Estimation and model selection in stagewise ranking: a representation story

Marina Meilă
University of Washington

NIPS AML 12/11/08

Joint with Bhushan Mandhani, Le Bao, Kapil Phadnis, Arthur Patterson, Jeff Bilmes

Overview

- Background
  - The consensus ranking problem ← An old problem
  - The code of a permutation ← The star of the show
  - The Mallows and GM Models ← Statistical formulation
- Exact algorithm for ML estimation — Theoretical solution
- Other statistical models on $\mathbb{S}_n$ — ...why Mallows?
- Extensions — Where else can it work?
  - “Model” selection

Outline

Background

- The consensus ranking problem
- The code of a permutation
- The Mallows and GM Models

Exact algorithm for ML estimation

Other statistical models on $\mathbb{S}_n$

Extensions

- "Model" selection

The Consensus Ranking problem

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 such that

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

for $d =$ distance on $\mathbb{S}_n$ the set of permutations of n objects

Relevance

- voting schemes Ireland, APA, panels
- aggregating user preferences (e.g in marketing)
- subproblem of other problems leaning to rank [Cohen, Schapire,Singer 99]

Equivalent to finding the “mean” or “median” of a set of points

The Inversion distance

Definition The Inversion distance
= the number of pairs on which π and π' disagree
= the minimum number of adjacent transpositions to turn π into π'
also called Kendall, or Kemeny distance

Example $\pi^{-1} = [\,1\,2\,3\,4\,],\ \ (\pi')^{-1} = [\,3\,1\,2\,4\,] \ \Rightarrow\ d = 2$

Fact: Consensus ranking for the inversion distance is NP hard

This talk Will make the problem even harder by phrasing it as ML estimation of a statistical model over $\mathbb{S}_n$

A decomposition for the inversion distance

$d(\pi, \mathrm{id})$ = number inversions between π and id

id = identity permutation

$$\boxed{\begin{aligned} d(\pi, \mathrm{id}) &= \#\underbrace{(\text{inversions w.r.t } 1)}_{V_1} \\ &+ \#\underbrace{(\text{ inversions w.r.t } 2)}_{V_2} \\ &+ \#\underbrace{(\text{ inversions w.r.t } 3)}_{V_3} \\ &+ \ldots \end{aligned}}$$

V_j = number inversions where j is disfavored

The code $V_{1:n-1}$

V_j = number inversions where j is disfavored

Definition $V_{1:n-1}(\pi)$ is called the **code** of permutation π

- $V_{1:n-1}(\mathrm{id}) = 0$
- $V_{1:n-1}(\pi)$ uniquely determines π

Example The code of $\pi^{-1} = [3\,5\,1\,4\,2]$

3	5	1	4	2	$V_1 = 2$
3	5	–	4	2	$V_2 = 3$
3	5	–	4	–	$V_3 = 0$
–	5	–	4	–	$V_4 = 1$
–	5	–	–	–	$V_5 = 0$

Reconstructing π from V

$$
\begin{array}{rcl}
\pi^{-1} & = & [6 \quad 1 \quad 3 \quad 5 \quad 2 \quad 4] \\
V & = & 1 \quad 3 \quad 1 \quad 2 \quad 1
\end{array}
$$

$$
\begin{array}{lcccccccl}
V_1 = 1 & \cdot & \mathbf{1} & \cdot & \cdot & \cdot & \cdot & \text{pay cost } \theta_1 V_1 \\
V_2 = 3 & \cdot & 1 & \cdot & \cdot & \mathbf{2} & \cdot & \text{pay cost } \theta_2 V_2 \\
V_3 = 1 & \cdot & 1 & \mathbf{3} & \cdot & 2 & \cdot & \text{pay cost } \theta_3 V_3 \\
V_4 = 2 & \cdot & 1 & 3 & \cdot & 2 & \mathbf{4} & \text{pay cost } \theta_4 V_4 \\
V_5 = 1 & \cdot & 1 & 3 & \mathbf{5} & 2 & 4 & \text{pay cost } \theta_5 V_5 \\
V_6 = 0 & 6 & 1 & 5 & 3 & 2 & 4 &
\end{array}
$$

