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

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 (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 $\{$ `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

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

$$\Pi = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$ permutation matrix

$$Q = \begin{bmatrix} - & 1 & 0 & 1 \\ 0 & - & 0 & 1 \\ 1 & 1 & - & 1 \\ 0 & 0 & 0 & - \end{bmatrix}$$ precedence matrix, $Q_{ij} = 1$ 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 $\mathrm{id} = [\,a\; b\; c\; d\,]$

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

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

$\overset{a}{2}\;\overset{b}{3}\;\overset{c}{1}\;\overset{d}{4}$ function on $\{a, b, c, d\}$

$\Pi =$ permutation matrix

0	0	1	0
1	0	0	0
0	1	0	0
0	0	0	1

$Q =$ precedence matrix, $Q_{ij} = 1$ if $i \prec_\pi j$

–	1	0	1
0	–	0	1
1	1	–	1
0	0	0	–

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

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

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

- item j has *weight* $w_j > 0$ — *Plackett-Luce*
  sample ranks 1, 2, ... sequentially $\propto$ remaining w_j's

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

- inversion between i and j has *cost* α_{ij} — *Bradley-Terry*

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

  interesting subclasses of the Bradley-Terry
  (Generalized) Mallows models (coming next)
  - are a subclass of Bradley-Terry models
  - do not suffer from these dependencies

	GM	B-T	P-L	Thurstone
Discrete parameter	yes	no	no	no
Tractable Z	yes	no	no	no
"Easy"* parameter 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

The precedence matrix Q

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

$Q(\pi) =$

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

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

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

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

The number of inversions of π and $Q(\pi)$

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

$Q(\pi) =$

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

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

The number of inversions of π and $Q(\pi)$

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

$Q(\pi) =$

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

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

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

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

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

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

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

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

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

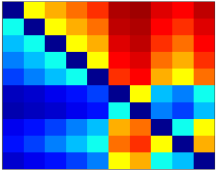
b	a	d	c	
–	0	1	0	b
1	–	1	0	a
0	0	–	0	d
1	1	1	–	c

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

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 π_0
3. Sum elements in lower triangle of Q



The code of a permutation

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

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

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

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

or

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

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

The code of a permutation

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

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

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

or

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

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

- For any π_0, the code $(V_1(\pi|\pi_0)\,\ldots\,V_{n-1}(\pi|\pi_0))$ defines π uniquely

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

Generalized Mallows(GM) model

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

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

Generalized Mallows(GM) model

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

- π_0 is the central permutation
  - π_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 $\mathbb{S}_n$
- $Z_j(\theta_j)$ is tractable

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

Generalized Mallows(GM) model

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

- π_0 is the central permutation
  - π_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 $\mathbb{S}_n$
- $Z_j(\theta_j)$ is tractable

Cost interpretation of the GM model

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

  pay price θ_j for every inversion w.r.t item j
- Assume stepwise construction of π: θ_j represents importance of step j

Outline

ML Estimation of π_0: costs and main results

Model	Data	Log-likelihood	
Mallows	complete rankings	$\sum_{j=1}^{n-1} \bar{V}_j(\pi_0)$	[M&al07] ${\pi_0}^{ML}$ estimated exactly by B&B search.
GM^V		$\sum_{j=1}^{n-1} \left[\theta_j \bar{V}_j(\pi_0) + \ln Z_j(\theta_j)\right]$	B&B=Branch-and-Bound
		$\bar{V}_j(\pi_0) = \frac{1}{N} \sum_{\pi \in \text{data}} V_j(\pi \vert \pi_0)$	

ML Estimation of π_0: costs and main results

Model	Data	Log-likelihood	
Mallows	complete rankings	$\sum_{j=1}^{n-1} \bar{V}_j(\pi_0)$	[M&al07] $\pi_0{}^{ML}$ estimated exactly by B&B search.
GM^V		$\sum_{j=1}^{n-1} \left[\theta_j \bar{V}_j(\pi_0) + \ln Z_j(\theta_j)\right]$	B&B=Branch-and-Bound
		$\bar{V}_j(\pi_0) = \frac{1}{N} \sum_{\pi \in \text{data}} V_j(\pi \mid \pi_0)$	
GM^S	complete rankings top-t rankings top-t rankings, $n = \infty$	$\sum_{j=1}^{t} \left[\theta_j \bar{S}_j(\pi_0) + \ln Z_j(\theta_j)\right]$	[MBao08] Local max for $\pi_0, \vec{\theta}$ by alternate maximization and B&B search.
		$\bar{S}_j(\pi_0) = \frac{1}{N} \sum_{\pi \in \text{data}} s_j(\pi \mid \pi_0)$	

Sufficient statistics [M&al07]

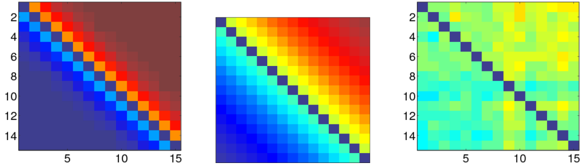
- Define $Q \equiv Q(\pi_{1:N}) = \frac{1}{N}\sum_{i=1}^{N} Q(\pi_i)$
- Sufficient statistics are sum of preference matrices for data

$Q(\pi)$

—	0	1	0
1	—	1	0
0	0	—	0
1	1	1	—

Q for large samples from Mallows models

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



Sufficient statistics [M&al07]

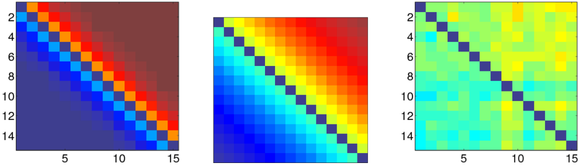
- Define $Q \equiv Q(\pi_{1:N}) = \frac{1}{N}\sum_{i=1}^{N} Q(\pi_i)$
- Sufficient statistics are sum of preference matrices for data

$Q(\pi)$

—	0	1	0
1	—	1	0
0	0	—	0
1	1	1	—

Q for large samples from Mallows models

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

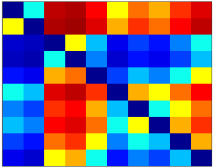


Consensus ranking $= \mathrm{argmin}_{\pi_0} L(\Pi_0^T Q \Pi_0) = \mathrm{argmin}_{\pi_0} L_{\pi_0}(Q)$

