# Bayesian methods for rank and preference data: from recommender systems to cancer genomics

Valeria Vitelli

Oslo Centre for Biostatistics and Epidemiology (OCBE),
Department of Biostatistics, University of Oslo, Norway,
valeria.vitelli@medisin.uio.no


UiO : University of Oslo

People involved:
Øystein Sørensen, Sylvia Liu, Manuela Zucknick
Thomas Fleischer, Vessela Kristensen, Ida Scheel,
Elja Arjas, and Arnoldo Frigessi

**Nordic Probabilistic AI School. NTNU, Trondheim, June 5, 2019**

# Outline

# Key collaborators

**Original method:**

Øystein Sørensen,
LCBC – UiO

Arnoldo Frigessi,
OCBE – UiO & BigInsight

Elja Arjas,
OCBE – UiO & Univ. Of Helsinki

**Extensions:**

Qinghua Liu,
math - UiO

Derbachew Asfaw,
Univ. of Hawassa, Ethiopia

Marta Crispino,
INRIA, Grenoble, France

**Cancer Genomics:**

Manuela Zucknick,
OCBE - UiO

Vessela Kristensen, UiO & OUS

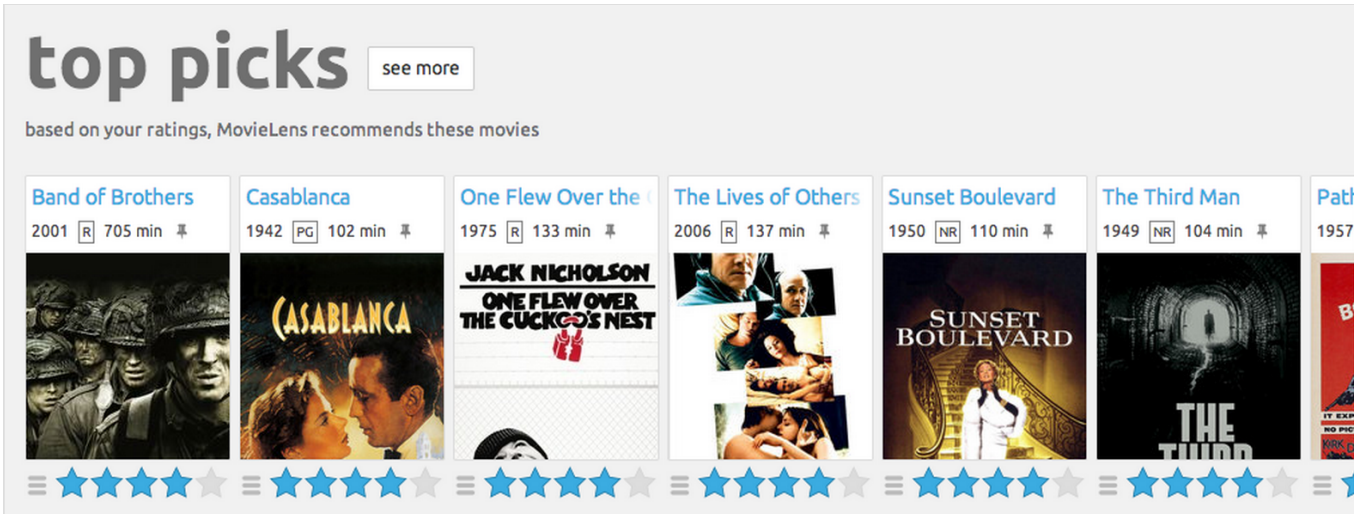
Thomas Fleischer, OUS

# Why preference learning matters?

- customers express preferences about products and services;

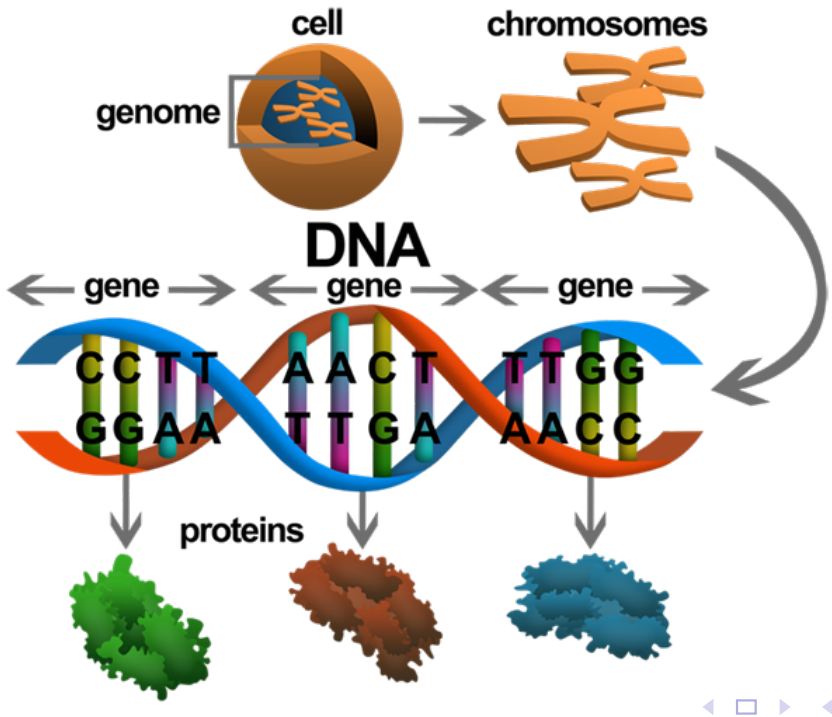
# Why preference learning matters?

- customers express preferences about products and services;
- users choose movies on an internet platform (e.g., Netflix);



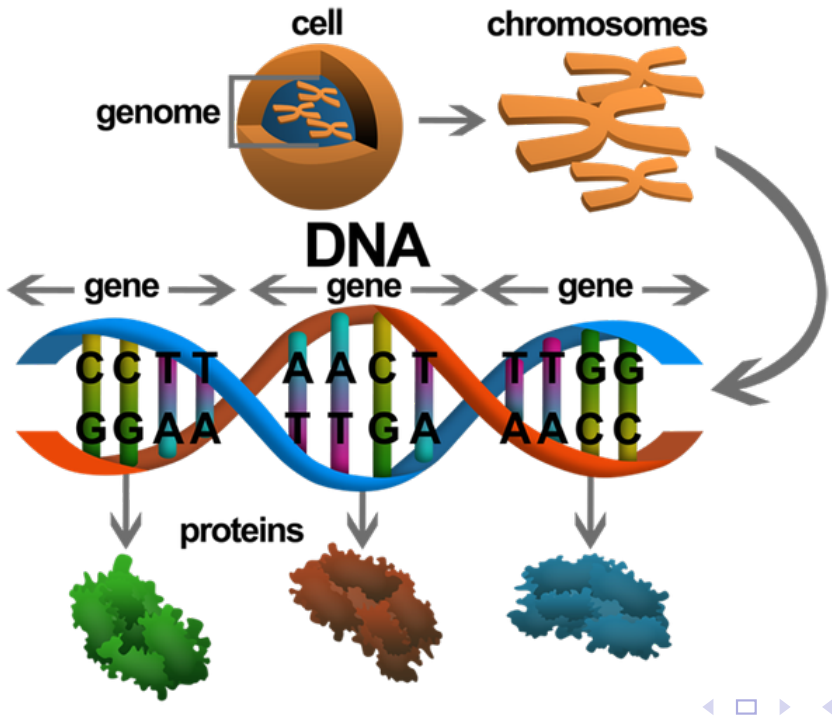
# Why preference learning matters?

- customers express preferences about products and services;
- users choose movies on an internet platform (e.g., Netflix);
- genes expression levels are related to their involvment in the biological process under study.



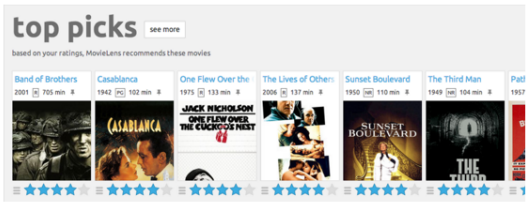
# Because preference data is everywhere!

- customers express preferences about products and services;
- users choose movies on an internet platform (e.g., Netflix);
- genes expression levels are related to their involvment in the biological process under study.

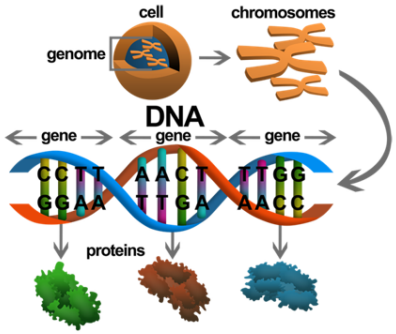


# Building blocks of a preference learning framework

sugar intake / taste / healthy




personal taste / time / popularity

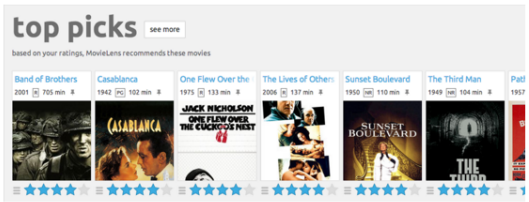



biological process / other factors

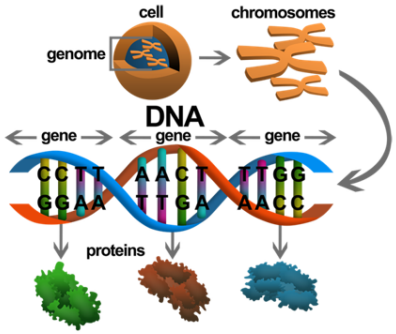
1. Set of ordered items
   - according to an unknown feature
   - not necessarily complete ranking

# Building blocks of a preference learning framework

sugar intake / taste / healthy




personal taste / time / popularity




biological process / other factors

1. Set of ordered items
   - according to an unknown feature
   - not necessarily complete ranking
2. Who provides the ordering?
   A set of assessors expressing their preference about items (as panels, users, patients, ...)

