Bayesian Mallows Model for Rank Data

장태영

서울대학교
통계학과, 베이즈통계 연구실

2021. 11. 12

CONTENTS

CONTENTS

Motivation

- Ranking and comparing items
  - Crucial for collecting information about preferences in many areas from marketing to politics
  - Netflix, Spotify, 배달의 민족, ...

Example data

```
> head(sushi_rankings)
     shrimp sea eel tuna squid sea urchin salmon roe egg fatty tuna tuna roll cucumber roll
[1,]      2       8   10     3           4           1   5           9         7              6
[2,]      1       8    6     4          10           9   3           5         7              2
[3,]      2       8    3     4           6           7  10           1         5              9
[4,]      4       7    5     6           1           2   8           3         9             10
[5,]      4      10    7     5           9           3   2           8         1              6
[6,]      4       6    2    10           7           5   1           9         8              3
```

What we want to do?

- Find consensus ranking of the items
- Extensions of model for pairwise comparisons, preference prediction and clustering.

CONTENTS

Elementary settings

- Setting : n items and N assessors. $\mathbf{R}_j \in \mathcal{P}_n$ denotes the ranking(the full set of ranks given to the n items) of assessor j for each $j = 1, \cdots, N$. ($\mathcal{P}_n$ is a permutation set)
- $d(\cdot\ ,\ \cdot) : \mathcal{P}_n \times \mathcal{P}_n \to [0, \infty)$ is a distance function between two rankings.
  - Kendall distance : number of pairs of distinct elements whose order in the two rankings are the opposite.
  - Footrule distance : ℓ_1 distance
  - Spearman's distance : ℓ_2 distance

Mallows model

- Mallows model is a class of non-uniform joint distributions for a ranking $\mathbf{r}$ on $\mathcal{P}_n$.

$$P(\mathbf{r}|\alpha, \boldsymbol{\rho}) = Z_n(\alpha, \boldsymbol{\rho})^{-1} \exp\{-\frac{\alpha}{n} d(\mathbf{r}, \boldsymbol{\rho})\} I(\mathbf{r} \in \mathcal{P}_n)$$

- $\boldsymbol{\rho} \in \mathcal{P}_n$ is the latent consensus ranking.
- $\alpha > 0$ is a scale (or precision) parameter.
  i.e. α represents the level of agreement between assesssors, so that as α gets larger, ranking $\mathbf{r}$ aggregates more to $\boldsymbol{\rho}$
- $Z_n(\alpha, \boldsymbol{\rho}) = \sum_{\mathbf{r} \in \mathcal{P}_n} e^{-\frac{\alpha}{n} d(\mathbf{r}, \boldsymbol{\rho})}$ is the partition function.
  - In physics, a 'partition function' describes the statistical properties of a system in thermodynamic equilibrium. (Source : Wikipedia)
  - Here, just consider this as a normalizing factor.

Likelihood function

- Assume that observed rankings $\mathbf{R}_1, \cdots, \mathbf{R}_N$ are conditionally independent given α and $\boldsymbol{\rho}$ and each of them is distributed according to the Mallows model with these parameters.
- Likelihood takes the form as

$$P(\mathbf{R}_1, \cdots, \mathbf{R}_N | \alpha, \boldsymbol{\rho}) = Z_n(\alpha, \boldsymbol{\rho})^{-N} \exp\{-\frac{\alpha}{n} \sum_{j=1}^{N} d(\mathbf{R}_j, \boldsymbol{\rho})\} \prod_{j=1}^{N} I(\mathbf{R}_j \in \mathcal{P}_n)$$

- For large n, finding the MLE of $\boldsymbol{\rho}$ given fixed α is not feasible because the space of permutations $\mathcal{P}_n$ has $n!$ elements.

Right-invariant distance and partition function

- For any right-invariant distance, it holds $d(\mathbf{r}_1, \mathbf{r}_2) = d(\mathbf{r}_1 \mathbf{r}_2^{-1}, \mathbf{1}_n)$ where $\mathbf{1}_n = \{1, 2, \cdots, n\}$ and $\mathbf{r}_1 \mapsto \mathbf{r}_1 \mathbf{r}_2^{-1}$ is relabelling map. Note that a right-invariant distance is unaffected by a relabelling of the items.
- Partition function $Z_n(\alpha, \boldsymbol{\rho})$ does not depend on $\boldsymbol{\rho}$.

$$\because Z_n(\alpha, \boldsymbol{\rho}) = \sum_{\mathbf{r} \in \mathcal{P}_n} \exp\{-\frac{\alpha}{n} d(\mathbf{r}, \boldsymbol{\rho})\} = \sum_{\mathbf{r} \in \mathcal{P}_n} \exp\{-\frac{\alpha}{n} d(\mathbf{r}\boldsymbol{\rho}^{-1}, \mathbf{1}_n)\}$$

$$= \sum_{\mathbf{r}' \in \mathcal{P}_n} \exp\{-\frac{\alpha}{n} d(\mathbf{r}', \mathbf{1}_n)\}$$

$$Z_n(\alpha, \boldsymbol{\rho}) = Z_n(\alpha) = \sum_{\mathbf{r} \in \mathcal{P}_n} \exp\{-\frac{\alpha}{n} d(\mathbf{r}, \mathbf{1}_n)\}$$

Right-invariant distance and partition function

- For some choice of right-invariant distance like Kendall distance, the partition function can be analytically computed.
- But there are important and natural right-invariant distances for which the computation of the partition function is not feasible, such as the footrule distance and the Spearman's distance.

Prior distributions

- Assume a priori that α and $\boldsymbol{\rho}$ are independent
- In this paper, the uniform prior $\pi(\boldsymbol{\rho}) = \frac{1}{n!} I(\boldsymbol{\rho} \in \mathcal{P}_n)$ is employed.
- Also, for the scale parameter, this paper used a truncated exponential prior with density $\pi(\alpha|\lambda) = \lambda e^{-\lambda\alpha} I(\alpha \in [0, \alpha_{max}])/(1 - e^{-\lambda\alpha_{max}})$ where the cut-off point $\alpha_{max} < \infty$ is large compared to the values supported by the data. In practice, in the computations involving the sampling of values for α, truncation was never applied. We assign λ a fixed value close to zero, implying a prior density for α which is quite flat.
  - In short, prior $\alpha \sim Exp(\frac{1}{\lambda})$ with small λ is used practically for α.

Posterior distributions

- The posterior distribution for $\boldsymbol{\rho}$ and α is given by

$$P(\boldsymbol{\rho}, \alpha | \mathbf{R}_1, \cdots, \mathbf{R}_N) \propto \frac{\pi(\boldsymbol{\rho})\pi(\alpha)}{Z_n(\alpha)^N} \exp\Big\{ -\frac{\alpha}{n} \sum_{j=1}^{N} d(\mathbf{R}_j, \boldsymbol{\rho}) \Big\} \qquad (1)$$

- The purpose of MCMC algorithm following is to obtain samples from this posterior.