$=$ argmin lower triangle of Q over all row and column permutations π_0

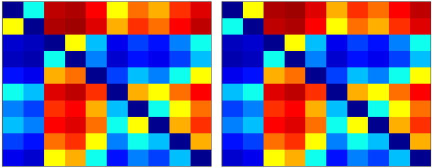
Search Algorithm Idea

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



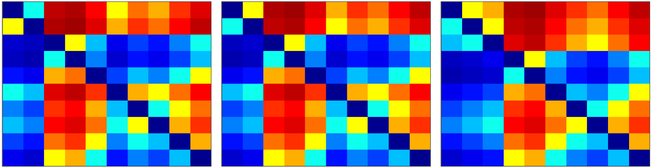
Search Algorithm Idea

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



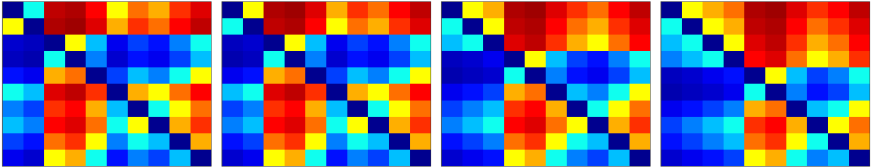
Search Algorithm Idea

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



Search Algorithm Idea

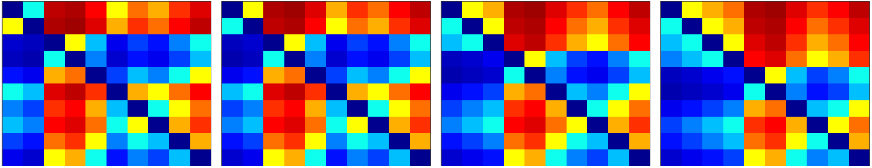
Wanted: $\text{argmin}_{\pi_0} \text{L}(\Pi_0^T \text{Q} \Pi_0) = \text{argmin}_{\pi_0} \text{L}_{\pi_0}(\text{Q})$ = argmin lower triangle of Q over all row and column permutations



...

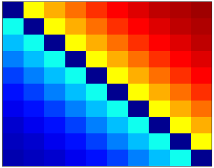
Search Algorithm Idea

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



. . .

. . .



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]

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

$$\mathcal{A} \;=\; \{Q \,|\, Q_{ik}+Q_{ki}=1,\; Q_{ik}>0\}$$

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
  $$\mathcal{A} = \{Q \mid Q_{ik} + Q_{ki} = 1,\ Q_{ik} > 0\}$$
- space of sufficient statistics = linear orderings polytope (difficult to describe [SturmfelsWelker11,Grötschel85])
  $$\mathcal{Q} = \{Q = \frac{1}{N}\sum_{1=1}^{N} Q(\pi_i)\}$$

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
$$\mathcal{A} = \{Q \mid Q_{ik} + Q_{ki} = 1,\ Q_{ik} > 0\}$$
- space of sufficient statistics = linear orderings polytope (difficult to describe [SturmfelsWelker11,Grötschel85])
$$\mathcal{Q} = \{Q = \frac{1}{N}\sum_{1=1}^{N} Q(\pi_i)\}$$
- space of means of GM model $\mathcal{M} = \{\, E_{\pi_0,\vec{\theta}}[Q]\,\}$
  - not a polytope
  - characterized algorithmically by [Mallows 57] for Mallows, [M&al07] for GMM

Consistency and rates of ML estimates

- $Q_{ij}/N \rightarrow P[\mathrm{item}\, i \prec_{\pi_0} \mathrm{item}_j]$ as $N \rightarrow \infty$ [FlignerVerducci86]
- Therefore
  - for any π_0 fixed, $\vec{\theta}^{ML}$ is consistent [FlignerVerducci86]
  - the discrete parameter ${\pi_0}^{ML}$ consistent when θ_j non-increasing [FlignerVerducci86, M–in prep]
    - is it "unbiased"?

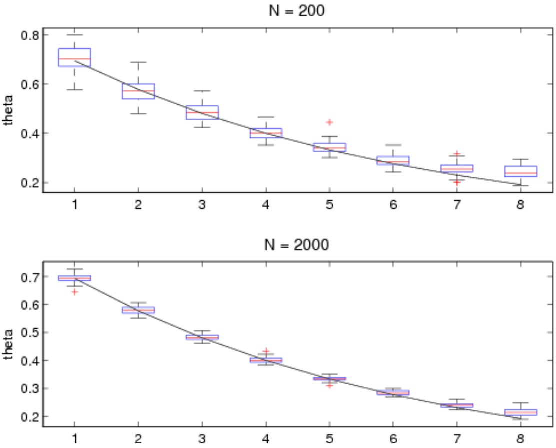
Theorem 1[M–in prep] For any N finite

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

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

The Bias of θ^{ML}

θ_j estimates for $j = 1, 8$



Outline

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

stat.columbia.edu

gsas.columbia.edu

colleges.niche.com/columbia-university/statistics

www.gocolumbialions.com/SportSelect.db..

grad-schools.usnews.rankingsandreviews.com

www.stat.sc.edu

...

- 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 $n \to \infty$
- **Observe** the top t ranks of an infinite permutation

Models for Infinite permutations

- **Domain** is countable, i.e $n \to \infty$
- **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

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

$\pi(1) = c \quad S_1 = 2$
$\pi(2) = a \quad S_2 = 0$
$\pi(3) = ? \quad S_3 = ?$

$\pi_0(1) = a \quad V_1 = 1$
$\pi_0(2) = b \quad V_2 \geq 1$
$\pi_0(3) = c \quad V_3 = 0$

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

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

sufficient statistics

no sufficient statistics

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

$Q(\pi) =$

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

The Infinite (Generalized) Mallows model (IGM)

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

- π is observed top-t ranking
- π_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 $Z(\theta_j) \;=\; 1/(1 - e^{-\theta_j})$
- $P_{\pi_0,\vec{\theta}}(\pi)$ is well defined marginal over the coset defined by π

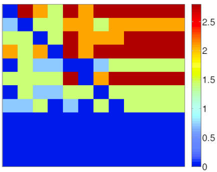
Sufficient statistics for top-t permutations [M,Bao 10]