# Where to use this? Recommender Systems

**Challenges / opportunities:**

- messy data, typical of internet-user activities (rating, clicking, ...)
- not only an aggregation problem, but inference at the individual level
- prone to non trivial generalizations (on-line inference, inconsistencies, covariates, ...)

# Where to use this? Recommender Systems

**Challenges / opportunities:**

- messy data, typical of internet-user activities (rating, clicking, ...)
- not only an aggregation problem, but inference at the individual level
- prone to non trivial generalizations (on-line inference, inconsistencies, covariates, ...)

## htpp://movielens.org

MovieLens is run by GroupLens, a research lab at the University of Minnesota.

# Where to use this? Recommender Systems

**Challenges / opportunities:**

- messy data, typical of internet-user activities (rating, clicking, ...)
- not only an aggregation problem, but inference at the individual level
- prone to non trivial generalizations (on-line inference, inconsistencies, covariates, ...)

## htpp://movielens.org

MovieLens is run by GroupLens, a research lab at the University of Minnesota.

- **MovieLens provides non-commercial, personalized movie recommendations:** first the user builds a custom taste profile by rating already watched movies, then the system starts recommending.

# Where to use this? Recommender Systems

**Challenges / opportunities:**

- messy data, typical of internet-user activities (rating, clicking, ...)
- not only an aggregation problem, but inference at the individual level
- prone to non trivial generalizations (on-line inference, inconsistencies, covariates, ...)

## htpp://movielens.org

MovieLens is run by GroupLens, a research lab at the University of Minnesota.

- **MovieLens provides non-commercial, personalized movie recommendations:** first the user builds a custom taste profile by rating already watched movies, then the system starts recommending.
- **MovieLens is also a web platform providing good data for researchers** who aim at trying out their recommender systems.

# Where to use this? Cancer Genomics

- Omics analyses have focused on tissue-specific cancers & have been successful in identifying "within-tissue" molecular subtypes.

# Where to use this? Cancer Genomics

- Omics analyses have focused on tissue-specific cancers & have been successful in identifying "within-tissue" molecular subtypes.
- Studies of single cancer types describe the fundamental alterations of each cancer (signatures) with respect to normal tissues.

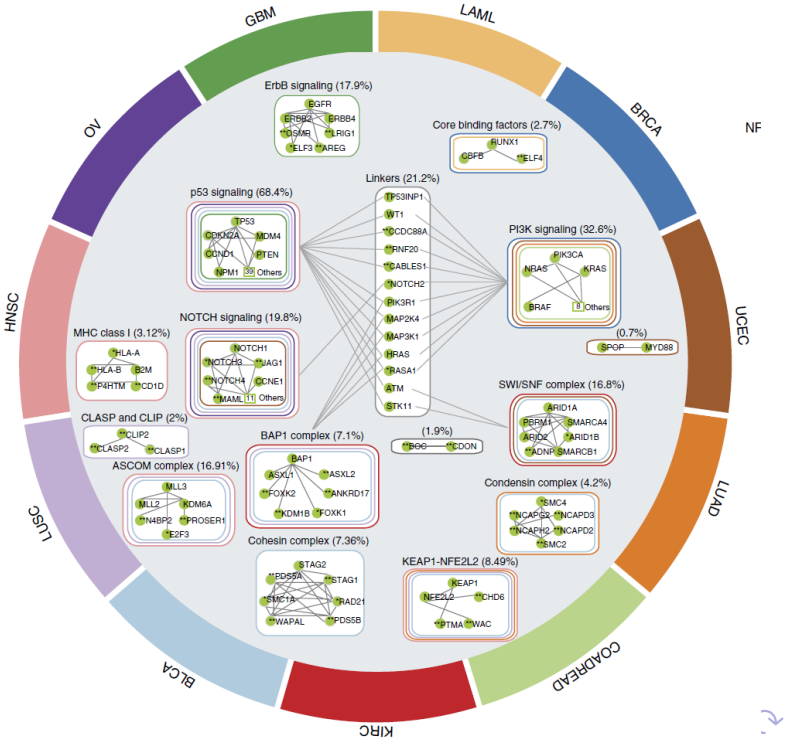
# Where to use this? Cancer Genomics

- Omics analyses have focused on tissue-specific cancers & have been successful in identifying "within-tissue" molecular subtypes.
- Studies of single cancer types describe the fundamental alterations of each cancer (signatures) with respect to normal tissues.

## Is cancer tissue-specific?

However, molecular disease mechanisms are known to be part of the development of diverse cancers across different tissues. Examples:

- **TP53 mutation:** serous ovarian, serous endometrial, basal-like breast cancers;
- **ERBB2-HER2 mutation/amplification:** subsets of glioblastoma, breast, gastric, serous endometrial, bladder and lung cancers.



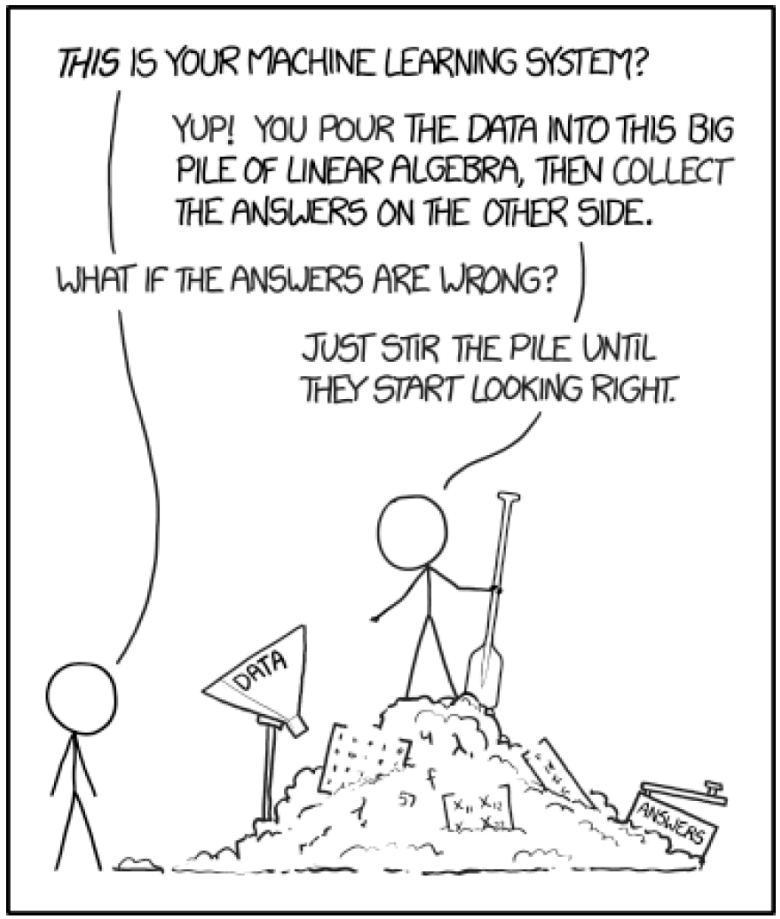
# Why model-based data analysis?

For the current problem-solving task:

- formalize reality in simple terms (good initial description of our beliefs)
- easy to control complexity (look into the box)
- neat probabilistic interpretations of the results

For the future:

- real applications are born to pass, models live forever
- sufficiently general to apply to a diversity of situations



# Ingredients for Preference Data

A set of **items**, to be evaluated...

# Ingredients for Preference Data

A set of **items**, to be evaluated...

...and a pool of **assessors** to evaluate them

# Objectives

1. aggregate, merge, summarize multiple rankings to estimate the **consensus ranking** across the assessors, and discover shared patterns and structures

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 3 | 2 | 3 | 3 |
| 2 | 3 | 2 | 1 | 1 | 2 | 1 |
| 3 | 2 | 3 | 2 | 3 | 1 | 2 |

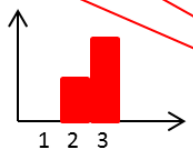


| |
|---|
| ? |
| ? |
| ? |

# Objectives

2 in case of **incomplete rankings** or preferences, predict the ranks of the missing items for each assessor

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 3 | 2 | 3 | 3 |
| 2 | 3 | ? | 1 | 1 | ? | 1 |
| 3 | 2 | ? | 2 | 3 | ? | 2 |

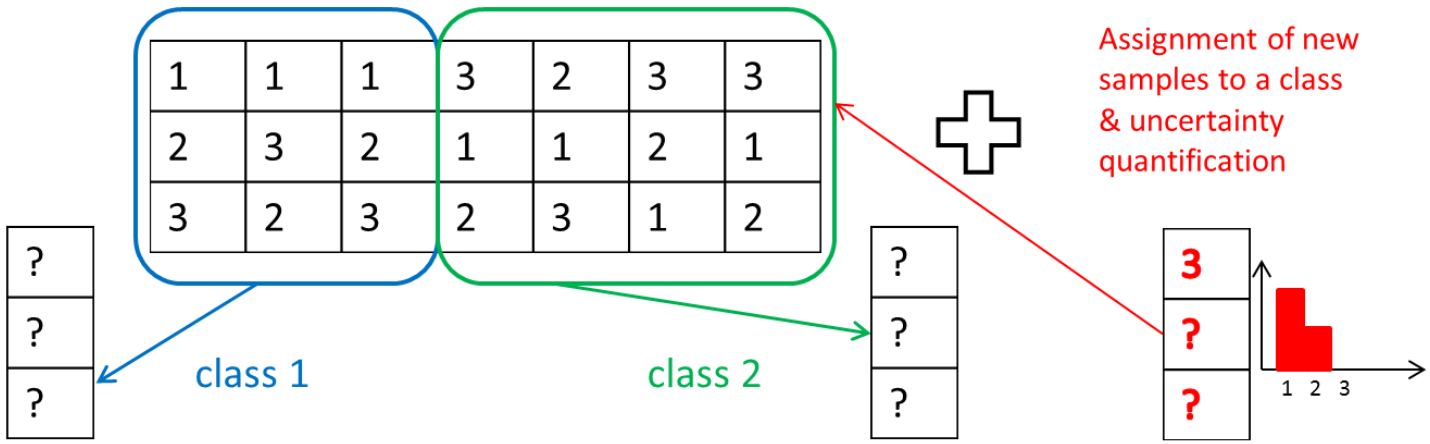



| |
|---|
| ? |
| ? |
| ? |

uncertainty quantification

# Objectives

3 partition the assessors into **classes**, each sharing a consensus ranking of the items, and classify new assessors to a class



# Possible approaches for handling preference data

Preference learning methods can be grouped according to their **purpose:**

1. **rank aggregation methods:** methods for learning the consensus ranking of a set of items by a group of assessors.