Metropolis-Hastings algorithm

- A general form of the Metropolis Hastings algorithm is as follows : Target probability distribution is $p_0(x)$ for r.v. X. Given a current value $x^{(s)}$ of X,
  1. Generate x^* from a proposal distribution $J_s(x^*|x^{(s)})$
  2. Compute the acceptance ratio

$$r = \frac{p_0(x^*)}{p_0(x^{(s)})} \Big/ \frac{J_s(x^*|x^{(s)})}{J_s(x^{(s)}|x^*)} = \frac{p_0(x^*)}{p_0(x^{(s)})} \frac{J_s(x^{(s)}|x^*)}{J_s(x^*|x^{(s)})}$$

  3. set $x^{(s+1)}$ to x^* with probability $\min(1, r)$
     i.e. Sample $u \sim unif(0, 1)$ and then if $u < r$ set $x^{(s+1)} = x^*$, else set $x^{(s+1)} = x^{(s)}$

Metropolis-Hastings algorithm

- The primary restriction placed on $J_s(x^*|x^{(s)})$ is that it does not depend on values in the sequence previous to $x^{(s)}$ so that the algorithm generates a Markov chain.
- By Ergodic Thm, the empirical distribution of samples generated from such a Markov chain will converge to the stationary distribution (of the Markov chain), which agrees with the target distribution.
- Source : Hoff 2009. (textbook)

Metropolis-Hastings Algorithm for Complete Rankings

- To obtain samples from the posterior distribution (1), we alternate between two steps.
  1. Given α and $\boldsymbol{\rho}$, update $\boldsymbol{\rho}$ by proposing $\boldsymbol{\rho}'$
  2. Then, given α and $\boldsymbol{\rho}'$, update α by proposing α'

Updating ρ

- Leap-and-Shift Proposal(L&S)
- Leap step
  1. Fix an integer $L \in \{1, 2, \cdots, \lfloor \frac{n-1}{2} \rfloor\}$
     (which is a tuning parameter for MCMC algorithm)
  2. Draw a random number $u \sim Unif\{1, 2, \cdots, n\}$
  3. Define $\mathcal{S} \subset \{1, 2, \cdots, n\}$ by
     $\mathcal{S} = \big[\max(1, \rho_u - L), \min(n, \rho_u + L)\big] \setminus \{\rho_u\}$
  4. Draw a random number $r \sim Unif(\mathcal{S})$
  5. Let $\boldsymbol{\rho}^* \in \{1, 2, \cdots, n\}^n$ have elements
     $$\begin{cases} \rho_i^* = \rho_i & i \in \{1, 2, \cdots, n\} \setminus \{u\} \\ \rho_u^* = r \end{cases}$$

Updating ρ

- Shift step
  1. Let $\Delta = \rho_u^* - \rho_u$. Note that $\Delta \neq 0$
  2. Define the proposed $\boldsymbol{\rho}' \in \mathcal{P}_n$ by below :
     1. If $\Delta > 0$ then

$$\begin{cases} \rho'_u = \rho_u^* & \\ \rho'_i = \rho_i - 1 & \text{if } \rho_u < \rho_i \leq \rho_u^* \\ \rho'_i = \rho_i & \text{otherwise} \end{cases}$$

     2. If $\Delta < 0$ then

$$\begin{cases} \rho'_u = \rho_u^* & \\ \rho'_i = \rho_i + 1 & \text{if } \rho_u > \rho_i \geq \rho_u^* \\ \rho'_i = \rho_i & \text{otherwise} \end{cases}$$

Updating ρ

- Example of Leap and Shift proposal

```
> #n=8, L=3
> print(r)
[1] 4 5 8 6 3 7 1 2
> print(u)
[1] 4
> setdiff(max(r[u]-L,1):min(r[u]+L,n),r[u])
[1] 3 4 5 7 8
> print(r.star)
[1] 4 5 8 3 3 7 1 2
> print(r.prime)
[1] 5 6 8 3 4 7 1 2
```

Updating ρ

- The probability mass function associated to the transition

$$P_L(\boldsymbol{\rho}'|\boldsymbol{\rho}) = \sum_{u=1}^{n} P_L(\boldsymbol{\rho}'|U=u,\boldsymbol{\rho})P(U=u)$$

$$= \frac{1}{n}\sum_{u=1}^{n}\left\{ I_{\{\boldsymbol{\rho}_{-u}\}}(\boldsymbol{\rho}^*_{-u}) I_{\{0<|\boldsymbol{\rho}_u-\boldsymbol{\rho}^*_u|\le L\}}(\boldsymbol{\rho}^*_u)\left[\frac{I_{\{L+1,\cdots,n-L\}}(\boldsymbol{\rho}_u)}{2L}+\sum_{z=1}^{L}\frac{I_{\{z\}}(\boldsymbol{\rho}_u)+I_{\{n-z+1\}}(\boldsymbol{\rho}_u)}{L+z-1}\right]\right\}$$

$$+ \frac{1}{n}\sum_{u=1}^{n}\left\{ I_{\{\boldsymbol{\rho}_{-u}\}}(\boldsymbol{\rho}^*_{-u}) I_{\{|\boldsymbol{\rho}_u-\boldsymbol{\rho}^*_u|=1\}}(\boldsymbol{\rho}^*_u)\left[\frac{I_{\{L+1,\cdots,n-L\}}(\boldsymbol{\rho}^*_u)}{2L}+\sum_{z=1}^{L}\frac{I_{\{z\}}(\boldsymbol{\rho}^*_u)+I_{\{n-z+1\}}(\boldsymbol{\rho}^*_u)}{L+z-1}\right]\right\}$$

Updating ρ

- Simple representation for the transition probability
  - As we calculate $P(\boldsymbol{\rho}'|\boldsymbol{\rho})$, we should consider two random draws
    - Draw $u \sim Unif\{1, 2, \cdots, n\}$
    - For S dependent on $\boldsymbol{\rho}_u$, draw $r \sim Unif(S)$
    - The other works including shift step involve no randomness.
  - Simply put, $P(\boldsymbol{\rho}'|\boldsymbol{\rho}) = \frac{1}{n} \cdot \frac{1}{|S|}$ for many cases.
  - However, if $|\boldsymbol{\rho}'_u - \boldsymbol{\rho}_u| = 1$ then we should consider something more.
  - When $|\boldsymbol{\rho}'_u - \boldsymbol{\rho}_u| > 1$ then u is the only possible index that proposes $\boldsymbol{\rho}'$ from $\boldsymbol{\rho}$. On the other hand, when $|\boldsymbol{\rho}'_u - \boldsymbol{\rho}_u| = 1$, there must be only one index u' other than u s.t. $|\boldsymbol{\rho}'_{u'} - \boldsymbol{\rho}_{u'}| = 1$ so that u' can also proposes $\boldsymbol{\rho}'$ from $\boldsymbol{\rho}$.
  - In this special case, $P(\boldsymbol{\rho}'|\boldsymbol{\rho}) = \frac{1}{n} \cdot \frac{1}{|S|} + \frac{1}{n} \cdot \frac{1}{|S'|}$ where S is produced from drawing u and S' is produced from drawing u'

Updating ρ

- example of leap and shift proposal when $|\rho'_u - \rho_u| = 1$

```
> #n=8, L=3
> print(r)
[1] 7 6 2 1 4 8 5 3
> print(u)
[1] 7
> setdiff(max(r[u]-L,1):min(r[u]+L,n),r[u])
[1] 2 3 4 6 7 8
> print(r.star)
[1] 7 6 2 1 4 8 6 3
> print(r.prime)
[1] 7 5 2 1 4 8 6 3
```

Updating ρ

- Using this logic, we can rewrite the equality about $P_L(\rho'|\rho)$ as the following

$$P_L(\rho'|\rho) = \sum_{u=1}^{n} P_L(\rho'|U=u,\rho)P(U=u)$$
$$= \frac{1}{n}\sum_{u=1}^{n} I(\rho',\rho,u)\frac{1}{|S^{(u)}|}$$

where $I(\rho',\rho,u)$ is an indicator for possibility of proposal from ρ to ρ' given u is drawed and $S^{(u)}$ is the set S given u is drawed
If ρ' is proposed from ρ then typically $I(\rho',\rho,u)=1$ for only one u but if $|\rho'_u - \rho_u| = 1$ then $I(\rho',\rho,u')=1$ also holds for another u' different from u

Updating ρ

- The acceptance probability when updating $\boldsymbol{\rho}$ is $\min\{1, r\}$ where r is given as

$$
\begin{aligned}
r &= \frac{P(\boldsymbol{\rho}', \alpha|\mathbf{R})}{P(\boldsymbol{\rho}, \alpha|\mathbf{R})} \cdot \frac{P_L(\boldsymbol{\rho}|\boldsymbol{\rho}')}{P_L(\boldsymbol{\rho}'|\boldsymbol{\rho})} \\
&= \frac{P_L(\boldsymbol{\rho}|\boldsymbol{\rho}')}{P_L(\boldsymbol{\rho}'|\boldsymbol{\rho})} \cdot \frac{\pi(\boldsymbol{\rho}')}{\pi(\boldsymbol{\rho})} \exp\Big\{ -\frac{\alpha}{n} \sum_{j=1}^{N} \big[d(\mathbf{R}_j, \boldsymbol{\rho}') - d(\mathbf{R}_j, \boldsymbol{\rho}) \big] \Big\}
\end{aligned}
$$