A parametrized divergence between permutations

- The Inversion distance to id

$$d(\pi, \mathrm{id}) \;=\; \sum_{j=1}^{n-1} V_j(\pi)$$

- The inversion distance between π, π'

$$d(\pi, \pi') \;=\; d(\pi(\pi')^{-1}) \;=\; \sum_{j=1}^{n-1} V_j(\pi(\pi')^{-1})$$

- **Definition** Generalized Inversion “distance”

$$d_{\vec{\theta}}(\pi, \pi') \;=\; \sum_{j=1}^{n-1} \theta_j V_j(\pi(\pi')^{-1}) \qquad \theta_j \geq 0$$

The Mallows Model

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

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

- π_0 is the central permutation
  it is the unique mode of $P_{\pi_0,\theta}$ whenever $\theta > 0$
- $\theta \geq 0$ is a dispersion parameter
- for $\theta = 0$, $P_{\pi_0,0}$ is the uniform distribution over $\mathbb{S}_n$
- $P_{\pi_0,\theta}$ is a product of independent univariate distributions

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

The Generalized Mallows Model (GMM)

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

An immediate generalization $\theta \; \rightarrow \; \vec{\theta} = (\theta_1, \, \theta_2, \ldots \, \theta_{n-1})$

Definition **The generalized Mallows Model (GMM)** [Fligner, Verducci 86]

$$P_{\pi_0,\vec{\theta}}(\pi) \; = \; \frac{1}{Z_{\vec{\theta}}} \exp \left[- \sum_{j=1}^{n-1} \theta_j V_j(\pi\pi_0^{-1}) \right]$$

The estimation problem

- **Data** $\{\pi_i\}_{i=1:N}$ i.i.d. sample from $\mathbb{S}_n$
- **Model** Mallows $P_{\pi_0,\theta}$ or GMM $P_{\pi_0,\vec{\theta}}$
- **Consensus ranking problem** Set $\theta = 1$ estimate π_0.

This problem is NP hard.

- **Parameter estimation problem:** Assume π_0 known, estimate the parameter θ or $\vec{\theta}$.

This problem is easy (convex, univariate)

- **General ML estimation:** estimate both π_0 and θ or $\vec{\theta}$.

...at least as hard as consensus ranking. Will show that it's no harder

Outline

Background

- The consensus ranking problem
- The code of a permutation
- The Mallows and GM Models

Exact algorithm for ML estimation

Other statistical models on $\mathbb{S}_n$

Extensions

- "Model" selection

The likelihood

Mallows

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

Generalized Mallows

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

- Likelihood is separable and concave in each $\theta_j \Longrightarrow$ estimation of θ_j is straightforward
  - No closed form solution
  - Numerical convex minimization of $\theta_j \bar{V}_j + \ln Z_j(\theta_j)$
- For Mallows Model
  - Numerical convex minimization of $\theta \sum_{j=1}^{n-1} \bar{V}_j + \sum_{j=1}^{n-1} Z_j(\theta)$

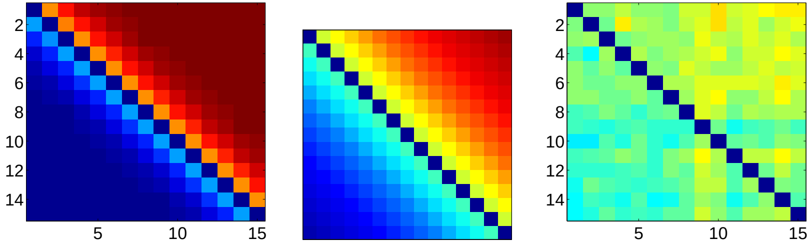
Sufficient statistics

- **Definition** Preference matrix $Q \in \mathbb{R}^{n \times n}$

$$Q_{kl} \;=\; \frac{1}{N}\sum_{i=1}^{N} 1_{[k \prec_{\pi_i} l]}$$

- Q_{kl} is the frequency of $k \prec l$ in the data
- Examples

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



Consensus Ranking: main result

- $\pi_0{}^{ML} = \operatorname*{argmin}_{\pi_0} \sum_{j=1}^{n-1} \frac{\sum_i V_j(\pi\pi_0^{-1})}{N} = \operatorname*{argmin}_{\pi_0} \sum_{j=1}^{n-1} \bar{V}_j(\pi_0)$