# Possible approaches for handling preference data

Preference learning methods can be grouped according to their **purpose:**

1. **rank aggregation methods:** methods for learning the consensus ranking of a set of items by a group of assessors. Examples: order statistics models (Thurstone 1927), Mallows-Bradley-Terry (MBT) model (Mallows 1957), Plackett-Luce (PL) model (Luce 1959; Plackett 1975), Bradley-Terry model (Zermelo 1929).

# Possible approaches for handling preference data

Preference learning methods can be grouped according to their **purpose:**

1. **rank aggregation methods:** methods for learning the consensus ranking of a set of items by a group of assessors. Examples: order statistics models (Thurstone 1927), Mallows-Bradley-Terry (MBT) model (Mallows 1957), Plackett-Luce (PL) model (Luce 1959; Plackett 1975), Bradley-Terry model (Zermelo 1929).
2. **preference learning methods:** methods for learning the individual rankings, when they are not readily available from the data, in order to perform personalized recommendations.

# Possible approaches for handling preference data

Preference learning methods can be grouped according to their **purpose:**

1. **rank aggregation methods:** methods for learning the consensus ranking of a set of items by a group of assessors. Examples: order statistics models (Thurstone 1927), Mallows-Bradley-Terry (MBT) model (Mallows 1957), Plackett-Luce (PL) model (Luce 1959; Plackett 1975), Bradley-Terry model (Zermelo 1929).
2. **preference learning methods:** methods for learning the individual rankings, when they are not readily available from the data, in order to perform personalized recommendations. Examples: the Hierarchical Bradley-Terry (HBT) model (Crispino & Frigessi, 2018, draft), which is based on the BT model. Collaborative Filtering (CF), which is grounded on matrix factorization techniques, and only infers on personal preferences (Koren et al. 2009).

# Possible approaches for handling preference data

- Most model-based methods assume that items are characterized by a real-valued **score** (or utility), on which the assessors' preferences (individual or shared) depend.

# Possible approaches for handling preference data

- Most model-based methods assume that items are characterized by a real-valued **score** (or utility), on which the assessors' preferences (individual or shared) depend.
- As an **alternative** to the above, **distance-based methods** are based on a **latent-ranking** of the items, with a parameter varying in the space of permutations $\mathcal{P}_n$: the Mallows model (Diaconis 1988), and its Bayesian version (Vitelli et al. 2018).

# Possible approaches for handling preference data

- Most model-based methods assume that items are characterized by a real-valued **score** (or utility), on which the assessors' preferences (individual or shared) depend.
- As an **alternative** to the above, **distance-based methods** are based on a **latent-ranking** of the items, with a parameter varying in the space of permutations $\mathcal{P}_n$: the Mallows model (Diaconis 1988), and its Bayesian version (Vitelli et al. 2018). The latter can handle both tasks, and it is thus **our preferred model.**

# Possible approaches for handling preference data

- Most model-based methods assume that items are characterized by a real-valued **score** (or utility), on which the assessors' preferences (individual or shared) depend.
- As an **alternative** to the above, **distance-based methods** are based on a **latent-ranking** of the items, with a parameter varying in the space of permutations $\mathcal{P}_n$: the Mallows model (Diaconis 1988), and its Bayesian version (Vitelli et al. 2018). The latter can handle both tasks, and it is thus **our preferred model.**
- See Liu et al. (2019a) for a review of possible alternative approaches to preference learning, with extensive comparisons of the Bayes Mallows performance with competitive methods on the very famous "potato data".

# Preliminaries

**Setting:**

- $N$ assessors rank $n$ items

# Preliminaries

**Setting:**

- $N$ assessors rank $n$ items
- let's start from the **complete data case**: for $j = 1, \ldots, N$, $\mathbf{R}_j$ is the ranking given by the $j$-th assessor to the full set of $n$ items

$$\mathbf{R}_j = (R_{1j}, R_{2j}, \ldots, R_{nj}) \in \mathcal{P}_n$$

where $\mathcal{P}_n$ is the space of all possible permutations of the vector of integers $(1, \ldots, n)$

# Preliminaries

**Setting:**

- $N$ assessors rank $n$ items
- let's start from the **complete data case**: for $j = 1, \ldots, N$, $\mathbf{R}_j$ is the ranking given by the $j$-th assessor to the full set of $n$ items

$$\mathbf{R}_j = (R_{1j}, R_{2j}, \ldots, R_{nj}) \in \mathcal{P}_n$$

  where $\mathcal{P}_n$ is the space of all possible permutations of the vector of integers $(1, \ldots, n)$
- in this setting, the only possible task is **rank aggregation,** i.e. finding a shared consensus ranking (and associated variability)

# Preliminaries

**Setting:**

- $N$ assessors rank $n$ items
- let's start from the **complete data case**: for $j = 1, \ldots, N$, $\mathbf{R}_j$ is the ranking given by the $j$-th assessor to the full set of $n$ items

$$\mathbf{R}_j = (R_{1j}, R_{2j}, \ldots, R_{nj}) \in \mathcal{P}_n$$

  where $\mathcal{P}_n$ is the space of all possible permutations of the vector of integers $(1, \ldots, n)$
- in this setting, the only possible task is **rank aggregation,** i.e. finding a shared consensus ranking (and associated variability)
- we consider the **Mallows model** (Mallows 1957)

# Mallows model (Mallows 1957)

Non-uniform joint distributions for $\mathbf{R}$ on $\mathcal{P}_n$, of the form

$$P(\mathbf{R}|\alpha, \boldsymbol{\rho}) = Z_n(\alpha, \boldsymbol{\rho})^{-1} \exp\{-(\alpha/n) d(\mathbf{R}, \boldsymbol{\rho})\} 1_{\mathcal{P}_n}(\mathbf{R}),$$

where:

- $\boldsymbol{\rho} \in \mathcal{P}_n$ is a location parameter,
- $\alpha$ is a positive scale parameter,
- $d(\cdot, \cdot)$ is a distance measure,
- and $Z_n(\alpha, \boldsymbol{\rho})$ is a normalizing constant (not computable for $n > 7$).

# Mallows model (Mallows 1957)

Non-uniform joint distributions for $\mathbf{R}$ on $\mathcal{P}_n$, of the form

$$P(\mathbf{R}|\alpha, \boldsymbol{\rho}) = Z_n(\alpha, \boldsymbol{\rho})^{-1} \exp\{-(\alpha/n) d(\mathbf{R}, \boldsymbol{\rho})\} 1_{\mathcal{P}_n}(\mathbf{R}),$$

where:

- $\boldsymbol{\rho} \in \mathcal{P}_n$ is a location parameter,
- $\alpha$ is a positive scale parameter,
- $d(\cdot, \cdot)$ is a distance measure,
- and $Z_n(\alpha, \boldsymbol{\rho})$ is a normalizing constant (not computable for $n > 7$).

If $d(\cdot, \cdot)$ is right-invariant, the normalizing constant does not depend on $\boldsymbol{\rho}$:

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

# $d(\mathbf{R}, \rho)$

The choice of the distance between rankings is crucial for the analysis.

# $d(\mathbf{R}, \rho)$

The choice of the distance between rankings is crucial for the analysis.
This choice also influences how $Z_n(\alpha)$ can be computed.

# $d(\mathbf{R}, \boldsymbol{\rho})$

The choice of the distance between rankings is crucial for the analysis.
This choice also influences how $Z_n(\alpha)$ can be computed.

- the **Kendall distance** measures the minimum number of pairwise adjacent transpositions which convert $\mathbf{R}$ into $\boldsymbol{\rho}$, and its normalizing constant simplifies to $Z_n(\alpha) = \prod_{i=1}^{n} \sum_{j=0}^{i-1} e^{-\alpha j/n}$, which is **computationally feasible also when $n$ is large.**
  For this reason, most applications of Mallows models have been restricted to Kendall distance.

# $d(\mathbf{R}, \boldsymbol{\rho})$

The choice of the distance between rankings is crucial for the analysis.
This choice also influences how $Z_n(\alpha)$ can be computed.

- the **Kendall distance** measures the minimum number of pairwise adjacent transpositions which convert $\mathbf{R}$ into $\boldsymbol{\rho}$, and its normalizing constant simplifies to $Z_n(\alpha) = \prod_{i=1}^{n} \sum_{j=0}^{i-1} e^{-\alpha j/n}$, which is **computationally feasible also when $n$ is large.**
  For this reason, most applications of Mallows models have been restricted to Kendall distance.
- However, other metrics have been suggested for particular applications, and important right-invariant metrics are
  the **footrule distance**, $d(\mathbf{R}, \boldsymbol{\rho}) = \sum_{i=1}^{n} |R_i - \rho_i|$,
  and the **Spearman distance** $d(\mathbf{R}, \boldsymbol{\rho}) = \sum_{i=1}^{n} (R_i - \rho_i)^2$.

# $d(\mathbf{R}, \rho)$

The choice of the distance between rankings is crucial for the analysis.
This choice also influences how $Z_n(\alpha)$ can be computed.

- the **Kendall distance** measures the minimum number of pairwise adjacent transpositions which convert $\mathbf{R}$ into $\boldsymbol{\rho}$, and its normalizing constant simplifies to $Z_n(\alpha) = \prod_{i=1}^{n} \sum_{j=0}^{i-1} e^{-\alpha j/n}$, which is **computationally feasible also when $n$ is large.**
  For this reason, most applications of Mallows models have been restricted to Kendall distance.
- However, other metrics have been suggested for particular applications, and important right-invariant metrics are
  the **footrule distance**, $d(\mathbf{R}, \boldsymbol{\rho}) = \sum_{i=1}^{n} |R_i - \rho_i|$,
  and the **Spearman distance** $d(\mathbf{R}, \boldsymbol{\rho}) = \sum_{i=1}^{n} (R_i - \rho_i)^2$.
  The computation of $Z_n(\alpha)$ in these cases is not feasible exactly.