- Leap and shift proposal is not a symmetric proposal distribution.
- The term $\sum_{j=1}^{N} \big[d(\mathbf{R}_j, \boldsymbol{\rho}') - d(\mathbf{R}_j, \boldsymbol{\rho}) \big]$ above can be computed efficiently since most elements of $\boldsymbol{\rho}$ and $\boldsymbol{\rho}'$ are equal and we can put aside indices i s.t. $\rho_i = \rho_i'$

Updating α

- Sample a proposal α' from a lognormal distribution $\log \mathcal{N}(\log(\alpha), \sigma_\alpha^2)$
  σ_α^2 is a tuning parameter for MCMC algorithm.
- Note that $X \sim \log \mathcal{N}(\mu, \sigma^2) \Leftrightarrow Y = \log X \sim N(\mu, \sigma^2)$
  The pdf of $X \sim \log \mathcal{N}(\log(\mu), \sigma^2)$ is written as

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(\log x - \log \mu)^2\Big) \frac{1}{x} I(x > 0)$$

- The probability density function associated to the transition is

$$J(\alpha'|\alpha) = \frac{1}{\sqrt{2\pi\sigma_\alpha^2}} \exp\Big(-\frac{1}{2\sigma_\alpha^2}(\log \alpha' - \log \alpha)^2\Big) \frac{1}{\alpha'}$$

  Accordingly, we have the ratio $\frac{J(\alpha'|\alpha)}{J(\alpha|\alpha')} = \frac{\alpha}{\alpha'}$

Updating α

- Acceptance probability is $\min\{1, r\}$ where r is given as

$$
\begin{aligned}
r &= \frac{P(\boldsymbol{\rho}, \alpha' | \mathbf{R})}{P(\boldsymbol{\rho}, \alpha | \mathbf{R})} \Big/ \frac{J(\alpha' | \alpha)}{J(\alpha | \alpha')} \\
&= \frac{\alpha'}{\alpha} \frac{\pi(\alpha')}{\pi(\alpha)} \frac{Z_n(\alpha)^N}{Z_n(\alpha')^N} \exp\Big\{ - \frac{\alpha' - \alpha}{n} \sum_{j=1}^{N} d(\mathbf{R}_j, \boldsymbol{\rho}) \Big\}
\end{aligned}
$$

- Additional parameter α_{jump} can be used to update α only every α_{jump} updates of $\boldsymbol{\rho}$

CONTENTS

Motivation for approximating $Z_n(\alpha)$

- Notice that MCMC algorithm to obtain samples from the posterior distribution, we need to know the value of $Z_n(\alpha)$ which appears in acceptance probability for updating α
- The partition function $Z_n(\alpha)$ is available in close form for Kendall's, Hamming, and Cayley distances.
- But this is not the case for footrule and Spearman distances.

$Z_n(\alpha)$ unavailable for some useful distances

- $Z_n(\alpha) = \sum_{\mathbf{r} \in \mathcal{P}_n} \exp\{-\frac{\alpha}{n} d(\mathbf{r}, \mathbf{1}_n)\}$
- Note that $d(\mathbf{r}, \mathbf{1}_n)$ takes only the finite number of discrte values $\mathcal{D} = \{d_1, \cdots, d_a\}$ where a depends on n and distance $d(\cdot, \cdot)$
- It can be rewritten as

$$Z_n(\alpha) = \sum_{d_i \in \mathcal{D}} |L_i| \exp\{-\frac{\alpha}{n} d_i\}$$

  where $L_i = \{\mathbf{r} \in \mathcal{P}_n : d(\mathbf{r}, \mathbf{1}_n) = d_i\}$
- To compute $Z_n(\alpha)$ we only need $|L_i|$ for all values $d_i \in \mathcal{D}$

$Z_n(\alpha)$ unavailable for some useful distances

- In the case of footrule distance
  - $\mathcal{D}$ includes all even numbers from 0 to $\lfloor n^2/2 \rfloor$
  - $|L_i|$ corresponds to the sequence A061869 available for $n \leq 50$ on the OEIS(Online Encyclopedia of Integer Sequences)
- In the case of Spearman's distance
  - $\mathcal{D}$ includes all even numbers from 0 to $2\binom{n}{3}$
  - $|L_i|$ corresponds to the sequence A175929 available for $n \leq 14$ on the OEIS
- What about the case where n is large?

Approximation of $Z_n(\alpha)$

- To handle these cases, we propose an approximation of the partition function $Z_n(\alpha)$ based on importance sampling.
- Recall that given right-invarant distances, the partition function does not depend on $\boldsymbol{\rho}$.

Off-line approximation

- Obtain an off-line approximation of the partition function on a grid of α values.
  - In computer science, an online algorithm is one that can process its input piece-by-piece in a serial fashion, i.e. in the order that the input is fed to the algorithm, without having the entire input available from the start.
  - In contrast, an offline algorithm is given the whole problem data from the beginning and is required to output an answer which solves the problem at hand.
    (Source : Wikipedia Online Algorithm)
- Then interpolate it to yield an estimate of $Z_n(\alpha)$ over a continuous range and read off needed values to compute the acceptance probabilities rapidly.

Importance sampling

- Suppose our goal is estimate $\mu = E_p[f(X)]$
  i.e. the expected value of $f(X)$ under $X \sim p$
- For a probability density q other than p, we can calculate that

$$\mu = E_p[f(X)] = \int f(x)p(x)\,dx$$
$$= \int \frac{f(x)p(x)}{q(x)} q(x)\,dx = E_q\left[\frac{f(X)p(X)}{q(X)}\right]$$

  i.e. μ equals the expected value of $\frac{f(X)p(X)}{q(X)}$ under $X \sim q$
- The importance sampling estimate of μ is

$$\hat{\mu}_q = \frac{1}{K}\sum_{k=1}^{K} \frac{f(X_i)p(X_i)}{q(X_i)} \qquad \textit{where } X_i \sim q$$

Importance sampling

- The basic idea of importance sampling is to sample the states from a different distribution to lower the variance of estimation of μ or when sampling from original density p is difficult.
- Reference : Wikipedia and Lecture note from standford.edu

Approximate $Z_n(\alpha)$ using IS approach

- For K rank vectors $\mathbf{R}^1, \cdots, \mathbf{R}^K$ sampled from an IS auxiliary distributution $q(\mathbf{R})$, the unbiased IS estimate of $Z_n(\alpha)$ is given by

$$\hat{Z}_n(\alpha) = \frac{1}{K} \sum_{k=1}^{K} \exp\left\{ -\frac{\alpha}{n} d(\mathbf{R}^k, \mathbf{1}_n) \right\} \frac{1}{q(\mathbf{R}^k)} \tag{2}$$