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

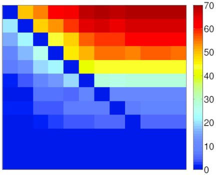
$R_j(\pi)$

	—			
		—		
$\pi(j)$	0	1	—	1
				—

$N = 2,\ t = 5$



$N = 100,\ t = 5$



$$S_j(\pi|\pi_0) \ = \ L_{\pi_0}(R_j(\pi))$$ [M,Bao 10]

Infinite GMM: ML estimation

Theorem [M,Bao 10]

- Sufficient statistics

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

- log-likelihood $l(\pi_0, \vec{\theta})$ = Sum(Lower triangle($\sum_j \theta_j R^{(j)}$) permuted by π_0) + constant

Infinite GMM: ML estimation

Theorem [M,Bao 10]

- Sufficient statistics

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

- log-likelihood $l(\pi_0, \vec{\theta})$ = Sum(Lower triangle($\sum_j \theta_j R^{(j)}$) permuted by π_0) + constant
- given π_0,

$$\theta_j{}^{ML} = \log\left(1 + N_j / L_{\pi_0}(R^{(j)})\right)$$

Infinite GMM: ML estimation

Theorem [M,Bao 10]

- Sufficient statistics

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

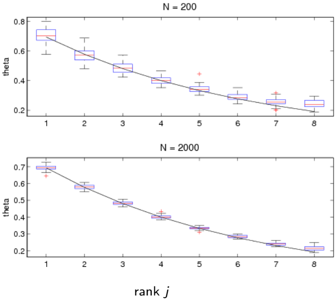
- log-likelihood $l(\pi_0, \vec{\theta}) =$ Sum(Lower triangle($\sum_j \theta_j R^{(j)}$) permuted by π_0) + constant
- given π_0,

$$\theta_j{}^{ML} \ = \ \log\left(1 + N_j / L_{\pi_0}(R^{(j)})\right)$$

- 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
  - π_0 restricted to the observed items
  - θ restricted to the observed ranks



GM are exponential family models

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
- $R_j^0 \in \mathbb{R}^{n \times n}$ equivalent sufficient statistics
  - informative prior for $\pi_0, \vec{\theta}$

Bayesian Inference: What operations are tractable?

Conjugate prior

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

Posterior

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

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

Bayesian Inference: What operations are tractable?

Conjugate prior

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

Posterior

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

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

- "Toolbox" of tractable Bayesian operations [M,Chen 10,16]
  - integrating out $\vec{\theta}$ parameters
  - sampling $\vec{\theta} \,|\, \pi_0$, $\pi_0 \,|\, \vec{\theta}$ from posterior
  - closed form posterior for $N = 1$

Bayesian Inference: What operations are tractable?

Conjugate prior

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

Posterior

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

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

- "Toolbox" of tractable Bayesian operations [M,Chen 10,16]
  Lemma 1[M,Bao 10] Posterior of π_0 and $\theta_j|\pi_0$

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

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

Bayesian Inference: What operations are tractable?

Conjugate prior

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

Posterior

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

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

- "Toolbox" of tractable Bayesian operations [M,Chen 10,16]
  Lemma 2[M, Chen 10, 16] Bayesian averaging over $\vec{\theta}$

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

Bayesian Inference: What operations are tractable?

Conjugate prior

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

Posterior

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

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

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

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

Posterior

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

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

- "Toolbox" of tractable Bayesian operations [M,Chen 10,16]
  Lemma 4 [M, Chen 10, 16] Exact sampling of $\pi_0 \mid \vec{\theta}$ from the posterior possible by stagewise sampling.

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

Bayesian Inference: What operations are tractable?

Conjugate prior

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

Posterior

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

- 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]

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

Bayesian Inference: What operations are tractable?

Conjugate prior

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

Posterior

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

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

- "Toolbox" of tractable Bayesian operations [M,Chen 10,16]
  - 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 [[Escobar,West95, 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

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

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

   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 θ_k from posterior $g_k(\theta) \propto g_0(\theta, \beta) \prod_{i \in C_k} f(x_i, \theta)$

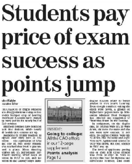
   g_k can be computed in closed form if g_0 is conjugate prior

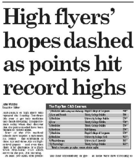
Output] a state with high posterior

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

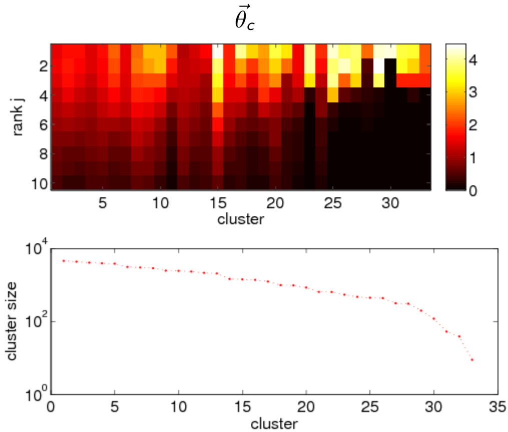

High flyers' hopes dashed as points hit record highs


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)
- [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
- $\vec{\theta}_c$ parameters large – cluster are concentrated
- number of significant ranks in σ_c, θ_c vary by cluster

College program rankings: are the clusters meaningful?