# $d(\mathbf{R}, \boldsymbol{\rho})$

The choice of the distance between rankings is crucial for the analysis.
This choice also influences how $Z_n(\alpha)$ can be computed.

- the **Kendall distance** measures the minimum number of pairwise adjacent transpositions which convert $\mathbf{R}$ into $\boldsymbol{\rho}$, and its normalizing constant simplifies to $Z_n(\alpha) = \prod_{i=1}^{n} \sum_{j=0}^{i-1} e^{-\alpha j/n}$, which is **computationally feasible also when $n$ is large.**
  For this reason, most applications of Mallows models have been restricted to Kendall distance.
- However, other metrics have been suggested for particular applications, and important right-invariant metrics are
  the **footrule distance**, $d(\mathbf{R}, \boldsymbol{\rho}) = \sum_{i=1}^{n} |R_i - \rho_i|$,
  and the **Spearman distance** $d(\mathbf{R}, \boldsymbol{\rho}) = \sum_{i=1}^{n} (R_i - \rho_i)^2$.
  The computation of $Z_n(\alpha)$ in these cases is not feasible exactly.
- Other examples are the Hamming distance, the Ulam distance and the Cayley distance (Marden 1995:pp. 23-27).

# Bayesian Mallows model

- Assume that $\mathbf{R}_1, \ldots, \mathbf{R}_N | \alpha, \boldsymbol{\rho} \overset{i.i.d.}{\sim} \text{Mallows}(\alpha, \boldsymbol{\rho})$.

# Bayesian Mallows model

- Assume that $\mathbf{R}_1, \ldots, \mathbf{R}_N | \alpha, \boldsymbol{\rho} \overset{i.i.d.}{\sim} \text{Mallows}(\alpha, \boldsymbol{\rho})$.
- The likelihood for the Mallows model is:

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

# Bayesian Mallows model

- Assume that $\mathbf{R}_1, \ldots, \mathbf{R}_N | \alpha, \boldsymbol{\rho} \overset{i.i.d.}{\sim} \text{Mallows}(\alpha, \boldsymbol{\rho})$.
- The likelihood for the Mallows model is:

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

- Priors: when **no prior information is available** about $\boldsymbol{\rho}$, we assign a uniform prior over $\mathcal{P}_n$: $\pi(\boldsymbol{\rho}) = \frac{1}{n!} 1_{\mathcal{P}_n}(\boldsymbol{\rho})$.

# Bayesian Mallows model

- Assume that $\mathbf{R}_1, \ldots, \mathbf{R}_N | \alpha, \boldsymbol{\rho} \overset{i.i.d.}{\sim} \text{Mallows}(\alpha, \boldsymbol{\rho})$.
- The likelihood for the Mallows model is:

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

- Priors: when **no prior information is available** about $\boldsymbol{\rho}$, we assign a uniform prior over $\mathcal{P}_n$: $\pi(\boldsymbol{\rho}) = \frac{1}{n!} 1_{\mathcal{P}_n}(\boldsymbol{\rho})$.
  Assigning $\alpha$ a **truncated exponential prior**, with density $\pi(\alpha|\lambda) = \lambda e^{-\lambda\alpha} 1_{[0,\alpha_{\text{max}}]}(\alpha)/(1 - e^{-\lambda\alpha_{\text{max}}})$, is a reasonable choice to a priori ensure a good dispersion around the consensus ranking. Note that the cut-off point $\alpha_{\text{max}} < \infty$ is large compared to the values supported by the data.

# Bayesian Mallows model

- Assume that $\mathbf{R}_1, \ldots, \mathbf{R}_N | \alpha, \boldsymbol{\rho} \overset{i.i.d.}{\sim} \text{Mallows}(\alpha, \boldsymbol{\rho})$.
- The likelihood for the Mallows model is:

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

- Priors: when **no prior information is available** about $\boldsymbol{\rho}$, we assign a uniform prior over $\mathcal{P}_n$: $\pi(\boldsymbol{\rho}) = \frac{1}{n!} 1_{\mathcal{P}_n}(\boldsymbol{\rho})$.
  Assigning $\alpha$ a **truncated exponential prior**, with density $\pi(\alpha|\lambda) = \lambda e^{-\lambda\alpha} 1_{[0,\alpha_{\text{max}}]}(\alpha)/(1 - e^{-\lambda\alpha_{\text{max}}})$, is a reasonable choice to a priori ensure a good dispersion around the consensus ranking. Note that the cut-off point $\alpha_{\text{max}} < \infty$ is large compared to the values supported by the data.
- Probabilistic statements about $\boldsymbol{\rho}$ and $\alpha$ are based on the joint posterior:

$$P(\boldsymbol{\rho}, \alpha | \mathbf{R}_1, \ldots, \mathbf{R}_N) \propto Z_n(\alpha)^{-N} \exp\left\{ -\alpha \left[ n^{-1} \sum_{j=1}^{N} d(\mathbf{R}_j, \boldsymbol{\rho}) + \lambda \right] \right\}.$$

# Metropolis-Hastings scheme

To obtain samples from the posterior distribution, we iterate two steps.

# Metropolis-Hastings scheme

To obtain samples from the posterior distribution, we iterate two steps.

1. starting at $\alpha \geq 0$, $\rho \in \mathcal{P}_n$, we propose $\boldsymbol{\rho}'$, and accept it with probability

$$\min\left\{1, \frac{P_L(\boldsymbol{\rho}|\boldsymbol{\rho}')\pi(\boldsymbol{\rho}')}{P_L(\boldsymbol{\rho}'|\boldsymbol{\rho})\pi(\boldsymbol{\rho})} \exp\left[-\frac{\alpha}{n}\sum_{j=1}^{N}\left\{d\left(\mathbf{R}_j, \boldsymbol{\rho}'\right) - d\left(\mathbf{R}_j, \boldsymbol{\rho}\right)\right\}\right]\right\},$$

# Metropolis-Hastings scheme

To obtain samples from the posterior distribution, we iterate two steps.

1. starting at $\alpha \geq 0$, $\rho \in \mathcal{P}_n$, we propose $\boldsymbol{\rho}'$, and accept it with probability

$$\min\left\{1, \frac{P_L(\boldsymbol{\rho}|\boldsymbol{\rho}')\pi(\boldsymbol{\rho}')}{P_L(\boldsymbol{\rho}'|\boldsymbol{\rho})\pi(\boldsymbol{\rho})} \exp\left[-\frac{\alpha}{n}\sum_{j=1}^{N}\left\{d\left(\mathbf{R}_j, \boldsymbol{\rho}'\right) - d\left(\mathbf{R}_j, \boldsymbol{\rho}\right)\right\}\right]\right\},$$

where $\boldsymbol{\rho}'$ is proposed according to the leap-and-shift distribution, which ensures a sufficient acceptance ratio by inducing small perturbations of elements of $\boldsymbol{\rho}$, while keeping $\boldsymbol{\rho}' \in \mathcal{P}_n$;

# Metropolis-Hastings scheme

To obtain samples from the posterior distribution, we iterate two steps.

1. starting at $\alpha \geq 0$, $\rho \in \mathcal{P}_n$, we propose $\boldsymbol{\rho}'$, and accept it with probability

$$\min\left\{1, \frac{P_L(\boldsymbol{\rho}|\boldsymbol{\rho}')\pi(\boldsymbol{\rho}')}{P_L(\boldsymbol{\rho}'|\boldsymbol{\rho})\pi(\boldsymbol{\rho})} \exp\left[-\frac{\alpha}{n}\sum_{j=1}^{N}\left\{d\left(\mathbf{R}_j, \boldsymbol{\rho}'\right) - d\left(\mathbf{R}_j, \boldsymbol{\rho}\right)\right\}\right]\right\},$$

   where $\boldsymbol{\rho}'$ is proposed according to the leap-and-shift distribution, which ensures a sufficient acceptance ratio by inducing small perturbations of elements of $\boldsymbol{\rho}$, while keeping $\boldsymbol{\rho}' \in \mathcal{P}_n$;

2. we propose $\alpha'$ according to $\log \mathcal{N}(\alpha, \sigma_\alpha^2)$, and accept it with probability

$$\min\left\{1, \frac{Z_n(\alpha)^N \pi(\alpha')\alpha'}{Z_n(\alpha')^N \pi(\alpha)\alpha} \exp\left[-\frac{(\alpha' - \alpha)}{n}\sum_{j=1}^{N} d\left(\mathbf{R}_j, \boldsymbol{\rho}\right)\right]\right\},$$

   where $\sigma_\alpha^2$ can be tuned to obtain a desired acceptance ratio.

# Metropolis-Hastings scheme: further details

To obtain samples from the posterior distribution, we iterate two steps.

1. starting at $\alpha \geq 0$, $\rho \in \mathcal{P}_n$, we propose $\boldsymbol{\rho}'$, and accept it with probability

$$\min\left\{1, \frac{P_L(\boldsymbol{\rho}|\boldsymbol{\rho}')\pi(\boldsymbol{\rho}')}{P_L(\boldsymbol{\rho}'|\boldsymbol{\rho})\pi(\boldsymbol{\rho})} \exp\left[-\frac{\alpha}{n}\sum_{j=1}^{N}\left\{d\left(\mathbf{R}_j, \boldsymbol{\rho}'\right) - d\left(\mathbf{R}_j, \boldsymbol{\rho}\right)\right\}\right]\right\},$$

   This term can be computed efficiently, since most elements of $\boldsymbol{\rho}$ and $\boldsymbol{\rho}'$ are equal. If $\rho_i = \rho_i'$ for $i \in E \subset \{1, \ldots, n\}$, each MCMC iteration involves computations only on $E^c$, making the algorithm scalable with $n$.