Approximate $Z_n(\alpha)$ using IS approach

- IS estimate of $Z_n(\alpha)$ in (2) is derived as the following

$$Z_n(\alpha) = \sum_{\mathbf{R}\in\mathcal{P}_n} \exp\{-\frac{\alpha}{n} d(\mathbf{R}, \mathbf{1}_n)\} = \sum_{\mathbf{R}\in\mathcal{P}_n} \frac{1}{P(\mathbf{R})} \exp\{-\frac{\alpha}{n} d(\mathbf{R}, \mathbf{1}_n)\} P(\mathbf{R})$$

$$= E_{R\sim P(R)}\Big[\frac{1}{P(\mathbf{R})} \exp\{-\frac{\alpha}{n} d(\mathbf{R}, \mathbf{1}_n)\}\Big] = E_{R\sim P(R)}\big[f(\mathbf{R})\big]$$

where $f(\mathbf{R}) = \frac{1}{P(\mathbf{R})} \exp\{-\frac{\alpha}{n} d(\mathbf{R}, \mathbf{1}_n)\}$
and $P(\mathbf{R})$ is abbreviation of $P(\mathbf{R}|\alpha, \mathbf{1}_n)$

$$\hat{Z}_n(\alpha) = \frac{1}{K} \sum_{k=1}^{K} \frac{f(\mathbf{R}^k) P(\mathbf{R}^k)}{q(\mathbf{R}^k)} = \frac{1}{K} \sum_{k=1}^{K} \exp\Big\{-\frac{\alpha}{n} d(\mathbf{R}^k, \mathbf{1}_n)\Big\} \frac{1}{q(\mathbf{R}^k)}$$

where $\mathbf{R}^k \sim q(\mathbf{R})$ for each $k = 1, \cdots, K$

Approximate $Z_n(\alpha)$ using IS approach

- We shall use the following psuedo-likelihood approximation for $q(\mathbf{R})$
- While we cannot sample $\mathbf{R}$ from $P(\mathbf{R}|\alpha, \mathbf{1}_n)$ ($\because$ we don't know the value of $Z_n(\alpha)$) it must be computationally feasible to sample $\mathbf{R}$ from $q(\mathbf{R})$.
- The more $q(\mathbf{R})$ resembles the Mallows likelihood $P(\mathbf{R}|\alpha, \mathbf{1}_n)$, the smaller is the variance of $\hat{Z}_n(\alpha)$.

Psuedo-likelihood approximation for $q(\mathbf{R})$

1. Sample $(i_1, \cdots, i_n) \in \mathcal{P}_n$, which gives the order of the psuedo-lieklihood factorization.
2. Factorization is given as

$$P(\mathbf{R}|\mathbf{1}_n) = P(R_{i_n}|\mathbf{1}_n)P(R_{i_{n-1}}|R_{i_n}, \mathbf{1}_n) \cdots P(R_{i_2}|R_{i_3}, \cdots, R_{i_n}, \mathbf{1}_n) \\ P(R_{i_1}|R_{i_2}, \cdots, R_{i_n}, \mathbf{1}_n)$$

Psuedo-likelihood approximation for $q(\mathbf{R})$

3 The conditional distributions are given by

$$P\left(R_{i_n}|\mathbf{1}_n\right) = \frac{\exp\left\{-(\alpha/n)d\left(R_{i_n}, i_n\right)\right\} \cdot 1_{[1,\ldots,n]}(R_{i_n})}{\sum_{r_n\in\{1,\ldots,n\}} \exp\left\{-(\alpha/n)d\left(r_n, i_n\right)\right\}},$$

$$P\left(R_{i_{n-1}}|R_{i_n}, \mathbf{1}_n\right) = \frac{\exp\left\{-(\alpha/n)d\left(R_{i_{n-1}}, i_{n-1}\right)\right\} \cdot 1_{[\{1,\ldots,n\}\setminus\{R_{i_n}\}]}(R_{i_{n-1}})}{\sum_{r_{n-1}\in\{1,\ldots,n\}\setminus\{R_{i_n}\}} \exp\left\{-(\alpha/n)d\left(r_{n-1}, i_{n-1}\right)\right\}},$$

$$\vdots$$

$$P\left(R_{i_2}|R_{i_3}, \ldots, R_{i_n}, \mathbf{1}_n\right) = \frac{\exp\left\{-(\alpha/n)d\left(R_{i_2}, i_2\right)\right\} \cdot 1_{\left[\{1,\ldots,n\}\setminus\{R_{i_3},\ldots,R_{i_n}\}\right]}(R_{i_2})}{\sum_{r_2\in\{1,\ldots,n\}\setminus\{R_{i_3},\ldots,R_{i_n}\}} \exp\left\{-(\alpha/n)d\left(r_2, i_2\right)\right\}},$$

$$P\left(R_{i_1}|R_{i_2}, \ldots, R_{i_n}, \mathbf{1}_n\right) = 1_{\left[\{1,\ldots,n\}\setminus\{R_{i_2},\ldots,R_{i_n}\}\right]}(R_{i_1}).$$

Each factor is a simple univariate distribution.

Psuedo-likelihood approximation for $q(\mathbf{R})$

4 For given value of α, sample R_{i_n} first, and then conditionally on that, $R_{i_{n-1}}$ and so on. The k-th full sample $\mathbf{R}^k$ has probability

$$q(\mathbf{R}^k) = P(R^k_{i_n}|\mathbf{1}_n)P(R^k_{i_{n-1}}|R^k_{i_n}, \mathbf{1}_n)\cdots P(R^k_{i_2}|R^k_{i_3}, \cdots, R^k_{i_n}, \mathbf{1}_n)$$

5 Iterate this process K times to calculate $\hat{Z}_n(\alpha)$ as in (2)

- $\surd$ Keeping the psuedo-likelihood with the same distance as the one in the target was most accurate and efficient so we shall use the distance in (2) as same as the distance in (1).

Estimate of $Z_n(\alpha)$ over a continuous range

- Over a discrete grid of 100 equally spaced α values between 0.01 and 10 (this is the range of α which turned out to be relevant in all our applications, typically $\alpha < 5$), we produce a smooth partition function simply using a polynomial of degree 10.
- What we have is 100 data points of $(\alpha^{(i)}, \hat{Z}_n(\alpha^{(i)}))$'s. A smooth partition function is produced by fitting multiple linear regression for the model

$$\log \hat{Z}_n(\alpha) = \beta_0 + \beta_1 \alpha + \beta_2 \alpha^2 + \cdots + \beta_{10} \alpha^{10}$$

so that only thing we should store before implementing MCMC for the partition function is those estimated beta parameter values.

Effect of approximation on the MCMC

- Theoretical results about the convergence of the MCMC when using the IS approximation of the partition function should be given.
- Algorithm using $\hat{Z}_n$ instead of Z_n converges to the posterior distribution proportional to

$$\frac{1}{\hat{C}(\mathbf{R})} \frac{\pi(\boldsymbol{\rho})\pi(\alpha)}{\hat{Z}_n(\alpha)^N} \exp\Big\{ -\frac{\alpha}{n} \sum_{j=1}^{N} d(\mathbf{R}_j, \boldsymbol{\rho}) \Big\}$$

where the normalizing factor
$\hat{C}(\mathbf{R}) = \int \sum_{\boldsymbol{\rho} \in \mathcal{P}_n} \frac{\pi(\boldsymbol{\rho})\pi(\alpha)}{\hat{Z}_n(\alpha)^N} \exp\Big\{ -\frac{\alpha}{n} \sum_{j=1}^{N} d(\mathbf{R}_j, \boldsymbol{\rho}) \Big\} d\alpha$

Effect of approximation on the MCMC

- The approximate posterior above converges to the exact posterior (1), if $K = K(N)$ increases with N and

$$\lim_{N\to\infty} \left(\frac{\hat{Z}_n^{K(N)}(\alpha)}{Z_n(\alpha)}\right)^N = 1 \quad \forall\, \alpha$$

- For this, it is sufficient that $K(N)$ grows faster than $c \cdot N^2$ where c depends on $\alpha, n, d(\ ,\ )$