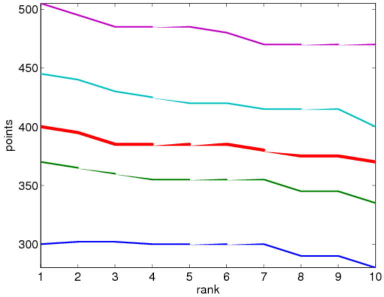
Cluster	Size	Description	Male (%)	Points avg(std)
1	4536	CS & Engineering	77.2	369 (41)
2	4340	Applied Business	48.5	366 (40)
3	4077	Arts & Social Science	13.1	384 (42)
4	3898	Engineering (Ex-Dublin)	85.2	374 (39)
5	3814	Business (Ex-Dublin)	41.8	394 (32)
6	3106	Cork Based	48.9	397 (33)
...	...	...	...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

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

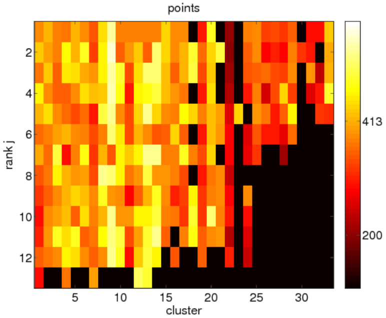
College program rankings: the "prestige" question

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

points overall (quantiles)



points for each cluster and rank

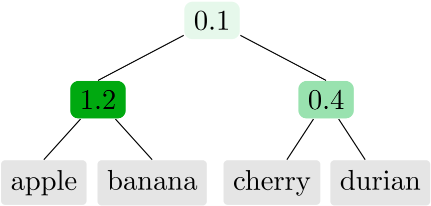


- Unclustered data: PR decreases monotonically with rankings
- Clustered data: PR not always monotonic
  - Simpson's paradox!

Outline

Recursive Inversion Models (RIM)

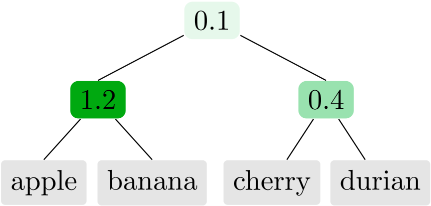
[Meek, M 14]



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

Recursive Inversion Models (RIM)

[Meek, M 14]

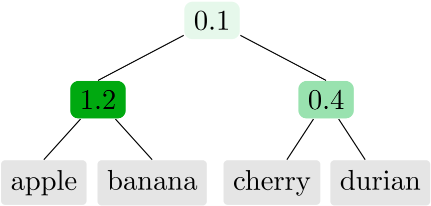


τ = tree structure
$\pi_0(\tau)$ = induced central ranking
$\theta_{1:n-1}$ = parameters at nodes
Inversions are penalized by θ_i parameters
Example: $\vec{\theta} = (0.1, 1.2, 0.4)$

$$\text{Cost}(a|b|c|d) = 0$$

Recursive Inversion Models (RIM)

[Meek, M 14]

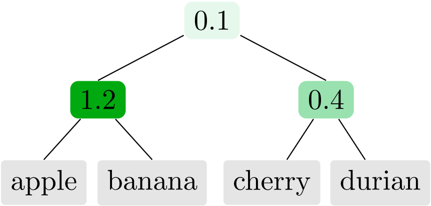


τ = tree structure
$\pi_0(\tau)$ = induced central ranking
$\theta_{1:n-1}$ = parameters at nodes
Inversions are penalized by θ_i parameters
Example: $\vec{\theta} = (0.1, 1.2, 0.4)$

$$\text{Cost}(a|b|c|d) = 0$$
$$\text{Cost}(b|a|c|d) = 1.2$$

Recursive Inversion Models (RIM)

[Meek, M 14]

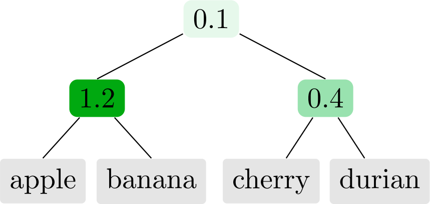


τ = tree structure
$\pi_0(\tau)$ = induced central ranking
$\theta_{1:n-1}$ = parameters at nodes
Inversions are penalized by θ_i parameters
Example: $\vec{\theta} = (0.1, 1.2, 0.4)$

$$\begin{aligned} \text{Cost}(a|b|c|d) &= 0 \\ \text{Cost}(b|a|c|d) &= 1.2 \\ \text{Cost}(c|b|a|d) &= 1.2 + 2 \times 0.1 \end{aligned}$$

Recursive Inversion Models (RIM)

[Meek, M 14]

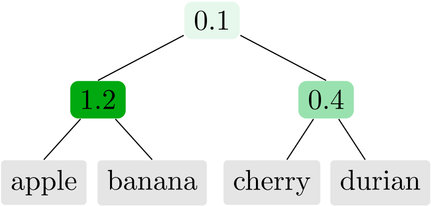


τ = tree structure
$\pi_0(\tau)$ = induced central ranking
$\theta_{1:n-1}$ = parameters at nodes
Inversions are penalized by θ_i parameters
Example: $\vec{\theta} = (0.1, 1.2, 0.4)$

$$\begin{aligned}
\text{Cost}(a|b|c|d) &= 0 \\
\text{Cost}(b|a|c|d) &= 1.2 \\
\text{Cost}(c|b|a|d) &= 1.2 + 2 \times 0.1
\end{aligned}$$

Recursive Inversion Models (RIM)

[Meek, M 14]



τ = tree structure
$\pi_0(\tau)$ = induced central ranking
$\theta_{1:n-1}$ = parameters at nodes
Inversions are penalized by θ_i parameters
Example: $\vec{\theta} = (0.1, 1.2, 0.4)$

$$\begin{aligned} \text{Cost}(a|b|c|d) &= 0 \\ \text{Cost}(b|a|c|d) &= 1.2 \\ \text{Cost}(c|b|a|d) &= 1.2 + 2 \times 0.1 \end{aligned}$$

$$\begin{aligned} P(a|b|c|d) &\propto e^0 \\ P(b|a|c|d) &\propto e^{-1.2} \\ P(c|b|a|d) &\propto e^{-1.2-2\times 0.1} \end{aligned}$$

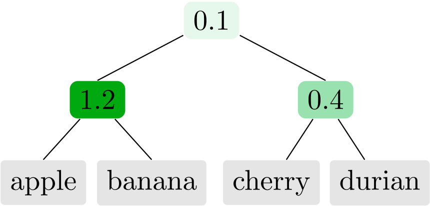
RIM distribution $P_{\tau,\vec{\theta}}$

Let v_i = number of inversions of π at node i

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

Recursive Inversion Models (RIM)

[Meek, M 14]