2. we propose $\alpha'$ according to $\log \mathcal{N}(\alpha, \sigma_\alpha^2)$, and accept it with probability

$$\min\left\{1, \frac{Z_n(\alpha)^N \pi(\alpha')\,\alpha'}{Z_n(\alpha')^N \pi(\alpha)\,\alpha} \exp\left[-\frac{(\alpha' - \alpha)}{n}\sum_{j=1}^{N} d\left(\mathbf{R}_j, \boldsymbol{\rho}\right)\right]\right\},$$

   This term is computationally heavy, and is thus approximated off-line.

# Off-line computation of $Z_n(\alpha)$

$Z_n(\alpha)$ is known for some choices of $d(\cdot,\cdot)$ and $n$ (see Vitelli et al. (2018)).

In all other cases, we propose the following **importance sampling (IS)** scheme to obtain an off-line approximation of

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

Note. $\mathbf{P}$ is any reference permutation, for instance $\mathbf{1} := \{1,\ldots,n\}$.

# Off-line computation of $Z_n(\alpha)$

$Z_n(\alpha)$ is known for some choices of $d(\cdot,\cdot)$ and $n$ (see Vitelli et al. (2018)).

In all other cases, we propose the following **importance sampling (IS)** scheme to obtain an off-line approximation of

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

Note. $\mathbf{P}$ is any reference permutation, for instance $\mathbf{1} := \{1,\ldots,n\}$.

1. Define a grid of resonable values for $\alpha$.

# Off-line computation of $Z_n(\alpha)$

$Z_n(\alpha)$ is known for some choices of $d(\cdot,\cdot)$ and $n$ (see Vitelli et al. (2018)).

In all other cases, we propose the following **importance sampling (IS)** scheme to obtain an off-line approximation of

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

Note. $\mathbf{P}$ is any reference permutation, for instance $\mathbf{1} := \{1,\ldots,n\}$.

1. Define a grid of resonable values for $\alpha$.
2. Invent a distribution $q(\mathbf{R})$ from which it is easy to sample permutations.

# Off-line computation of $Z_n(\alpha)$

$Z_n(\alpha)$ is known for some choices of $d(\cdot,\cdot)$ and $n$ (see Vitelli et al. (2018)).

In all other cases, we propose the following **importance sampling (IS)** scheme to obtain an off-line approximation of

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

Note. $\mathbf{P}$ is any reference permutation, for instance $\mathbf{1} := \{1,\ldots,n\}$.

1. Define a grid of resonable values for $\alpha$.
2. Invent a distribution $q(\mathbf{R})$ from which it is easy to sample permutations.
3. Fix $K$ sufficiently large, and sample $\mathbf{R}^1,\ldots,\mathbf{R}^K$.

# Off-line computation of $Z_n(\alpha)$

$Z_n(\alpha)$ is known for some choices of $d(\cdot,\cdot)$ and $n$ (see Vitelli et al. (2018)).

In all other cases, we propose the following **importance sampling (IS)** scheme to obtain an off-line approximation of

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

Note. $\mathbf{P}$ is any reference permutation, for instance $\mathbf{1} := \{1,\ldots,n\}$.

1. Define a grid of resonable values for $\alpha$.
2. Invent a distribution $q(\mathbf{R})$ from which it is easy to sample permutations.
3. Fix $K$ sufficiently large, and sample $\mathbf{R}^1,\ldots,\mathbf{R}^K$.
4. Compute

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

# Off-line computation of $Z_n(\alpha)$

$Z_n(\alpha)$ is known for some choices of $d(\cdot,\cdot)$ and $n$ (see Vitelli et al. (2018)).

In all other cases, we propose the following **importance sampling (IS)** scheme to obtain an off-line approximation of

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

Note. $\mathbf{P}$ is any reference permutation, for instance $\mathbf{1} := \{1,\dots,n\}$.

1. Define a grid of resonable values for $\alpha$.
2. Invent a distribution $q(\mathbf{R})$ from which it is easy to sample permutations.
3. Fix $K$ sufficiently large, and sample $\mathbf{R}^1,\dots,\mathbf{R}^K$.
4. Compute

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

5. Interpolate the values of $\hat{Z}_n(\alpha)$ obtained for each $\alpha$ on the grid, so to obtain a smooth estimate to be used in the MCMC.

# Off-line computation of $Z_n(\alpha)$

Note 1. $\mathbb{E}\left\{\hat{Z}_n(\alpha)\right\} = Z_n(\alpha)$.

# Off-line computation of $Z_n(\alpha)$

Note 1. $\mathbb{E}\left\{\hat{Z}_n(\alpha)\right\} = Z_n(\alpha)$. Due to the IS construction, the closer $q(\cdot)$ is to the actual Mallows model distribution, the better the approximation is.

# Off-line computation of $Z_n(\alpha)$

Note 1. $\mathbb{E}\left\{\hat{Z}_n(\alpha)\right\} = Z_n(\alpha)$. Due to the IS construction, the closer $q(\cdot)$ is to the actual Mallows model distribution, the better the approximation is.

Note 2. A good candidate for the IS distribution $q(\cdot)$ is what we call **randomized pseudo-likelihood** distribution:

- sample a random permutation $\{i_1, \ldots, i_n\}$ in $\mathcal{P}_n$;

# Off-line computation of $Z_n(\alpha)$

Note 1. $\mathbb{E}\left\{\hat{Z}_n(\alpha)\right\} = Z_n(\alpha)$. Due to the IS construction, the closer $q(\cdot)$ is to the actual Mallows model distribution, the better the approximation is.

Note 2. A good candidate for the IS distribution $q(\cdot)$ is what we call **randomized pseudo-likelihood** distribution:

- sample a random permutation $\{i_1, \ldots, i_n\}$ in $\mathcal{P}_n$;
- start with the $i_1$-th item:
  $P(R_{i_1}|\mathbf{1}) \propto \exp\{-(\alpha/n)d(R_{i_1}, i_1)\} \times 1_{[1,\ldots,n]}(R_{i_1})$;

# Off-line computation of $Z_n(\alpha)$

Note 1. $\mathbb{E}\left\{\hat{Z}_n(\alpha)\right\} = Z_n(\alpha)$. Due to the IS construction, the closer $q(\cdot)$ is to the actual Mallows model distribution, the better the approximation is.

Note 2. A good candidate for the IS distribution $q(\cdot)$ is what we call **randomized pseudo-likelihood** distribution:

- sample a random permutation $\{i_1, \ldots, i_n\}$ in $\mathcal{P}_n$;
- start with the $i_1$-th item:
  $P(R_{i_1}|\mathbf{1}) \propto \exp\{-(\alpha/n)d(R_{i_1}, i_1)\} \times 1_{[1,\ldots,n]}(R_{i_1})$;
- then proceed to the $i_2$-th
  $P(R_{i_2}|R_{1_1}, \mathbf{1}) \propto \exp\{-(\alpha/n)d(R_{i_2}, i_2)\} \times 1_{[\{1,\ldots,n\}\setminus\{R_{i_1}\}]}(R_{i_2})$;

# Off-line computation of $Z_n(\alpha)$

Note 1. $\mathbb{E}\left\{\hat{Z}_n(\alpha)\right\} = Z_n(\alpha)$. Due to the IS construction, the closer $q(\cdot)$ is to the actual Mallows model distribution, the better the approximation is.

Note 2. A good candidate for the IS distribution $q(\cdot)$ is what we call **randomized pseudo-likelihood** distribution:

- sample a random permutation $\{i_1, \ldots, i_n\}$ in $\mathcal{P}_n$;
- start with the $i_1$-th item:
  $P(R_{i_1}|\mathbf{1}) \propto \exp\{-(\alpha/n)d(R_{i_1}, i_1)\} \times 1_{[1,\ldots,n]}(R_{i_1})$;
- then proceed to the $i_2$-th
  $P(R_{i_2}|R_{1_1}, \mathbf{1}) \propto \exp\{-(\alpha/n)d(R_{i_2}, i_2)\} \times 1_{\left[\{1,\ldots,n\}\setminus\{R_{i_1}\}\right]}(R_{i_2})$;
- and so on like this until you get to the $i_n$-th:
  $P\left(R_{i_n}|R_{i_1}, \ldots, R_{i_{n-1}}, \mathbf{1}\right) = 1_{\left[\{1,\ldots,n\}\setminus\{R_{i_1},\ldots,R_{i_{n-1}}\}\right]}(R_{i_n})$.

# Off-line computation of $Z_n(\alpha)$

Note 1. $\mathbb{E}\left\{\hat{Z}_n(\alpha)\right\} = Z_n(\alpha)$. Due to the IS construction, the closer $q(\cdot)$ is to the actual Mallows model distribution, the better the approximation is.

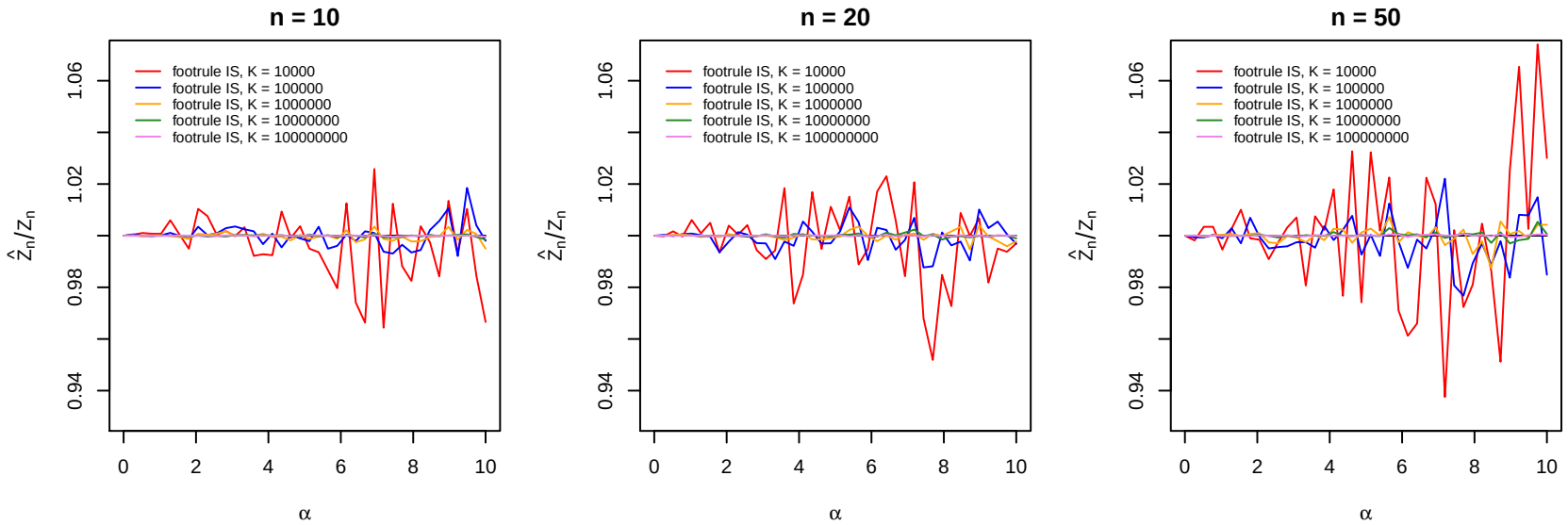
Note 2. A good candidate for the IS distribution $q(\cdot)$ is what we call **randomized pseudo-likelihood** distribution:

- sample a random permutation $\{i_1, \ldots, i_n\}$ in $\mathcal{P}_n$;
- start with the $i_1$-th item:
  $P(R_{i_1}|\mathbf{1}) \propto \exp\{-(\alpha/n)d(R_{i_1}, i_1)\} \times 1_{[1,\ldots,n]}(R_{i_1})$;
- then proceed to the $i_2$-th
  $P(R_{i_2}|R_{1_1}, \mathbf{1}) \propto \exp\{-(\alpha/n)d(R_{i_2}, i_2)\} \times 1_{\left[\{1,\ldots,n\}\setminus\{R_{i_1}\}\right]}(R_{i_2})$;
- and so on like this until you get to the $i_n$-th:
  $P\left(R_{i_n}|R_{i_1}, \ldots, R_{i_{n-1}}, \mathbf{1}\right) = 1_{\left[\{1,\ldots,n\}\setminus\{R_{i_1},\ldots,R_{i_{n-1}}\}\right]}(R_{i_n})$.

Note 3. A good alternative to the IS construction, for large $n$, is the asymptotic off-line approximation $Z_{lim}(\alpha)$ given in Mukherjee (2016) for $n \to \infty$.

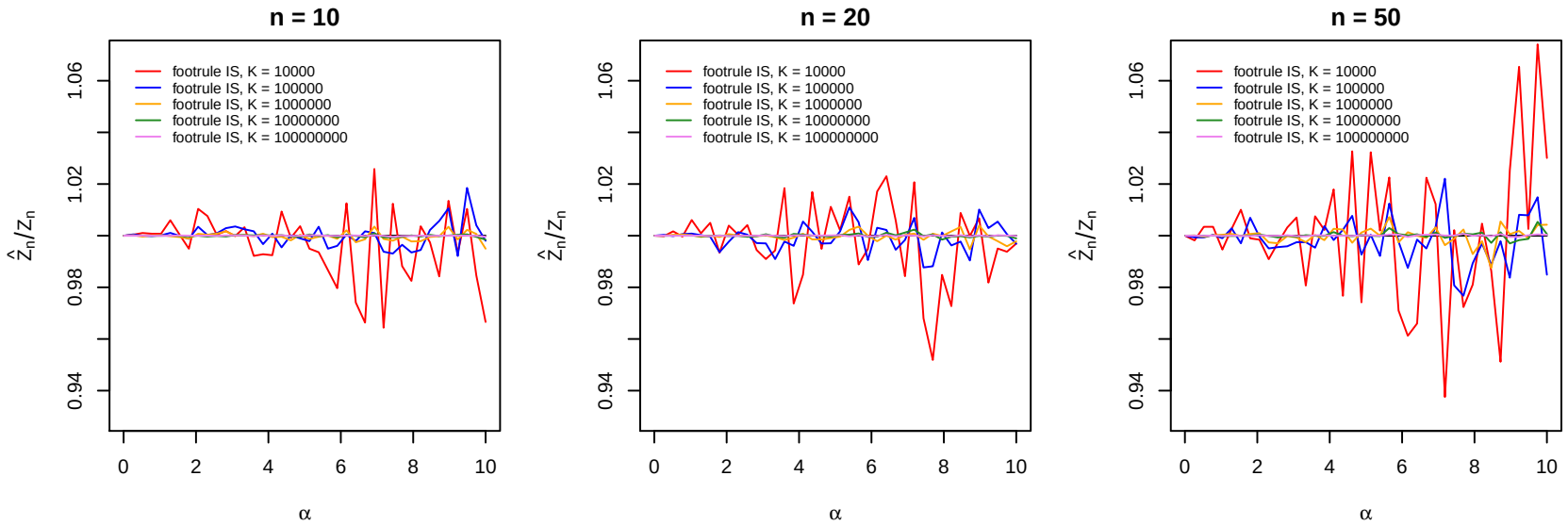
# Testing the importance sampler

Ratio of the approximate partition function computed via IS to the exact:



# Testing the importance sampler

Ratio of the approximate partition function computed via IS to the exact:




Maximum relative error between the current and the previous $K$:

| $K$ | $10^2$ | $10^3$ | $10^4$ | $10^5$ | $10^6$ | $10^7$ | $10^8$ |
|---|---|---|---|---|---|---|---|
| $n = 75$ | 152.036 | 0.921 | 0.373 | 0.084 | 0.056 | 0.005 | 0.004 |
| $n = 100$ | 67.487 | 1.709 | 0.355 | 0.187 | 0.045 | 0.018 | 0.004 |

# Effect of the approximation of $Z_n(\alpha)$ on inference

**Proposition 1.** The previous M-H algorithm using $\hat{Z}_n(\alpha)$ instead of $Z_n(\alpha)$ converges to the posterior distribution proportional to

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

with the normalizing factor
$\hat{C}(\mathbf{R}) = \int_0^\infty \pi(\boldsymbol{\rho})\pi(\alpha)\hat{Z}_n(\alpha)^{-N}\sum_{\boldsymbol{\rho}\in\mathcal{P}_n}\exp\left\{-\frac{\alpha}{n}\sum_{j=1}^{N}d(\mathbf{R}_j,\boldsymbol{\rho})\right\}d\alpha.$

# Effect of the approximation of $Z_n(\alpha)$ on inference

**Proposition 1.** The previous M-H algorithm using $\hat{Z}_n(\alpha)$ instead of $Z_n(\alpha)$ converges to the posterior distribution proportional to

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

with the normalizing factor
$\hat{C}(\mathbf{R}) = \int_0^\infty \pi(\boldsymbol{\rho})\pi(\alpha)\hat{Z}_n(\alpha)^{-N}\sum_{\boldsymbol{\rho}\in\mathcal{P}_n}\exp\left\{-\frac{\alpha}{n}\sum_{j=1}^{N}d(\mathbf{R}_j,\boldsymbol{\rho})\right\}d\alpha.$

**Proposition 2.** There exists a factor $c(\alpha, n, d(\cdot,\cdot))$ not depending on $N$, such that, if $K = K(N)$ tends to infinity as $N \to \infty$ faster than $c(\alpha, n, d(\cdot,\cdot)) \cdot N^2$, then it holds that

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

# Effect of the approximation of $Z_n(\alpha)$ on inference

**Proposition 1.** The previous M-H algorithm using $\hat{Z}_n(\alpha)$ instead of $Z_n(\alpha)$ converges to the posterior distribution proportional to

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

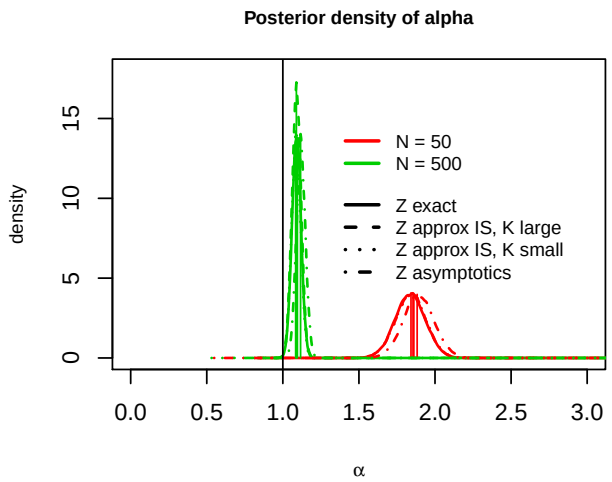
with the normalizing factor
$\hat{C}(\mathbf{R}) = \int_0^\infty \pi(\boldsymbol{\rho})\pi(\alpha)\hat{Z}_n(\alpha)^{-N}\sum_{\boldsymbol{\rho}\in\mathcal{P}_n}\exp\left\{-\frac{\alpha}{n}\sum_{j=1}^{N}d(\mathbf{R}_j,\boldsymbol{\rho})\right\}d\alpha.$

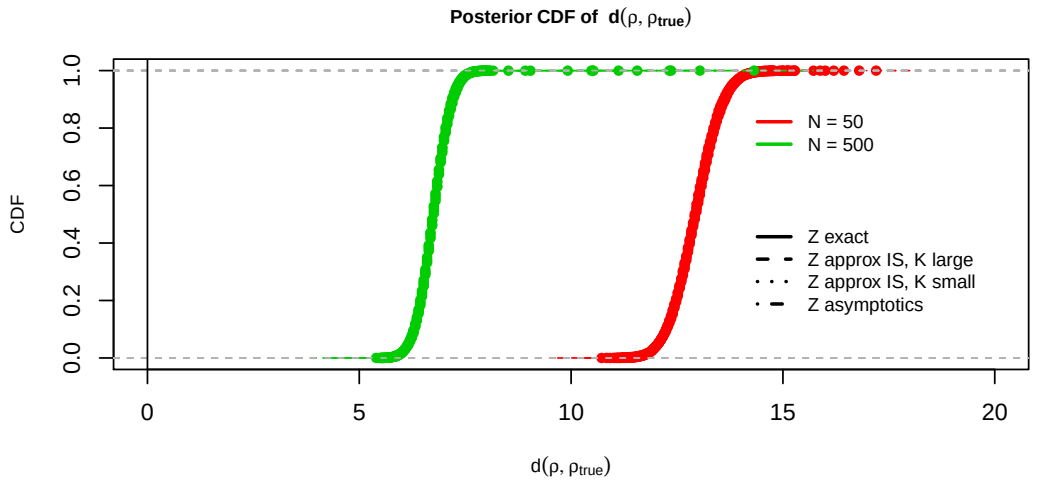
**Proposition 2.** There exists a factor $c(\alpha, n, d(\cdot,\cdot))$ not depending on $N$, such that, if $K = K(N)$ tends to infinity as $N \to \infty$ faster than $c(\alpha, n, d(\cdot,\cdot)) \cdot N^2$, then it holds that