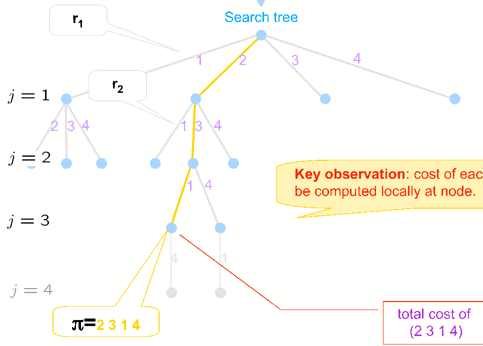
Theorem[M,Phadnis,Patterson,Bilmes 07] The optimal $\pi_0{}^{ML}$ can be found exactly by a branch-and-bound (B&B) algorithm searching on matrix Q.

- ...the search may not be tractable
- Intuition
  - The cost equals Sum (Lower triangle (Q permuted by π_0))
  - Columns of lower triangle = $\bar{V}_j(\pi_0)$

The search tree

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

Total cost of a permutation



$$\text{Cost} = \bar{V}_1(r_1)$$
$$+ \bar{V}_2(r_2|r_1)$$
$$+ \bar{V}_3(r_3|r_2, r_1)$$

Key observation: cost of each decision can be computed locally at node.

Simultaneous estimation of $\vec{\theta}$ and π_0

Cost

$$\sum_{j=1}^{n-1} \left[\theta_j \frac{\sum_i V_j(\pi_i \pi_0^{-1})}{N} + \ln Z_j(\theta_j) \right]$$

Theorem [MPPB07] The optimal ${\pi_0}^{ML}$ and $\vec{\theta}^{ML}$ can be found exactly by a B&B algorithm searching on matrix Q.

- same search tree as before
- at a node of depth j, an additional estimation of θ_j is needed (constant computational increase per node)

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

Related work

ML Estimation

[Fligner,Verducci 86] $\vec{\theta}$ estimation; heuristic for π_0

FV ALGORITHM

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

Related work (2)

Consensus Ranking $(\theta = 1)$

[CSS99] CSS $\textsc{algorithm}$ $=$ 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, 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

Is B&B practical?

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

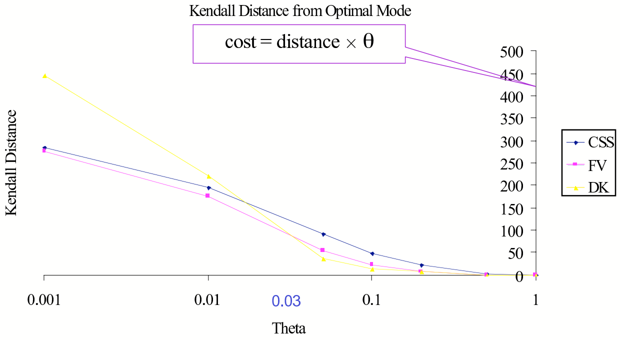
[MPPB07] admissible heuristic for Mallows Model

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

Experiments, estimate Mallows model

Data from Mallows model with $n = 100$, $N = 100$, various θ's

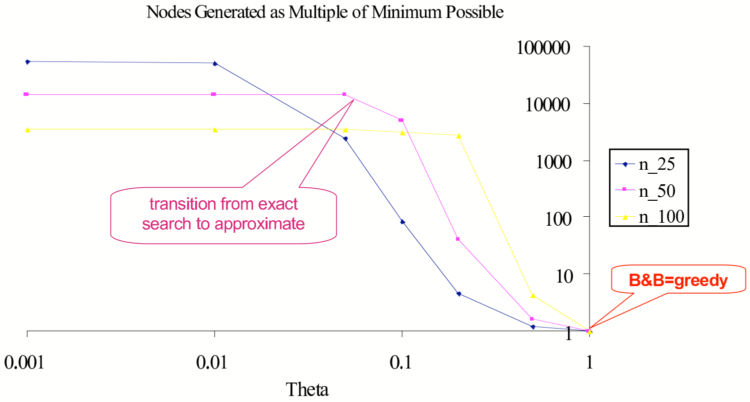
Inversion distance between B&B result and FV, DK, Random