Studying the effect of approximation by simulations

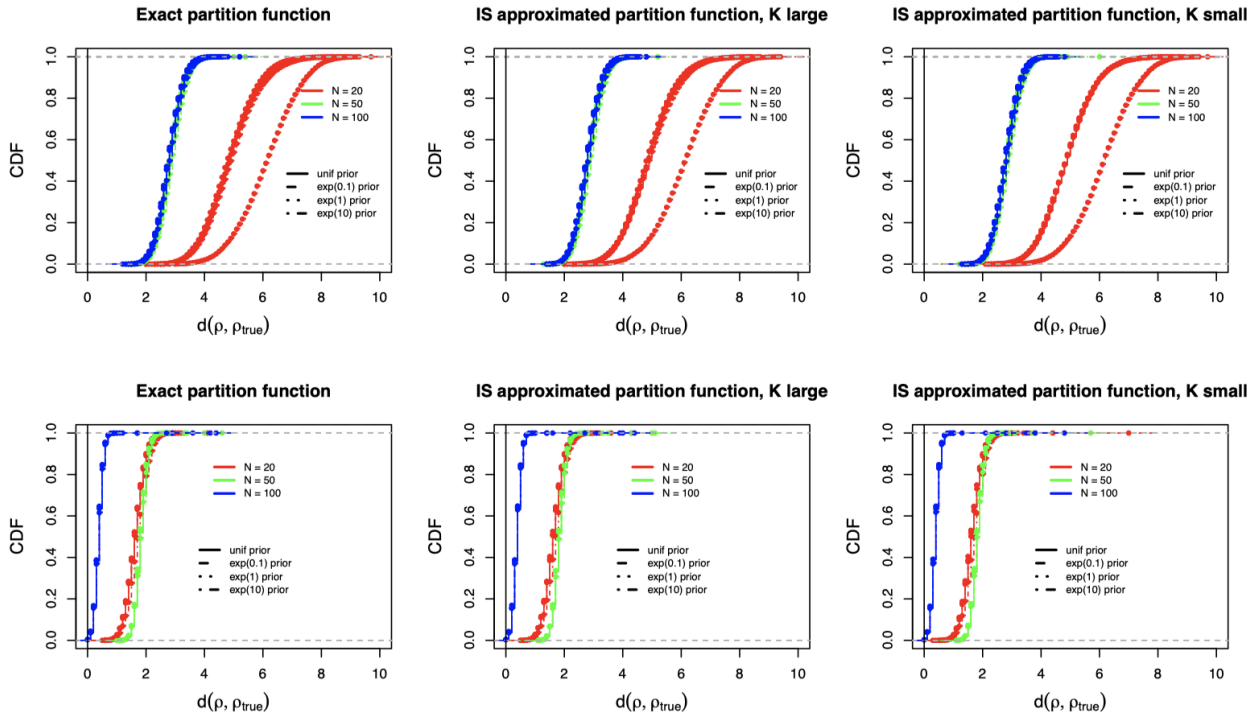


Figure 3: Results of the simulations described in Section 3.3, when $n = 20$. In each plot, posterior CDF of $d(\boldsymbol{\rho}, \boldsymbol{\rho}_{\text{true}})$ obtained for various choices of N (different colors), and for different choices of the prior for α (different line types), as stated in the legend. From left to right, MCMC run with the exact $Z_n(\alpha)$, with the IS approximation $\hat{Z}_n^K(\alpha)$ with $K = 10^8$, and with the IS approximation $\hat{Z}_n^K(\alpha)$ with $K = 10^4$. First row: $\alpha_{\text{true}} = 1$; second row: $\alpha_{\text{true}} = 3$.

Studying the effect of approximation by simulations

- Consider the performance of the method when using the IS approximation $\hat{Z}_n(\alpha)$with $K = 10^4$ and 10^8 then comparing the results with those based on the exact $Z_n(\alpha)$
- The precision and the accuracy of the marginal posterior distriutions increasing for $\boldsymbol{\rho}$ with N becoming larger.
- For smaller values of α_{true}, $\boldsymbol{\rho}$ is stochastically farther from ρ_{true}. These results are stable against varying choices of the prior for α
- Most importantly, inference on $\boldsymbol{\rho}$ is completely unaffected by the approximation of $Z_n(\alpha)$ already when $K = 10^4$
- Similar results is still yielded when n becomes larger as 50 or 100

Studying the effect of approximation by simulations

- The main postive result from the perspective of practical applications
  1. The relative lack of sensitivity of the posterior inferences to the specification of the prior for the scale parameter α
  2. The apprarent robustness of the marginal posterior inferences on $\boldsymbol{\rho}$ on the choice of the approximation of the partition function $Z_n(\alpha)$

CONTENTS

Assumptions that we will relax in this section

- We will relax two assumptions of the previous sections.
  1. Each assessor ranks all n items.
  2. The assessors are homogeneous, all sharing a common consensus ranking.

Top-k ranks

- Often only a subset of the items is ranked.
- These situations can be handled conveninetly in Bayesian framework by applying data augmentation techniques.
- We shall consider the case of the top-k ranks.

Setting for top-k ranks case

- Among n items $\{A_1, \cdots, A_n\}$, each assessor j has ranked the subset of items $\mathcal{A}_j \subset \{A_1, \cdots, A_n\}$ giving them top ranks from 1 to $n_j = |\mathcal{A}_j|$.
- Before, we had complete ranking $\mathbf{R}_j \in \mathcal{P}_n$, but now, we denote $\mathbf{R}_j$ as partial ranking.
- We have augmented ranking vectors $\tilde{\mathbf{R}}_j \in \mathcal{P}_n$ where unknown part follows a symmetric prior on the permutations of $(n_j + 1, \cdots, n)$ for each $j = 1, \cdots, N$

MCMC Algorithms for top-k ranks case

- $\mathcal{S}_j$: set of all possible augmented ranking vectors given original partially ranked items together with the allowable 'fill-ins' of the missing ranks, for each $j = 1, \cdots, N$
- Our goal is to sample from the posterior distribution

$$P(\alpha, \boldsymbol{\rho} \,|\, \mathbf{R}_1, \cdots, \mathbf{R}_N) = \sum_{\tilde{\mathbf{R}}_1 \in \mathcal{S}_1} \cdots \sum_{\tilde{\mathbf{R}}_N \in \mathcal{S}_N} P(\alpha, \boldsymbol{\rho}, \tilde{\mathbf{R}}_1, \cdots, \tilde{\mathbf{R}}_N \,|\, \mathbf{R}_1, \cdots, \mathbf{R}_N)$$

- Our MCMC algorithm alternates between
  1. sampling the augmented ranks given the current values of α and $\boldsymbol{\rho}$
  2. sampling α and $\boldsymbol{\rho}$ given the current values of the augmented ranks.
- The latter is done similar as in Section 2, where in this case $\mathbf{R}_1, \cdots, \mathbf{R}_N$ is replaced by $\tilde{\mathbf{R}}_1, \cdots, \tilde{\mathbf{R}}_N$

MCMC Algorithms for top-k ranks case

- For the former, given the current $\tilde{\mathbf{R}}_j$ (which embeds info contained in $\mathbf{R}_j$) and the current values of α and $\boldsymbol{\rho}$, $\tilde{\mathbf{R}}'_j$ is sampled in $\mathcal{S}_j$ from a uniform proposal distribution which is obviously symmetric.
  The proposed $\tilde{\mathbf{R}}'_j$ is accepted with probability $\min\{1, r\}$ with

$$
\begin{aligned}
r &= \frac{P(\tilde{\mathbf{R}}_1, \cdots, \tilde{\mathbf{R}}'_j, \cdots, \tilde{\mathbf{R}}_N \,|\, \alpha, \boldsymbol{\rho})}{P(\tilde{\mathbf{R}}_1, \cdots, \tilde{\mathbf{R}}_j, \cdots, \tilde{\mathbf{R}}_N \,|\, \alpha, \boldsymbol{\rho})} \\
&= \exp\left[-\frac{\alpha}{n}\left\{d(\tilde{\mathbf{R}}'_j, \boldsymbol{\rho}) - d(\tilde{\mathbf{R}}_j, \boldsymbol{\rho})\right\}\right]
\end{aligned}
$$

- Note that we can generalize this algorithm to generic partial ranking, where items partially ranked by each assessor are not necessarily the top ranked items.

Effects of Unranked Items on the Top-k Consensus Ranking

- It is possible that the number of items is large and there are items which none of the assessors included in their top-list.
- Can we ignore such 'left-over' items and consider only the items explicitly ranked by at least one assessor?
- The two main points are that
  - Only items explicitly ranked by the assessors appear in top positions of the consesnsus ranking.
  - When considering the MAP(maximum a posteriori) consensus ranking, excluding the left-over items from the ranking procedure already at the start has no effect on how the remaining ones will appear in such consensus ranking.