τ = tree structure
$\pi_0(\tau)$ = induced central ranking
$\theta_{1:n-1}$ = parameters at nodes
Inversions are penalized by θ_i parameters
Example: $\vec{\theta} = (0.1, 1.2, 0.4)$

$$\begin{aligned} \text{Cost}(a|b|c|d) &= 0 \\ \text{Cost}(b|a|c|d) &= 1.2 \\ \text{Cost}(c|b|a|d) &= 1.2 + 2 \times 0.1 \end{aligned}$$

$$\begin{aligned} P(a|b|c|d) &\propto e^{0} \\ P(b|a|c|d) &\propto e^{-1.2} \\ P(c|b|a|d) &\propto e^{-1.2-2\times 0.1} \end{aligned}$$

RIM distribution $P_{\tau,\vec{\theta}}$

Let v_i = number of inversions of π at node i

$$P_{\tau,\vec{\theta}}(\pi) \propto \prod_{i \in nodes} \exp\left(-\theta_i v_i\right)$$

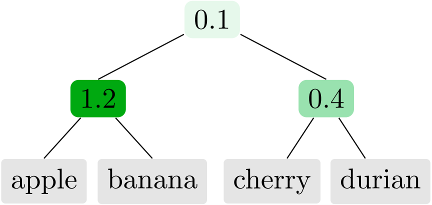
Normalization constant

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

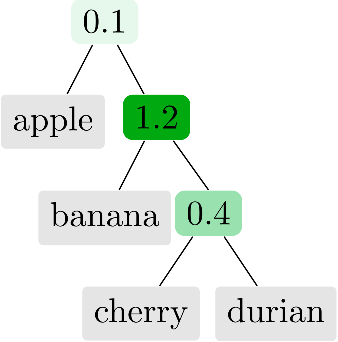
$$\text{with} \quad G(L,R,q) = \frac{(q)_{L+R}}{(q)_L (q)_L}, \quad (q)_n = \prod_{i=1}^{n}(1-q^i).$$

Structure τ 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

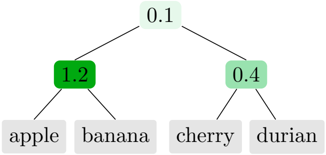
[M,Meek 14]

- Problem Given permutations $\pi_1, \ldots \pi_N$, infer τ, θ

Max Likelihood Estimation for RIM

[M,Meek 14]

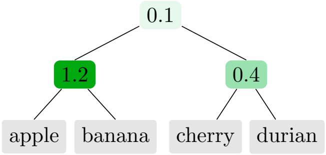
- Problem Given permutations $\pi_1, \ldots \pi_N$, infer τ, θ
- Identifiability and estimation of θ
  - reorder to obtain cannonical representation, with $\theta_i \geq 0$ for all $i \in nodes$
  - given τ, θ_i can be estimated by convex univariate minimization



Max Likelihood Estimation for RIM

[M,Meek 14]

- Problem Given permutations $\pi_1, \ldots \pi_N$, infer τ, θ
- Identifiability and estimation of θ
  - reorder to obtain cannonical representation, with $\theta_i \geq 0$ for all $i \in nodes$
  - given τ, θ_i can be estimated by convex univariate minimization



- Identifiability of τ

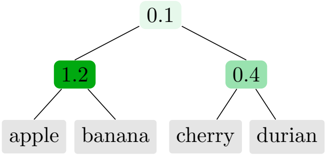
Theorem[M, Meek 14] A model τ, θ is identifiable iff

1. $\theta_i > 0$ for all $i \in nodes$
2. $\theta_i \neq \theta_{pa(i)}$ for all $i \in nodes$ ($pa(i)$ is the parent of node i in τ)

Max Likelihood Estimation for RIM

[M,Meek 14]

- Problem Given permutations $\pi_1, \ldots \pi_N$, infer τ, θ
- Identifiability and estimation of θ
  - reorder to obtain cannonical representation, with $\theta_i \geq 0$ for all $i \in nodes$
  - given τ, θ_i can be estimated by convex univariate minimization

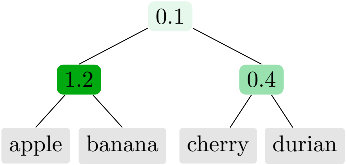


- Identifiability of τ

  Theorem[M, Meek 14] A model τ, θ is identifiable iff

  1. $\theta_i > 0$ for all $i \in nodes$
  2. $\theta_i \neq \theta_{pa(i)}$ for all $i \in nodes$ ($pa(i)$ is the parent of node i in τ)
- Hardness of τ estimation
  - Estimating π_0 is NP-hard [Duchi, Mackey, Jordan 13]
  - Estimating τ structure given π_0 is tractable

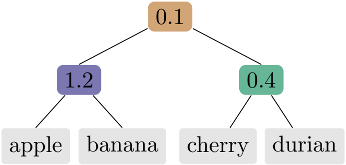
Sufficient statistics



$Q(d|a|b|c) =$

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

Sufficient statistics

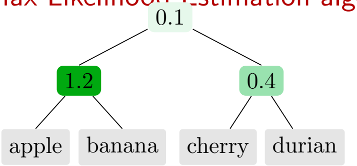


$Q(d|a|b|c) =$

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

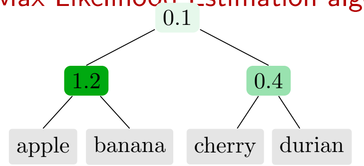
$$\text{Cost}(d|a|b|c) = 0.1 \times 2 + 1.2 \times 0 + 0.4 \times 1$$

Max Likelihood Estimation algorithm(s)



- Estimating τ given π_0 is tractable

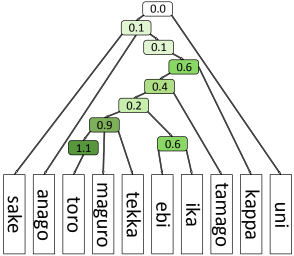
Max Likelihood Estimation algorithm(s)



- Estimating τ given π_0 is tractable
  - by Dynamic Programming (DP) algorithm, similar to Matrix Chain Multiplication, Inside(-Outside) algorithm $\mathcal{O}(n^4)$
  - contains θ_j estimation at each DP "partial solution"

- Estimating π_0: Stochastic local search over π_0 space, similar to Simulated Annealing
  1. Sample ${\pi_0}^{new}$ from proposal distribution current $P_{\tau,\theta}$
  2. Given ${\pi_0}^{new}$, find τ^{opt}, θ^{opt} by Dynamic Programming
  3. Bring to cannonical form $\Rightarrow \tau^{new}, \theta^{new} \succeq 0$
  4. Compute *log-likelihood score*, accept/reject like in Metropolis-Hastings, return to step 1

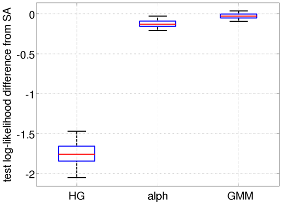
Experiments - Sushi preferences data



Data
$N = 5000$ permutations of $n = 10$ items
Compared with:

- **alph** π_0 fixed, $\tau, \theta | \pi_0$ optimize
- **GM** fixed τ, optimize π_0, θ
- **HG** fixed τ from [Huang,Guestrin,12], optimize θ
- **SA** Simulated Annealing

Test set log-likelihood w.r.t SA

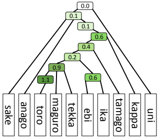


$N_{test} = 300$, $N_{train} = 4700$, 30 replicates

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"
...