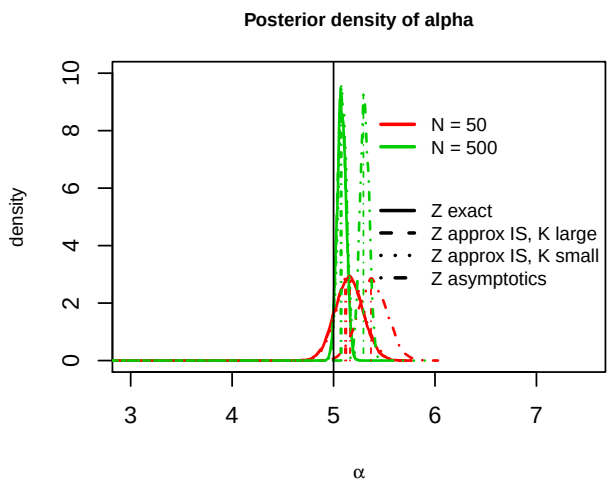
$$\lim_{N\to\infty}\left(\frac{\hat{Z}_n^{K(N)}(\alpha)}{Z_n(\alpha)}\right)^N = 1, \quad \text{for all } \alpha.$$

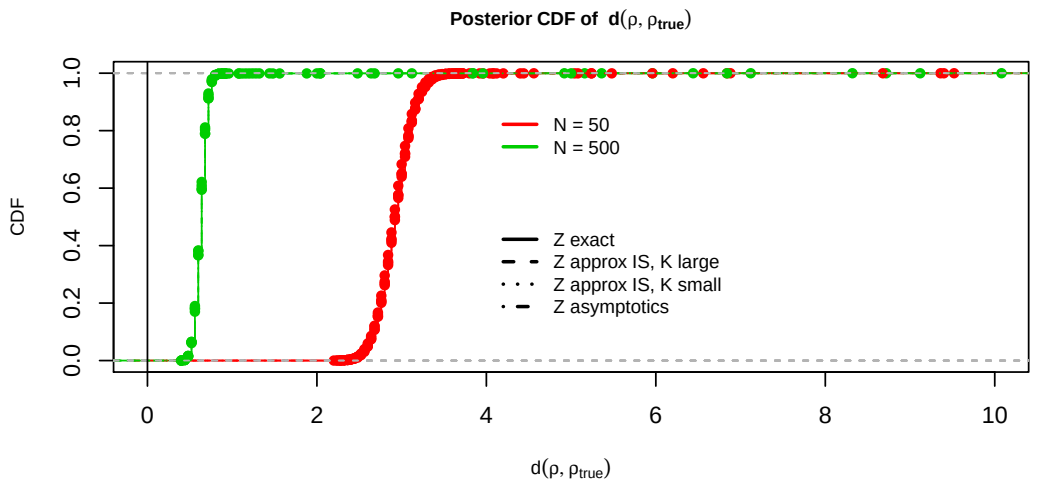
Hence, things should be just fine for reasonable values of $K$, $n$, and $N$

# Effect of the approximation of $Z_n(\alpha)$ on inference

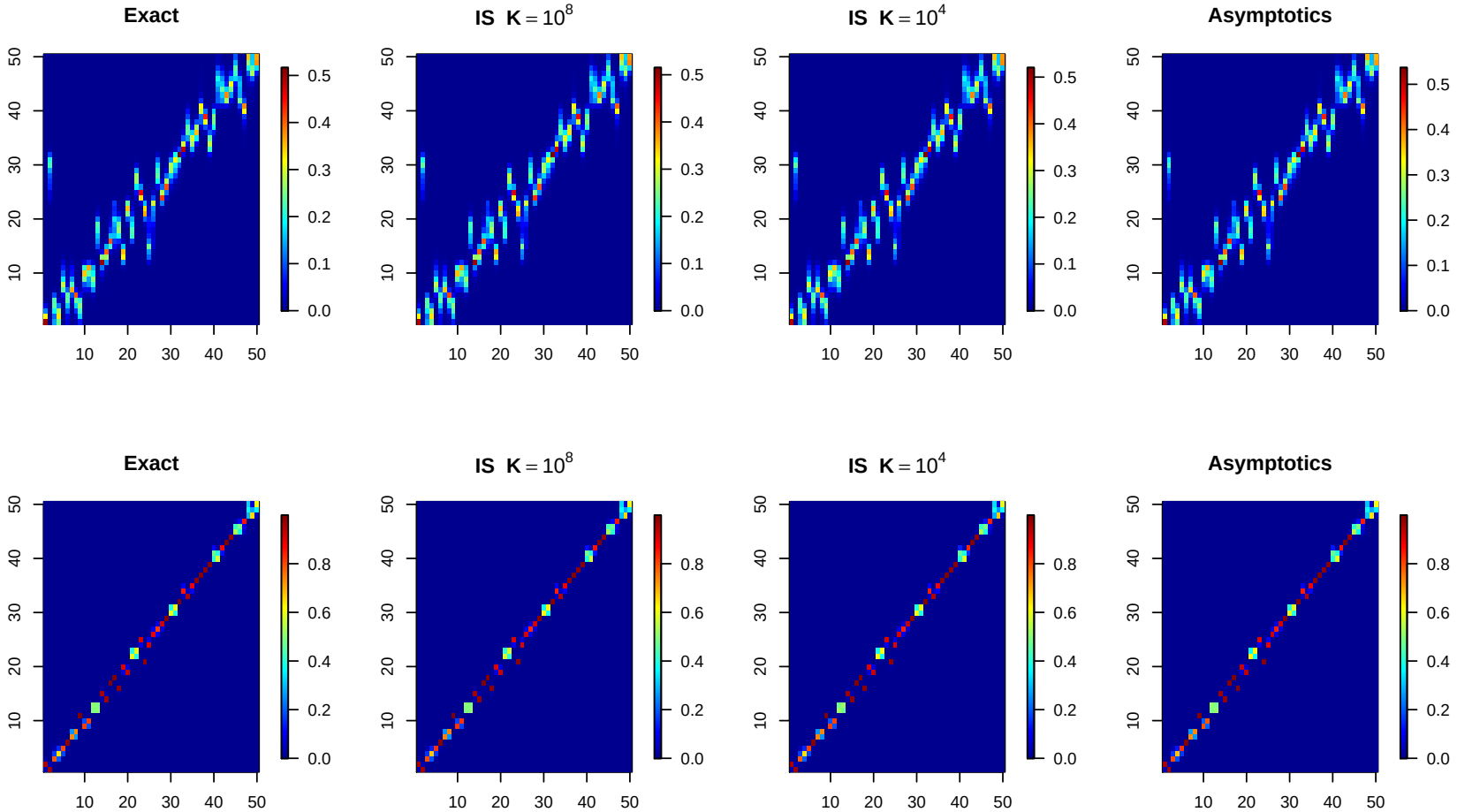


# Effect of the approximation of $Z_n(\alpha)$ on inference



# Partial Rankings

- Only a subset of the items are ranked.

# Partial Rankings

- Only a subset of the items are ranked.
- Ranks can be missing at random, or the assessors may only have ranked the in-their-opinion top-$k$ items.

# Partial Rankings

- Only a subset of the items are ranked.
- Ranks can be missing at random, or the assessors may only have ranked the in-their-opinion top-$k$ items.
- This is a very common situation in the applications, especially if $n$ is large.

# Partial Rankings

- Only a subset of the items are ranked.
- Ranks can be missing at random, or the assessors may only have ranked the in-their-opinion top-$k$ items.
- This is a very common situation in the applications, especially if $n$ is large.

Partial rankings can be handled easily in the Bayesian framework, by applying **data augmentation** techniques: estimating the lacking ranks consistently with the partial observations.

# Partial Rankings

- Only a subset of the items are ranked.
- Ranks can be missing at random, or the assessors may only have ranked the in-their-opinion top-$k$ items.
- This is a very common situation in the applications, especially if $n$ is large.

Partial rankings can be handled easily in the Bayesian framework, by applying **data augmentation** techniques: estimating the lacking ranks consistently with the partial observations. Let $\tilde{\mathbf{R}}_1, \ldots, \tilde{\mathbf{R}}_N$ be the augmented rankings, the MCMC alternates between two steps:

1. update the parameters given the augmented rankings (basic Mallows model)
$$P(\boldsymbol{\rho}, \alpha | \tilde{\mathbf{R}}_1, \ldots, \tilde{\mathbf{R}}_N);$$
2. sample the augmented ranks from a Mallows model given the current parameter values
$$P(\tilde{\mathbf{R}}_1, \ldots, \tilde{\mathbf{R}}_N | \boldsymbol{\rho}, \alpha; \mathbf{R}_1, \ldots, \mathbf{R}_N).$$

# Pairwise Preferences

- The assessors do not even see all the items, but just compare their level of preference between two of them.

# Pairwise Preferences

- The assessors do not even see all the items, but just compare their level of preference between two of them.
- Very sparse data, typical of internet users activities (ex. MovieLens).

# Pairwise Preferences

- The assessors do not even see all the items, but just compare their level of preference between two of them.
- Very sparse data, typical of internet users activities (ex. MovieLens).

Pairwise preferences can also be handled via data augmentation.