Experiments, estimate Mallows model

Data from Mallows model with $n = 100$, $N = 100$, various θ's

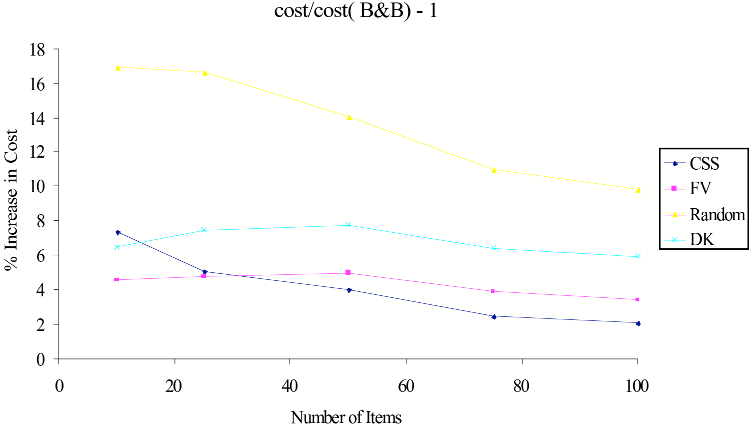
Nodes explored as a multiple of greedy search



Experiments, estimate Mallows model

Data $= Q$ with random entries in $[0,1]$, variable n

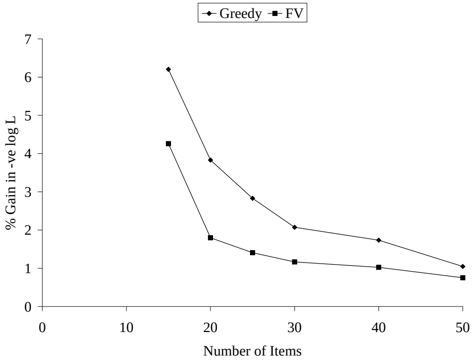
Relative improvement in cost between of B&B over the other algorithms



Experiments with GMM

Data from GMM with θ_j decreasing linearly, $N = 1000$

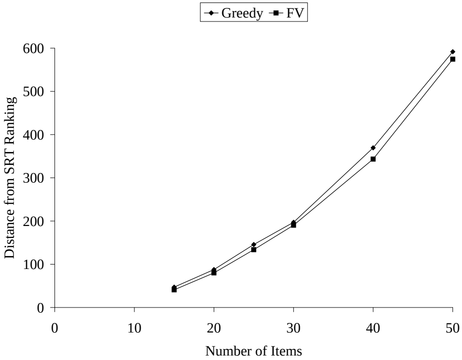
Nodes explored as multiple of minimum possible