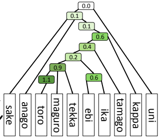
Three good things about the RIM

- RIM is a general model (includes Mallows, generalized Mallows)
- likelihood $P(\pi|\tau(\vec{\theta}))$ 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
$\underbrace{\{\text{sake,anago,}\ldots\}}_{E_1} \,|\, \underbrace{\{\text{tamago,ika,}\ldots\}}_{E_2}$



Beyond sufficient statistics – handling partial rankings

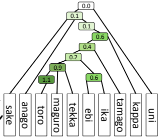
"Sushi preference" data $n = 12$

types of sushi

ika|maguro|tekka|{all other types}

{all but ebi}|ebi

$\underbrace{\{\text{sake,anago,}\ldots\}}_{E_1} \mid \underbrace{\{\text{tamago,ika,}\ldots\}}_{E_2}$



Partial ranking σ [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 $(n_1, \ldots n_K)$, $n_k = |E_k|$, $\sum n_k = n$

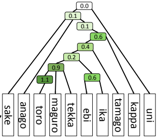
Beyond sufficient statistics – handling partial rankings

"Sushi preference" data $n = 12$ types of sushi

ika|maguro|tekka|{all other types}

{all but ebi}|ebi

$\underbrace{\{\text{sake,anago,}\ldots\}}_{E_1} \mid \underbrace{\{\text{tamago,ika,}\ldots\}}_{E_2}$



Partial ranking σ [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 $P(\pi|\tau(\vec{\theta}))$ 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 π_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 π_0 for RIM is intractable in the worst case
- We define tractable as $\mathcal{O}(N\, poly(n))\times$ time (memory) for complete data

Main technical difficulty

- marginal probability of a partial ranking σ

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

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

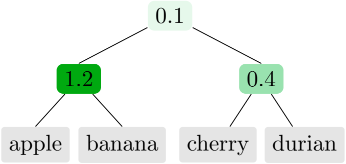
Contributions

1. for marginal probability $P(\sigma|\tau(\vec{\theta}))$
   - 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 $E[Q_{ab}] = Pr[a \text{ precedes } b \mid \sigma, \tau(\vec{\theta})]$
   - exact recursive (polynomial) algorithm
   - proved algorithm no more costly than for complete permutations
3. for parameter $\vec{\theta}$ estimation (Maximum Likelihood)
   - convex univariate minimization algorithm for each θi
   - 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 $P(\sigma|\tau, \vec{\theta})$



$$P(a|b|c|d) \propto e^{0}$$
$$P(b|a|c|d) \propto e^{-1.2}$$
$$P(c|b|a|d) \propto e^{-1.2-2\times 0.1}$$

RIM probability for complete data $P(\pi|\tau, \vec{\theta})$

(with v_i = number of inversions of π_0 at node i)

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

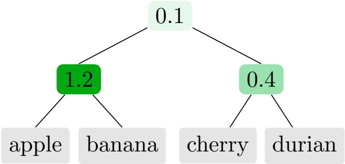
$$\text{with } G_{L,R}(q) = \frac{(q)_{L+R}}{(q)_L (q)_R}, \quad (q)_n = \prod_{i=1}^{n} (1-q^i).$$

RIM probability for partial ranking σ

[M, Meek in prep]

$$P_{\tau,\vec{\theta}}(\sigma) = \prod_{i \in nodes} (\text{factor at node } i)$$

Marginal $P(\pi|\tau,\vec{\theta})$ for partial ranking σ



Sufficient to consider root node

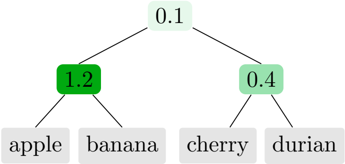
Complete ranking $\pi = (c|a|b|d)$

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

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 $P(\pi|\tau,\vec{\theta})$ for partial ranking σ



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

$$\text{factor} = \frac{e^{-2\theta} G_{0,1}(e^{-\theta}) G_{2,1}(e^{-\theta})}{G_{2,2}(e^{-\theta})}$$

In general, at some internal node where

- set $\mathcal{L}$ is merged with set $\mathcal{R}$
- partial ranking σ 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 $P(\sigma|\tau(\vec{\theta}))$ 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 σ at node $\leq \#$ inversions in $\pi \sim \sigma$

Marginal $P(\pi|\tau,\vec{\theta})$ – how much extra computation?

How many additional factors?

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

Marginal $P(\pi|\tau, \vec{\theta})$ – 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 $P(\pi|\tau,\vec{\theta})$ – 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 $\sigma = (ika|maguro|sake|\{\text{everything else}\})$ $P(\sigma|\tau,\vec{\theta})$ has at most $t-1$ non-trivial factors

Marginal $P(\pi|\tau, \vec{\theta})$ – how much extra computation?

How many additional factors?

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

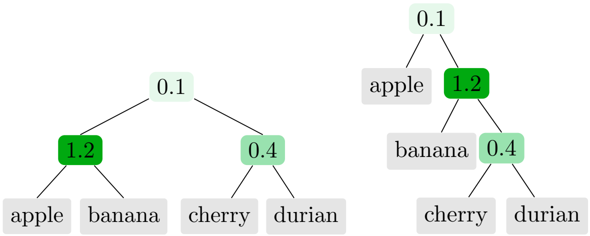
Rem 2 at each node, at least one of L_k, R_k decreases (and their initial sum is n)

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

How much additional computation?