Effects of Unranked Items on the Top-k Consensus Ranking

Proposition 4 *Consider two latent consensus rank vectors* $\boldsymbol{\rho}$ *and* $\boldsymbol{\rho}'$ *such that*

(i) in the ranking $\boldsymbol{\rho}$ *all items in* $\mathcal{A}$ *have been included among the top-n-ranked, while those in* $\mathcal{A}^c$ *have been assigned ranks between* $n+1$ *and* n^*,

(ii) $\boldsymbol{\rho}'$ *is obtained from* $\boldsymbol{\rho}$ *by a permutation, where the rank in* $\boldsymbol{\rho}$ *of at least one item belonging to* $\mathcal{A}$ *has been transposed with the rank of an item in* $\mathcal{A}^c$.

Then, $P_{n^*}(\boldsymbol{\rho}|data) \geq P_{n^*}(\boldsymbol{\rho}'|data)$, *for the footrule, Kendall and Spearman distances in the* full analysis *mode.*

Corollary 2 *Denote by* $\boldsymbol{\rho}^{MAP*}$ *the MAP estimate for consensus ranking obtained in a* full analysis, $\boldsymbol{\rho}^{MAP*} := \text{argmax}_{\boldsymbol{\rho}\in\mathcal{P}_{n^*}} P_{n^*}(\boldsymbol{\rho}|data)$, *and by* $\boldsymbol{\rho}^{MAP}$ *the MAP estimate for consensus ranking obtained in a* restricted analysis, $\boldsymbol{\rho}^{MAP} := \text{argmax}_{\boldsymbol{\rho}\in\mathcal{P}_n} P_n(\boldsymbol{\rho}|data)$. *Then,* $\boldsymbol{\rho}^{MAP*}|_{i:A_i\in\mathcal{A}} \equiv \boldsymbol{\rho}^{MAP}$.

Effects of Unranked Items on the Top-k Consensus Ranking

- The above proposition says that the MAP estimate for consensus ranking assigns n highest ranks to explicitly ranked items in the data (Be aware that here we denote the number of total items as n^*)
- Note that full analysis, which includes the complete set of all items, cannot always be carried out in practice due to the fact that left-over items might be unknown or too many for realistic computation. The corollary guarantees that the top-n items in the MAP consensus ranking do not depend on whether we include left-over items in the analysis.

Pairwise Comparison

- Often, assessors compare pairs of items rather than ranking all or a subset of items.
- Notation for pairwise comparison
  - $A_r \prec A_s$: A_s is preferred to A_r, so that A_s has a higher rank than A_r
  - $\mathcal{B}_j$: pairwise orderings or preferences stated by assessor j
  - $\mathcal{A}_j$: set of items constrained by asessor j
  - $tc(\mathcal{B}_j)$: the transitive closure of $\mathcal{B}_j$, containing all pairwise orderings of the elements in $\mathcal{A}_j$ induced by $\mathcal{B}_j$.

$$
\begin{aligned}
\mathcal{B}_j &= \{A_1 \prec A_2, A_2 \prec A_5\} \\
&\Rightarrow tc(\mathcal{B}_j) = \{A_1 \prec A_2, A_2 \prec A_5, A_1 \prec A_5\} \\
\mathcal{B}_k &= \{A_1 \prec A_2, A_2 \prec A_5, A_4 \prec A_5\} \\
&\Rightarrow tc(\mathcal{B}_k) = \{A_1 \prec A_2, A_2 \prec A_5, A_1 \prec A_5, A_4 \prec A_5\}
\end{aligned}
$$

MCMC Algorithms for pairwise comparison

- In the MCMC algorithm, we need to propose augmented ranks which obey the partial ordering constraints given by each assessor, to avoid a large number of rejections, with the difficulty that none of the items is now fixed to a given rank.
- We can also handle the case when assessors give ties : in such a situation, each pair of items resulting in a ties is randomized to a preference at each data augmentation step inside the MCMC.

MCMC Algorithms for pairwise comparison

- The main idea of MCMC algorithm remains the same as the one for the top-k ranks
- The difference is that here, a 'modified' leap-and-shift proposal distribution, rather than a uniform proposal distribution, is used to sample augmented ranks.

Modified Leap-and-Shift proposal

- Only leap step is modified and the shift step remains unchanged.
- Given a full augmented rank vector $\tilde{\mathbf{R}}_j$ compatible with $tc(\mathcal{B}_j)$, we shall propose $\tilde{\mathbf{R}}'_j$
  1. Draw a random number $u \sim Unif\{1, 2, \cdots, n\}$
  2. If $A_u \notin \mathcal{A}_j$ then complete the leap step by drawing $\tilde{R}^*_{uj} \sim Unif\{1, 2, \cdots, n\}$
  3. If $A_u \in \mathcal{A}_j$ then complete the leap step by drawing $\tilde{R}^*_{uj} \sim Unif\{l_j + 1, \cdots, r_j - 1\}$ where l_j and r_j are defined by
     - $l_j = \max\{\tilde{R}_{kj} : A_k \in \mathcal{A}_j, k \neq u, (A_k \succ A_u) \in tc(\mathcal{B}_j)\}$ with convention that $l_j = 0$ if the set is empty
     - $r_j = \min\{\tilde{R}_{kj} : A_k \in \mathcal{A}_j, k \neq u, (A_k \prec A_u) \in tc(\mathcal{B}_j)\}$ with convention that $r_j = n + 1$ if the set is empty
     - Briefly, l_j is given rank of the item whose rank is closest to A_u among all assessed items preferred to A_u, and r_j is given rank of the item whose rank is closest to A_u among all assessed items less preferred than A_u
- Note that this modified leap-and-shift is symmetric proposal. Hence we use the same acceptance probability as in the top-k ranks case.

Motivation for clustering

- So far we have assumed that there exists a unique consensus ranking shared by all assessors.
- The possibility of dividing assessors into more homogeneous subsets, each sharing a consesnsus ranking of the items, brings the model closer to reality.
- We introduce a mixture of Mallows models to handle heterogeneity.

Mixture of Mallows models

- Assume that the data consist of complete rankings.
- $z_j \in \{1, \cdots, C\}$ assigns assessor j to one of C clusters for each $j = 1, \cdots, N$ i.e. $z_1, \cdots, z_N$ are cluster labels.
- The assessments $\mathbf{R}$ within each cluster $c \in \{1, \cdots, C\}$ are described by a Mallows model with parameters α_c and $\boldsymbol{\rho}_c$ which is the cluster consensus.
- Assume conditional independence across the clusters.
- Likelihood for the observed rankings $\mathbf{R}_1, \cdots, \mathbf{R}_N$ is given by

$$P(\mathbf{R}_1, \cdots, \mathbf{R}_N | \{\alpha_c, \boldsymbol{\rho}_c\}_{c=1,\cdots,C}, z_1, \cdots, z_N)$$
$$= \prod_{j=1}^{N} \frac{1}{Z_n(\alpha_{z_j})} \exp\{-\frac{\alpha_{z_j}}{n} d(\mathbf{R}_j, \boldsymbol{\rho}_{z_j})\}$$

Mixture of Mallows models

- Assumption for priors
  1. $\boldsymbol{\rho}_1, \cdots \boldsymbol{\rho}_C \overset{indep}{\sim} \pi_{\boldsymbol{\rho}}$ where $\pi_{\boldsymbol{\rho}}$ is a uniform prior on $\mathcal{P}_n$ as before.
  2. $\alpha_1, \cdots, \alpha_C \overset{indep}{\sim} \pi_\alpha$ where π_α is a truncated exponential prior with shared λ
  3. τ_c is the probability that an assessor belongs to the c-th cluster. $\tau_c \geq 0 \quad \forall c = 1, \cdots, C$ and $\sum_{c=1}^{C} \tau_c = 1$. $(\tau_1, \cdots \tau_C)$ are assigned the standard symmetric Dirichlet prior $\mathcal{D}(\psi, \cdots, \psi)$
  4. $P(z_j = c \,|\, \tau_1, \cdots, \tau_C) = \tau_c \quad \forall c = 1, \cdots C$ and $z_1, \cdots z_N$ are conditionally i.i.d.
- λ and ψ are tuning parameters. What about C?