Experiments with GMM

Data from GMM with θ_j decreasing linearly, $N = 1000$

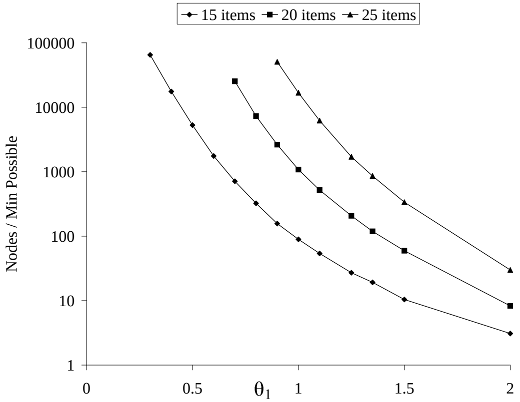
Running time



Experiments with GMM

Data from GMM with θ_j decreasing linearly, $N = 1000$

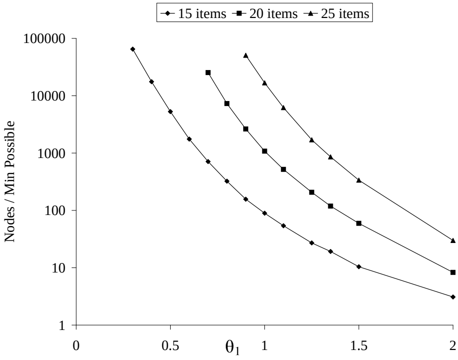
Running time



Experiments with GMM

Data from GMM with θ_j decreasing linearly, $N = 1000$

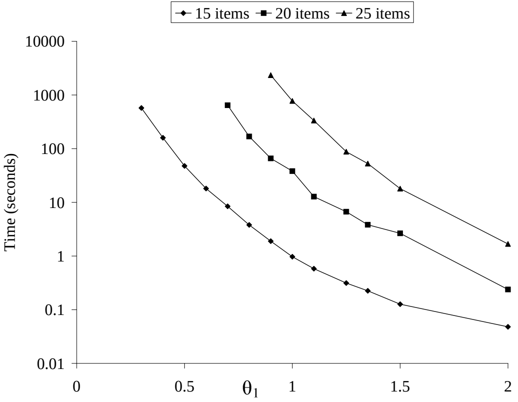
Relative improvement (%) of B&B over other algorithms



Experiments with GMM

Data from GMM with θ_j decreasing linearly, $N = 1000$

Inversion distance between B&B result and FV, DK, Random



Outline

Background
- The consensus ranking problem
- The code of a permutation
- The Mallows and GM Models

Exact algorithm for ML estimation

Other statistical models on $\mathbb{S}_n$

Extensions
- "Model" selection

Other statistical models on permutations

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*
- $P(\pi) \propto \prod_{j=1}^{n-1} \beta_{jV_j(\pi)}$ with $\beta_{j:}$ the distribution of V_j *Full model*
- $P(\pi) \propto \exp\left(-\sum_{i<j} \alpha_j 1_{[i \succ j]}\right)$ *Bradley-Terry*

$\boxed{\text{Mallows} \subset \text{GMM} \subset \text{Full} \subset \text{Bradley-Terry}}$

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

$$P([\mathrm{item}_a, \mathrm{item}_b, \ldots \mathrm{item}_n]) \propto \frac{w_a}{\sum_{i'} w_{i'}} \frac{w_b}{\sum_{i'} w_{i'} - w_a} \ldots$$

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

	GMM	Full	B-T	P-L	T
Tractable Z	yes	yes	no	no	no
"Easy" param estimation	yes	sometimes	no	no	Gauss
Tractable marginals	yes	yes	no	no	Gauss
Params "interpretable"	yes	yes	no	no	Gauss

The GM model's advantage comes from the code: the V_j's are functionally independent

Outline

Background
- The consensus ranking problem
- The code of a permutation
- The Mallows and GM Models

Exact algorithm for ML estimation

Other statistical models on $\mathbb{S}_n$

Extensions
- "Model" selection

Extensions

Can we extend the exact $\pi_0, \vec{\theta}$ estimation to other classes of problems?

- Generalized Mallows GMM ✓
- Top-t rankings
- Infinite permutations ✓
- "Model selection": what are the stages? ✓
- Signed permutations ...

Infinite permutations

- **Domain** of items to be ranked is countable, i.e $n \rightarrow \infty$
- **Observed** the top t ranks of an infinite permutation
- Examples
  - Google: UW Statistics
    www.stat.washington.edu/
    www.stat.washington.edu/www/jobs/
    www.stat.wisc.edu/
    www.washington.edu/admin/factbook/
    ...
  - searches in data bases of biological sequences (by e.g Blast, Sequest, etc)
  - open-choice polling, "grassroots elections"
- Mathematically more natural
  - for large n, models should not depend on n
  - models can be simpler, more elegant than for finite n

Definitions

Assume we have

- a countable set of items
- an infinite central ranking ${\pi_0}^{-1} = [\mathrm{item}_a, \mathrm{item}_b, \mathrm{item}_c, \ldots]$
- a top-t ranking: $\pi^{-1}(1:t) = [\mathrm{item}_1, \mathrm{item}_2, \ldots \mathrm{item}_t]$
- **Define** $\boxed{s_j + 1 = \text{rank of item}_j \text{ of } \pi \text{ in } \pi_0}$
  - relation to V_j: $\boxed{s_j(\pi) = V_j(\pi^{-1})}$
- The divergence becomes

$$d_{\vec{\theta}}(\pi, \pi_0) = \sum_{j=1}^{t} \theta_j s_j(\pi \mid \pi_0)$$

The Infinite Generalized Mallows model (IGMM)

For simplicity we assume t is fixed and the same for all observed top-t rankings.