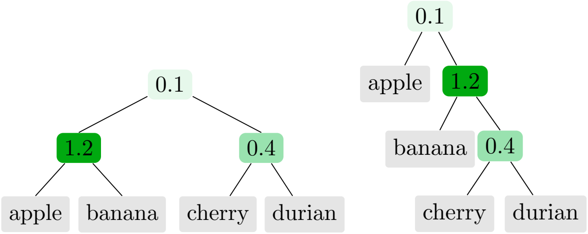
  - $G_{L,R}$ is computed recursively over $l = 0, \ldots L$, $r = 1, \ldots R$
  - Hence, all $G_{l,r}(\theta)$ in numerator are cached while computing the denominator
- Overhead for whole sample of size N is no more than nN lookups+multiplications
- For comparison, for a complete whole sample
  - computation of sufficient statistics is $\mathcal{O}(n^2 N)$
  - computation of Z given $\vec{\theta}$ 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 $\text{path}(a, b) \cap \text{path}(c, d) = \emptyset$
- Indepence checking can reveal the "branching structure" (but not π_0)
- In progress: combine independence tests with local search to estimate τ

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, identifyable, intepretable models
- Exact and tractable algorithms, closed form expression

Outline

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

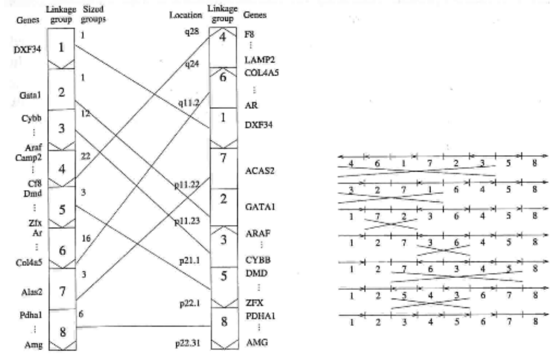


Figure 1.5: "Transformation" of a human X chromosome into a mouse X chromosome.

- These transfomations define $\mathbb{W}_n$ the hyperoctahedral group on $\{1{:}n\}$
- The elements of $\mathbb{W}_n$ are called signed permutations

Signed permutations. Three representations

- Signed permutation $\pi = [4\,\underline{2}\,\underline{1}\,3]$
- Reflected representation of π: $\pi^{ref} = [4\,\underline{2}\,\underline{1}\,3 \mid \underline{3}\,1\,2\,\underline{4}]$
- Precedence matrix $C(\pi)$

$$C_{ii'} = 1_{i \prec i'}$$

e	1	2	3	4	$\underline{4}$	$\underline{3}$	$\underline{2}$	$\underline{1}$
1	-	1	0	0	1	0	1	0
2	0	-	0	0	1	0	0	0
3	1	1	-	0	1	1	0	0
4	1	1	1	-	1	1	1	1
$\underline{4}$	0	0	0	0	–	0	0	0
$\underline{3}$	1	1	0	0	1	–	0	0
$\underline{2}$	1	1	1	0	1	1	–	1
$\underline{1}$	1	1	1	0	1	1	0	–

Group theory

hyperoctaedral group $\mathbb{W}_n$ = group of signed permutations of order n

Generators $\{\tau_1, \tau_2, \ldots \tau_{n-1}, w_n\}$ with

w_n = sign change at rank n
τ_j = elementary transposition of ranks j and $j+1$

let $\mathcal{I} = [1,\, 2,\, \ldots,\, n,\, \underline{n} \ldots \underline{2},\, \underline{1}]$

π^{ref} = permutation of $\mathcal{I}$ such that $\pi_j^{ref} = \pi_j$ and $\pi_{j+n}^{ref} = \underline{\pi_j}$ for $j \leq n$.
E.g. identity gives $\mathrm{id}^{ref} = [1 \ldots n\, \underline{n} \ldots \underline{1}]$

Inversion distance – algorithmic view

- Inversion distance $d(\pi, \pi_0)$ = # steps to bubble sort π into π_0
- $c_j(\pi|\pi_0)$ = #steps to bring item $i = \pi_0(j)$ to j'th position in π^{ref}

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

- Code of π w.r.t π_0 $c(\pi|\pi_0) = (\, c_j(\pi|\pi_0)\,)_{j=1:n}$

Example $\pi = [4\,\underline{2}\,\underline{1}\,3]$, $\pi_0 = [3\,1\,\underline{2}\,4]$

j	$\pi_0(j)$	action	current π^{ref}	c_j
			$[4\,\underline{2}\,\underline{1}\,3 \mid \underline{3}\,1\,2\,\underline{4}]$	
1	3	move 3 left 3 steps, delete $\underline{3}$	$[3\,4\,\underline{2}\,\underline{1} \mid 1\,2\,\underline{4}]$	3
2	1	move 1 left 3 steps, delete $\underline{1}$	$[3\,1\,4\,\underline{2} \mid 2\,\underline{4}]$	3
3	$\underline{2}$	move $\underline{2}$ left 1 step, delete 2	$[3\,1\,\underline{2}\,4 \mid \underline{4}]$	1
4	4	4 already in place, delete $\underline{4}$	$[3\,1\,\underline{2}\,4]$	0
				$7 = d(\pi, \pi_0)$

π_0^{ref}	3	1	$\underline{2}$	4	$\underline{4}$	2	$\underline{1}$	$\underline{3}$
1	-	1	0	0	1	1	0	1
2	0	-	0	0	1	1	1	1
3	1	1	-	0	1	1	1	1
4	1	1	1	-	1	1	1	1
$\underline{4}$	0	0	0	0	–	0	0	0
$\underline{3}$	0	0	0	0	1	–	0	0
$\underline{2}$	1	1	0	0	1	1	–	1
$\underline{1}$	0	1	0	0	1	1	0	–

Algorithm DISTANCE(π, π_0)