How to determine the number of clusters

- The number of clusters C is often unknown, and the selection of C can be based on different criteria.
  1. The within cluster sum of distances $\sum_{c=1}^{C} \sum_{j:z_j=c} d(\tilde{\mathbf{R}}_j, \boldsymbol{\rho}_c)$
  2. The within-cluster indicator of mis-fit to the data $\sum_{c=1}^{C} \sum_{j:z_j=c} \left|\{B \in tc(\mathcal{B}_j) : \text{B is not consistent with } \boldsymbol{\rho}_c\}\right|$ (valid for pairwise comparison case)
  - $\surd$ While the former only depends on MCMC outputs $\tilde{\mathbf{R}}_j$ and $\boldsymbol{\rho}_c$, the latter partly depends on the observed data $\mathcal{B}_j$

How to determine the number of clusters

- Here we use the posterior distribution of the within-cluster sum of distances of the observed ranks from the corresponding cluster consensus.
- Separate analyses were performed for $C = 1, 2, \cdots, \mathcal{C}$ for some $\mathcal{C}$
- We expect to observe an 'elbow' in the within-cluster distance posterior distribution as a function of C, identifying the optimal number of clusters.

MCMC Algorithm for Mixture Mallows model

- The algorithm alternates between
  1. sampling $\boldsymbol{\rho}_1, \cdots, \boldsymbol{\rho}_C$ and $\alpha_1, \cdots, \alpha_C$ in a Metropolis-Hastings step
  2. sampling $\tau_1, \cdots, \tau_C$ and $z_1, \cdots, z_N$ in a Gibbs sampler step.
- The former is straightforward. Update is proceeded element-wisely and the acceptance probability is slightly changed according to the cluster index $c \in \{1, \cdots C\}$

Setting for top-k ranks case

- For the latter
  1. Gibbs step for $(\tau_1, \cdots, \tau_C)$
     Dirichlet prior is conjugate to the multinomial conditional prior.
     Since $(\tau_1, \cdots \tau_C) \sim \mathcal{D}(\psi, \cdots, \psi)$ &
     $(n_1, \cdots, n_C)|(\tau_1, \cdots, \tau_C) \sim Multi(N, (\tau_1, \cdots, \tau_C))$ where
     $n_c = \sum_{j=1}^{N} I(z_j = c)$ for each $c = 1, \cdots, C$, we sample $(\tau_1, \cdots, \tau_C)$
     from $\mathcal{D}(\psi + n_1, \cdots, \psi + n_C)$ in the Gibbs step.
  2. Gibbs step for $(z_1, \cdots, z_N)$
     We sample z_j from $P(z_j = c \,|\, \tau, \boldsymbol{\rho}, \alpha, \mathbf{R}_j) \quad \forall c = 1, \cdots C$ for each
     $j = 1, \cdots N$ where $\tau, \boldsymbol{\rho}, \alpha$ are C-dim vectors.

$$
\begin{aligned}
P(z_j = c \,|\, \tau, \boldsymbol{\rho}, \alpha, \mathbf{R}_j) &\propto P(z_j = c \,|\, \tau) P(\mathbf{R}_j \,|\, \boldsymbol{\rho}, \alpha, z_j = c) \\
&\quad \because \text{prior} * \text{likelihood} \\
&= P(z_j = c \,|\, \tau) P(\mathbf{R}_j \,|\, \boldsymbol{\rho}_c, \alpha_c) \\
&= \tau_c Z_n(\alpha_c)^{-1} \exp\left\{ -\frac{\alpha_c}{n} d(\mathbf{R}_j, \boldsymbol{\rho}_c) \right\}
\end{aligned}
$$

Remark for the extension models

- Merging two algorithms in this section, we can treat situations where incomplete ranking data are observed and assessors must be divided into separate clusters.

CONTENTS

Benefit of Bayesian Approach

- Estimation for consensus ranking by MAP estimator
- In applications, the interest often lies in computing posterior probabilities of more complex functions of the consensus ρ.
  - (Ex) The posterior probability that a certain item has consensus rank higher than a given level ("among the top 5", say)
  - (Ex) The posterior probability that the consensus rank of a certain item is higher than the consensus rank of another one.
- Bayesian approach naturally allows to estimate any posterior summary of interest by means of MCMC.

Benefit of Bayesian Approach - Preference Prediction

- One such thing is a preference prediction
- Situation : assessors have been asked to respond to some queries containing different sets of pairwise comparisons. One may then ask how the assesssors would have ranked for pairwise comparisons when such comparison could not be concluded directly from the data they provided.

Benefit of Bayesian Approach - Preference Prediction

- For example, suppose assessor j did not compare A_1 to A_2. We might be interested in computing $P(A_1 \prec_j A_2 \,|\, data)$, the predictive probability that this assessor would have preferred item A_2 to item A_1. This probability is then readily obtained from the MCMC output as a marginal of the posterior $P(\tilde{\mathbf{R}}_j \,|\, data)$
  i.e. If we have 10^5 MCMC posterior outputs for $\tilde{\mathbf{R}}_j$ then compute the ratio of the number of outputs satisfying $A_1 \prec_j A_2$ to the number of total outputs, 10^5.
- This type of problems is called as preference learning or personalized ranking, which is a step towards personalized recommendation.

Example : Sushi Data

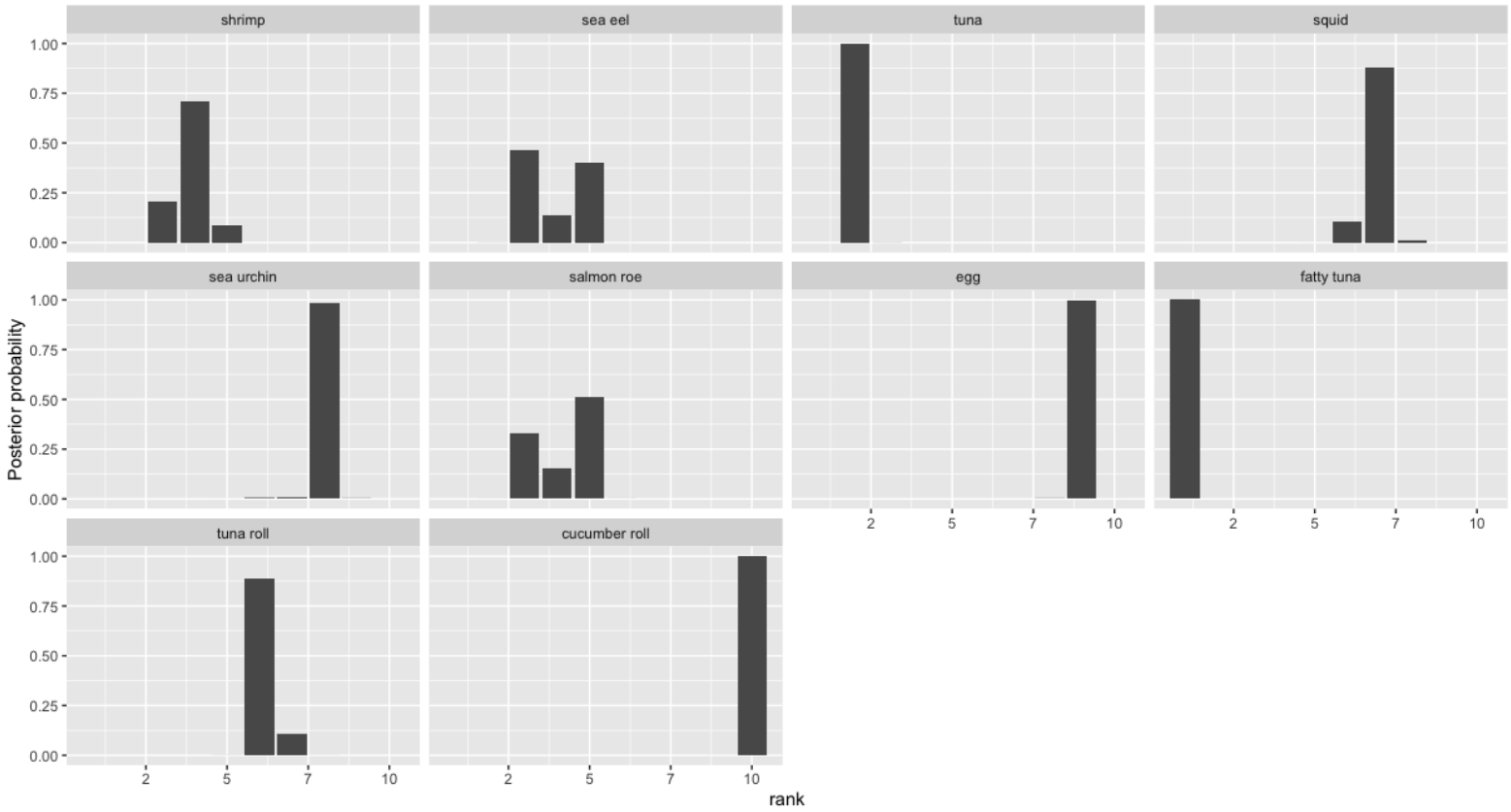
- $N = 5000$ people were interviewed, each giving a complete ranking of $n = 10$ sushi variants.

```
› head(sushi_rankings)
     shrimp sea eel tuna squid sea urchin salmon roe egg fatty tuna tuna roll cucumber roll
[1,]      2       8   10     3           4           1   5           9         7             6
[2,]      1       8    6     4          10           9   3           5         7             2
[3,]      2       8    3     4           6           7  10           1         5             9
[4,]      4       7    5     6           1           2   8           3         9            10
[5,]      4      10    7     5           9           3   2           8         1             6
[6,]      4       6    2    10           7           5   1           9         8             3
```