- **Definition The Infinite GM model** [MBao08] is a distribution over top-t rankings with

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

- π_0 is a discrete infinite "location" parameter
- $\theta_{1:t} > 0$ dispersion parameter
- product of independent univariate distributions
- $P_{\pi_0,\vec{\theta}}(\pi)$ is well defined marginal over the coset defined by π
- Normalization constant $Z(\theta_j) \;=\; 1/(1 - e^{-\theta_j})$

Infinite Mallows model

Definition **The Infinite Mallows Model**

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

it is the IGMM with $\theta_1 = \theta_2 = \ldots = \theta$

Infinite Mallows: 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

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

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

Infinite GMM: ML estimation

Theorem [M,Bao 08]

- Sufficient statistics

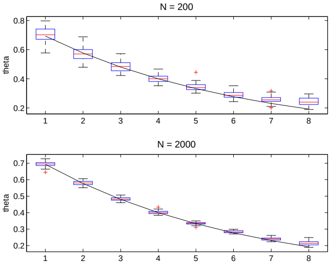
n	# distinct items observed in data
N_j	# total permutations with length $\geq j$
$Q^{(j)} = [Q^{(j)}_{kl}]_{k,l=1:n,\, j=1:t}$	frequency of $1_{[\pi(k)=j,\, \pi(l)<j]}$ in data
$q^{(j)} = [q^{(j)}_k]_{k=1:n}$	frequency of k in rank j in data
$R^{(j)} = q^{(j)}\mathbf{1}^T - Q^{(j)}$	sufficient statistics 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 $L_{\pi_0}(R) = \text{Sum}(\text{Lower triangle}(R(\vec{\theta})$ 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
  - $\vec{\theta}$ restricted to the observed ranks



- Similar result holds for finite domains

Model selection: What are the stages?

- $\theta_j V_j$ = penalty for placing j-th item, after items $1 : j-1$ are placed
- One can also define $\theta_j V_j^{reverse}$ = penalty for j-th item, after items $j+1 : n$ are placed
  The GMM model based on $V_j^{reverse}$ has similar properties to the standard GMM
- In general, given some permutation $\sigma \in \mathbb{S}_n$ one can define

$$\theta_j V_j^\sigma \;=\; \text{penalty for placing } \sigma(j) \text{ after items } \sigma_{1:j-1} \text{ are placed}$$

- σ represents the ordering of the stages
- Each σ defines a model class $\{P^\sigma_{\pi_0,\vec{\theta}} \,|\, \pi_0 \in \mathbb{S}_n, \, \vec{\theta} \in [0,\infty)^n\}$
- Can we estimate σ and π_0 from data?
  This is a "model selection" + estimation problem
- **Identifiability** Can the data distinguish between different σ's ?
- **Algorithm** Can we find an algorithm to solve the problem?

Identifiability: A few results

- Mallows $\rightarrow$ not identifiable
- Generalized Mallows
  - sometimes identifiable (Ex: $n = 3$, $\sigma = \pi_0 = \mathrm{id}$, $\theta_1 \ggg 0$, $\theta_2 = 0$)
  - sometimes not identifiable
    $Q_{ij} = 0.5$ for all j is always unidentifiable

Estimation: a few results

- V_j^σ can be defined consistently for all σ and π_0
- for general σ sufficient statistics not known
- for σ **skew-merged**
  - $[Q_{ij}]$ are sufficient statistics
  - a B&B algorithm can estimate exactly σ, π_0, θ
  - Algorithm examines both column and row sums of Q
  - strictly more complexity than the standard estimation

A skew-merged permutation:

$$\sigma = [1\,5\,4\,2,\,3] \;\Rightarrow\; 1\boxed{\boxed{\boxed{2\,\boxed{3}}\,4}\,5}$$

picks items from the "free ends" of the sequence only

Conclusions

- B&B-type algorithm
  - are theoretical solutions to estimation, consensus ranking
  - but are also practical when a mode exists
- Mallows, GMM
  - are simple models with good properties
  - well understood now
  - $\Rightarrow$ to be used as components for more realistic data generation mechanisms (mixtures, kernel density estimation, ...)
- The code grants GM it's tractability
  - because the V_j's are independent