Represent π in reflected form π^{ref}

For $j = 1 : n$ ranks in π_0

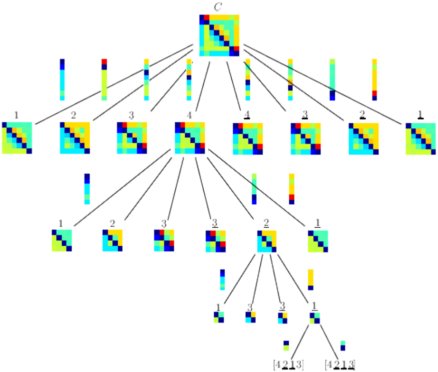
1. let $i = (\pi_0)_j$ the rank j element of π_0
2. move i left in π^{ref} to rank j by adjacent transpositions
3. delete $\underline{i}$ from the list

Ouput: $d(\pi, \pi_0)$=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 $\mathbb{W}_n$
- one can define Mallows, GM models, conjugate priors on $\mathbb{W}_n$
- sufficient statistics are (subtriangle) of precedence matrix
- estimation/consensus ranking by B&B algorithm

π_0^{ref}	3	1	$\underline{2}$	4	$\underline{4}$	2	$\underline{1}$	$\underline{3}$
1	-	1	0	0	1	1	0	1
2	0	-	0	0	1	1	1	1
3	1	1	-	0	1	1	1	1
4	1	1	1	-	1	1	1	1
$\underline{4}$	0	0	0	0	–	0	0	0
$\underline{3}$	0	0	0	0	1	–	0	0
$\underline{2}$	1	1	0	0	1	1	–	1
$\underline{1}$	0	1	0	0	1	1	0	–



A surrogate for the reversal median

- Reversal distance $r(\pi, \pi_0) = \#$ reversals to turn π into π_0
  (one reversal = several inversions)

- Reversal median problem: find π_0 minimizing

$$R(\pi_0) \;=\; \min_{\pi_0 \in \mathbb{W}_n} \sum_{k=1}^{m} r(\pi_k, \pi_0) \tag{2}$$

- 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

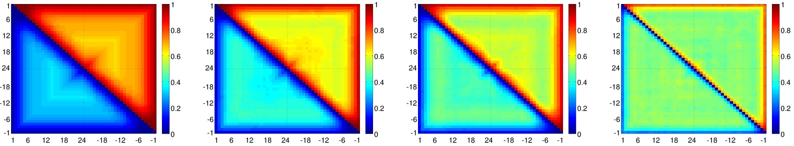
- A1 π generated by r random reversals from π_0
- A2 sample size $N \to \infty$ (asymptotic regime)
- A3 each reversal independent of previous ones
- A4 "number inversions/reversal not too large"

Theorem[M,in preparation] Under A1–4, we can show numerically that $\mathrm{argmin}_{\mathbb{W}_n} E[d(\pi, \pi$
$\mathrm{argmin}_{\mathbb{W}_n} E[r(\pi, \pi_0)]$

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.

		$n = 24$					
r	N	Objective $D(\hat{\pi}_0)$			Distance $d(\hat{\pi}_0, \pi^{true})$		
		AStar	Greedy	Rand	AStar	Greedy	Rand
1	50	125.0	125.6	370	0	1.2	135
1	100	120.8	129.0	370	0	16.5	134.7
1	1000	125.5	125.5	365	0	0	140.7
1	2000	119.1	129.9	362	0	25.2	136.9
2	50	168.8	170.1	338	0	4.4	139.3
2	100	175.4	186.1	336	0	43.3	153.4
2	1000	174.5	175.0	337	0	1.5	146.4
2	2000	171.4	182.5	340	9	47.3	149.4
3	50	203.0	205.6	325	0	15.3	143.2
3	100	198.1	206.4	330	21.1	57.1	135.7
3	1000	202.9	205.3	326	0	14.3	125.5
3	2000	201.1	210.7	324	49.4	94.5	132.6

		$n = 50$					
		ASTAR	GREEDY	RAND	ASTAR	GREEDY	RAND
1	50	372.4	383.5	1684.3	0	17.1	612
1	100	363.4	414.0	1668.8	0	77.1	636
1	1000	370.3	370.3	1674.3	0	0	627
1	2000	382.8	455.1	1699.8	0	116.7	622
2	50	601.5	619.6	1565.4	0	39.5	619
2	100	613.0	676.3	1555.7	0	147.2	623
2	1000	601.5	613.5	1557.8	0	27	596
2	2000	595.0	666.6	1536.4	0	164	619
3	50	746.6	772.8	1480.8	0	76.6	608
3	100	739.5	798.8	1485.9	0	209.4	624
3	1000	748.2	768.8	1474.7	0	64.3	633
3	2000	744.2	806.1	1480.1	0	224.1	585

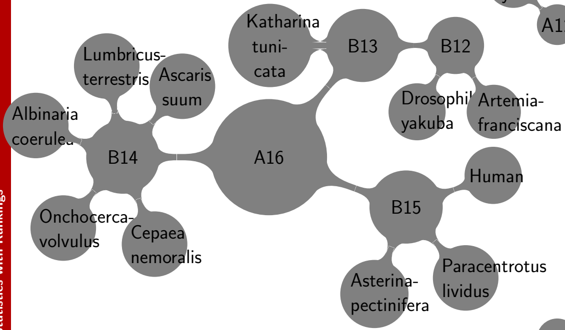
Median of (runtime ASTAR /runtime GREEDY) over 10 runs

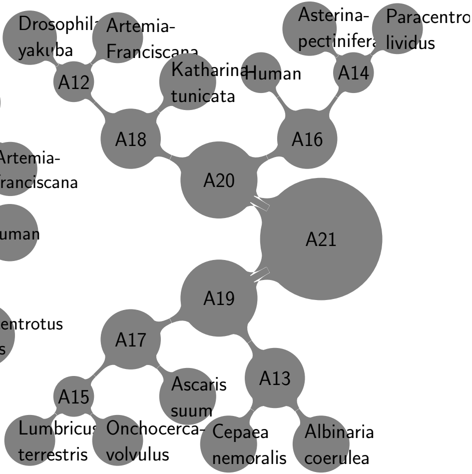
N	100	1000	100	1000	100	1000
r	1	1	2	2	3	3
$n = 50$	3.5	3.5	3.4	3.4	3.4	3.4
$n = 24$	2.25	3	3	3	5	3

Results on Metazoan mtDNA data [Bourque & Pevzner 2002]

tree built using B&B

[Bourque & Pevzner 2002] binary tree





Conclusions

Why models based on inversions?

- 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 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