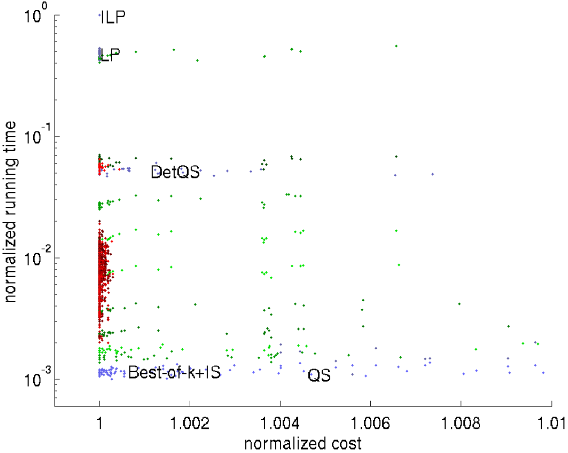
- Local Search, B&B,other



Extensive comparisons II

Websearch data, all relevant algorithms (detail)

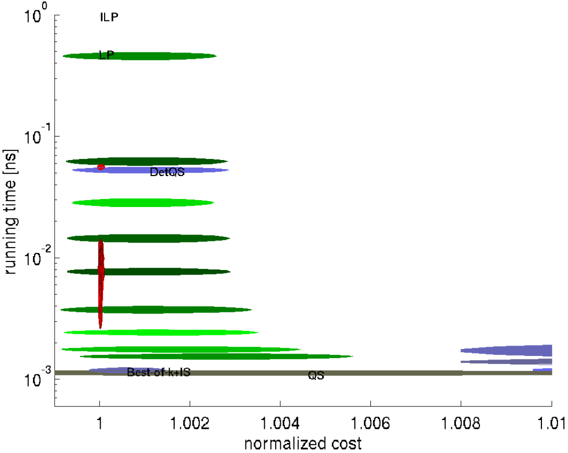
- Local Search, B&B, other



Extensive comparisons III

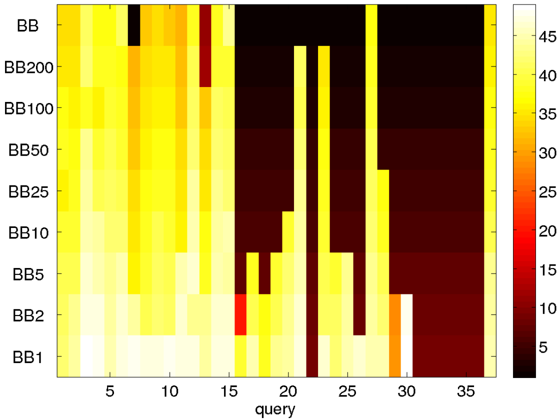
`Websearch` data, all relevant algorithms (more detail)

- Local Search, B&B, other



Extensive comparisons IV

Ranks of B&B algorithms among all other algorithms (cost)



Sufficient statistics spaces I

- space of sufficient statistics $\mathcal{Q} = \{Q = \sum_{1=1}^{n} Q(\pi_i)\} = \text{convex}(\mathbb{S}_n)$

  $\boxed{\mathcal{Q} = \text{convex}_{1+n(n-1)/2}(\mathbb{S}_n) \text{ by Caratheodory's Thm}}$
- space of means (marginal polytope) of GM model $\mathcal{M} = \{ E_{\pi_0,\theta}[Q] \}$

characterized algorithmically [M&al07]; [Mallows 57] for Mallows

- GM model is **curved** exponential family
- Full exponential family = Bradley-Terry model
  - not tractable/ loses nice computational/ interpretational properties
- GM $\subset$ full model [Fligner, Verducci 88] $\subset$ Bradley-Terry
  - open problem: tractable (exact) ML estimation of full model, Bradley-Terry model $\propto \exp\left(-\sum_{i<j} \alpha_{ij} Q_{ij}(\pi)\right)$
    - heuristic [Fligner, Verducci 88] works reasonably well for full model

Consistency and unbiasedness of ML estimates I

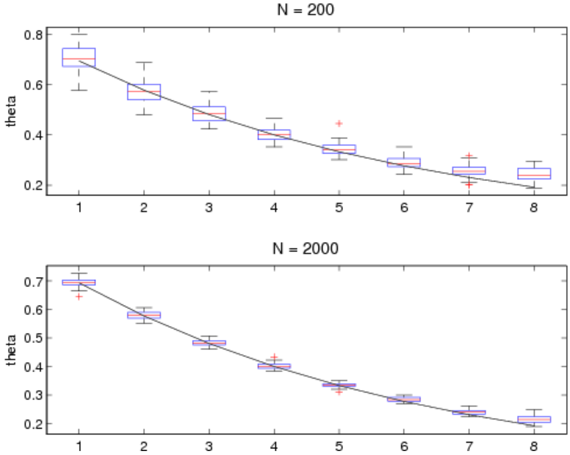
- $Q_{ij}/N \rightarrow P[\text{item } i \prec_{\pi_0} \text{item} j]$ as $N \rightarrow \infty$ [FV86]
- Therefore
  - for any π_0 fixed, $\vec{\theta}^{ML}$ is consistent [FV86]
  - the discrete parameter ${\pi_0}^{ML}$ consistent when θ_j non-increasing [FV86, M in preparation] (joint work with Hoyt Koepke)
    - is it "unbiased"?
- **Theorem 1**[M,in preparation] For any N finite

$$E[\theta^{ML}] > \theta \qquad \boxed{\textbf{Bias!}}$$

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

The Bias of θ^{ML}

- artificial data from Infinite GM
- θ_j estimates for $j = 1 : 8$ and sample sizes $N = 200, 2000$



Convergence rates [M, in preparation] I

Theorem 2 For the Mallows (single θ) model, and sample size N sufficiently large

$$(\sqrt{2ch(\theta)})^{-N} \leq P[{\pi_0}^{ML} \neq \pi_0] \leq \frac{n(n-1)}{2}\left(\sqrt{2ch(\theta)}\right)^{-N}$$

Theorem 3 For the GM model, with $\vec{\theta} > 0$ strongly unimodal, $\vec{\theta}, \pi_0$ unknown

$$P[{\pi_0}^{ML} \neq \pi_0] = \mathcal{O}\left(e^{-c(\vec{\theta})N}\right)$$

- confidence interval for θ in the Mallows model from Theorem 2
- confidence interval for $\vec{\theta}$? in progress