- We want to figure out the consensus ranking of sushi.
- For convenience, use only a subset of 250 people.

Example : Sushi Data

- Consensus ranking for sushi is estimated as
  1. fatty tuna 2. tuna 3. sea eel 4.shrimp 5. salmon roe
  6. tuna roll 7. squid 8. sea urchin 9. egg 10. cucumber roll

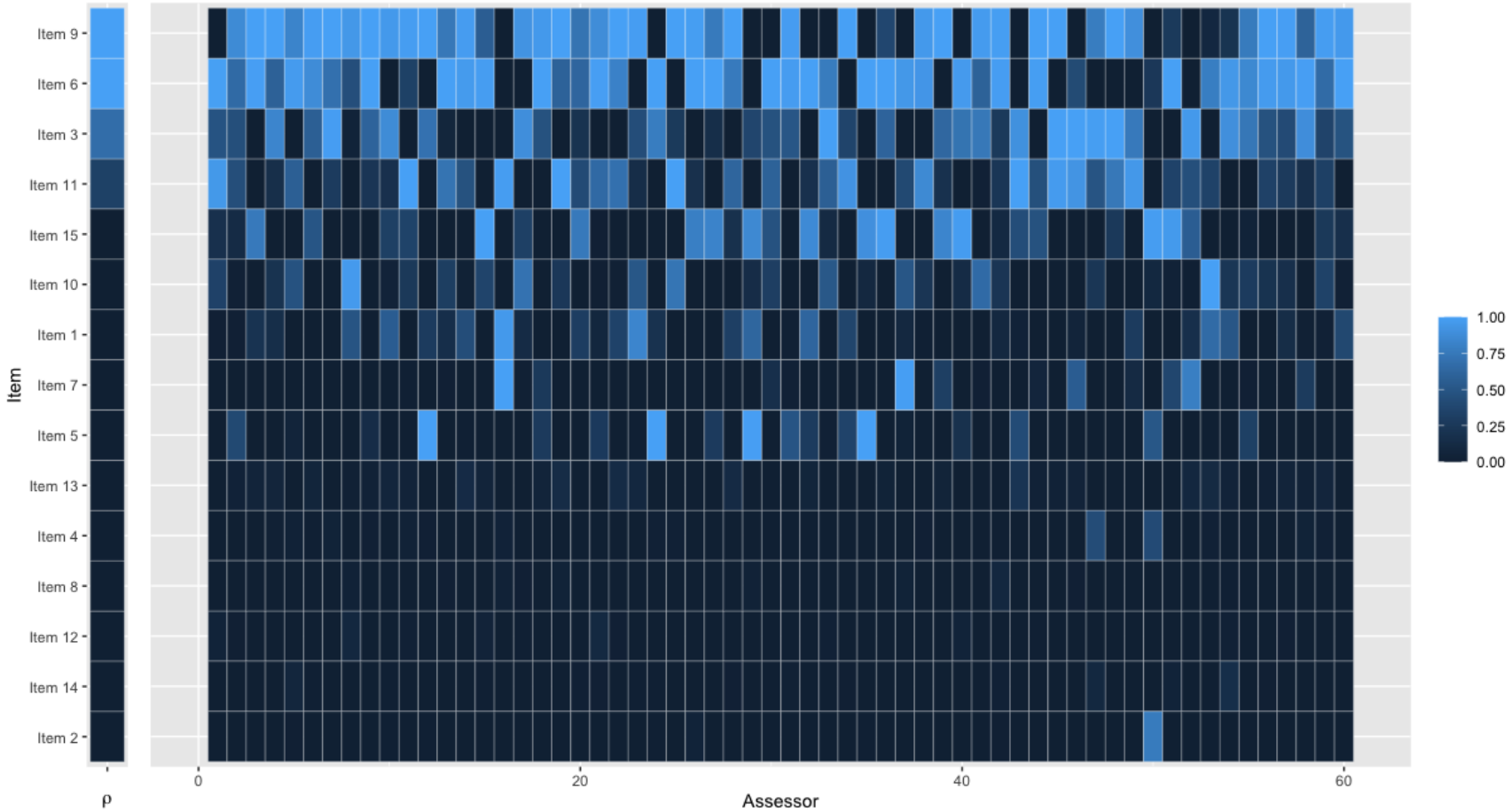


Example : Beach Preference Data

- This is the case in the pairwise comparison
- There are $n = 15$ images of tropical beaches s.t. they differ in terms of presence of building and people.
- Each assessor answers for comparing a random set of 25 pairs of images. $N = 60$ answers are collected.
- Nine assessors returned orderings which contained at least one non-transitive pattern of comparisons. (This refers to the case like $A_1 \prec A_2$, $A_2 \prec A_3$ but $A_3 \prec A_1$).
- In this analysis we dropped the non-transitive patterns from the data. Systematic methods for dealing with non-transitive rank data will be considered in another article.

Example : Beach Preference Data

- We want to get consensus ranking for beaches and also prediction for the top-3 beaches for each individual.



Discussion

- The Mallows model performs very well with a large number of assessors N
- But it may not be computationally feasible when the number of items is extremely large, for example $n \geq 10^4$, which is not uncommon in certain applications. MCMC algorithm converges slowly in such large spaces.
- There are many situations where rankings vary over time. We assume to observe ranks at discrete time-points indexed by $t = 0, 1, \cdots, T$ and let $\boldsymbol{\rho}^{(t)}$ and $\alpha^{(t)}$ denote the parameters of the Mallows model at time t.
- A natural generalization of our model is to allow for item-specific α's. The Mallows model with footrule and Spearman distance has not yet been generalized to handle item specific α's mostly due to the obvious computational difficulties. Within our famework, this appears as feasible.

Reference

Valeria Vitelli, Øystein Sørensen, Marta Crispino, Arnoldo Frigessi, and Elja Arjas. Probabilistic preference learning with the mallows rank model. 2017.