# Pairwise Preferences

- The assessors do not even see all the items, but just compare their level of preference between two of them.
- Very sparse data, typical of internet users activities (ex. MovieLens).

Pairwise preferences can also be handled via data augmentation. Suppose to observe the following preferences for assessor $j$: $\mathbf{R}_j = \{A > B, B > C, D > C\}$. Then, inside the MCMC

1. compute the associated **transitive closure**

$$tc_j = \mathbf{R}_j \cup \{A > C\}$$

2. sample $\tilde{\mathbf{R}}_j \in \mathcal{P}_n$ consistent with $tc_j$.

# Clustering

- Assessors cannot be assumed to form one homogeneous group, but possibly $C$ groups.

# Clustering

- Assessors cannot be assumed to form one homogeneous group, but possibly $C$ groups.
- We use a mixture of Mallows models to cluster the $N$ assessors according to how they rank the $n$ items.

# Clustering

- Assessors cannot be assumed to form one homogeneous group, but possibly $C$ groups.
- We use a mixture of Mallows models to cluster the $N$ assessors according to how they rank the $n$ items.
- We estimate a latent ranking of the items $\boldsymbol{\rho}_c$ with its dispersion parameter $\alpha_c$ for each cluster of assessors.

# Clustering

- Assessors cannot be assumed to form one homogeneous group, but possibly $C$ groups.
- We use a mixture of Mallows models to cluster the $N$ assessors according to how they rank the $n$ items.
- We estimate a latent ranking of the items $\boldsymbol{\rho}_c$ with its dispersion parameter $\alpha_c$ for each cluster of assessors.
- The latent augmented variables $z_1, \ldots, z_N \in \{1, \ldots, C\}$ assign each assessor to one of the clusters.

# Clustering

- Assessors cannot be assumed to form one homogeneous group, but possibly $C$ groups.
- We use a mixture of Mallows models to cluster the $N$ assessors according to how they rank the $n$ items.
- We estimate a latent ranking of the items $\boldsymbol{\rho}_c$ with its dispersion parameter $\alpha_c$ for each cluster of assessors.
- The latent augmented variables $z_1, \ldots, z_N \in \{1, \ldots, C\}$ assign each assessor to one of the clusters.

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

# Clustering

- Assessors cannot be assumed to form one homogeneous group, but possibly $C$ groups.
- We use a mixture of Mallows models to cluster the $N$ assessors according to how they rank the $n$ items.
- We estimate a latent ranking of the items $\boldsymbol{\rho}_c$ with its dispersion parameter $\alpha_c$ for each cluster of assessors.
- The latent augmented variables $z_1, \ldots, z_N \in \{1, \ldots, C\}$ assign each assessor to one of the clusters.

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

Note. Label switching is handled via post-hoc re-ordering of MCMC outputs.

# R package BayesMallows

This is just to say: the package implementing the method (with all extensions) is finally on CRAN! Pretty easy to use...

## short example:

```
library(BayesMallows)
load("../../data/valeriv/pancancer/data/Ciriello2013.Rdata")
fitMallows <- compute_mallows(rankings = R, nmc = 1.1e7,
n_clusters = 2:20, include_wcd = TRUE, logz_estimate = estimate)
plot_elbow(fitMallows)
```

**Note:** See Sørensen et al. (2019) for more details on the implementation.

# MovieLens Data

**Data characteristics:**

- $n = 200$ most rated movies
- $N = 6004$ users who rated (not equally) at least 3 movies
- each user compared an average of 30.2 movies
- rating $\rightarrow$ pairwise preferences

# MovieLens Data

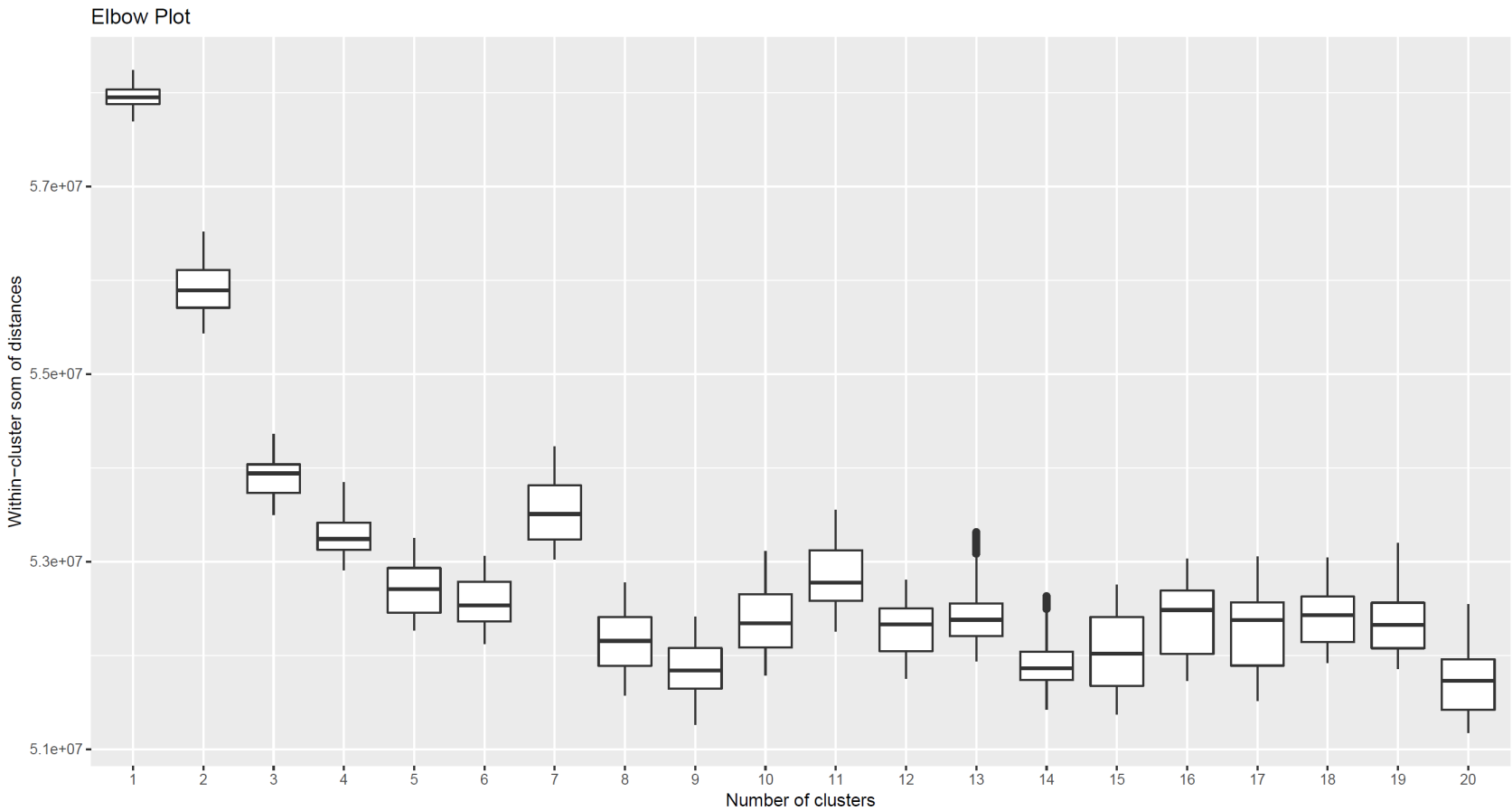
**Data characteristics:**

- $n = 200$ most rated movies
- $N = 6004$ users who rated (not equally) at least 3 movies
- each user compared an average of 30.2 movies
- rating $\rightarrow$ pairwise preferences

**Strategy:**

- **very sparse incomplete data:**
  use the Mallows model with data augmentation;
- perform **clustering:**
  cannot assume homogeneity across so many assessors!
- run for reasonable $C$ and **a posteriori decide the number of groups;**
- perform **preference prediction** with uncertainty quantification

# Choosing the right $C$ – within-cluster distance from $\rho_c$



# Preference prediction for the model with $C = 9$ clusters

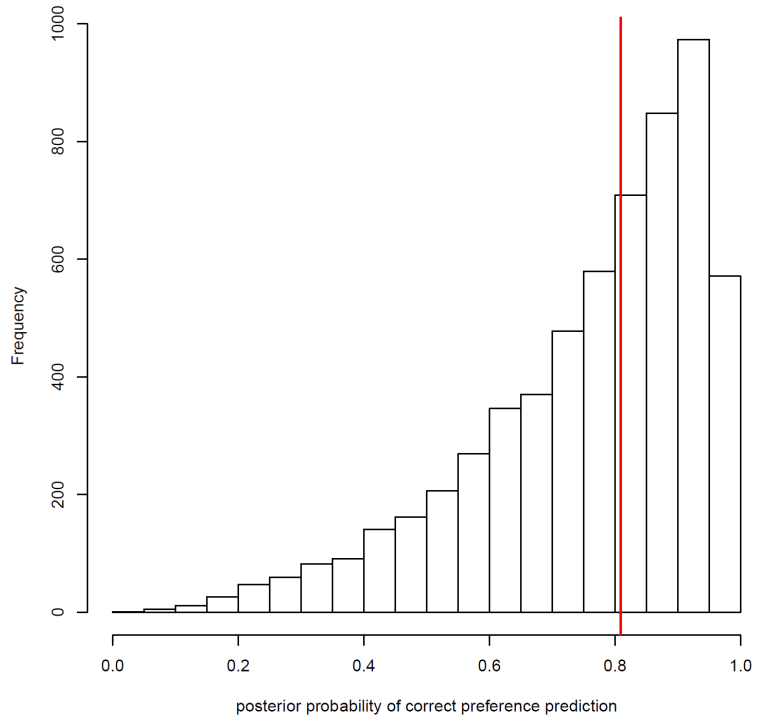
- inspect the posterior predictive probabilities $P(\tilde{\mathbf{R}}_j|\text{data})$ for each assessor

# Preference prediction for the model with $C = 9$ clusters

- inspect the posterior predictive probabilities $P(\tilde{\mathbf{R}}_j|\text{data})$ for each assessor
- compute the posterior probability of guessing the discarded preference **right**

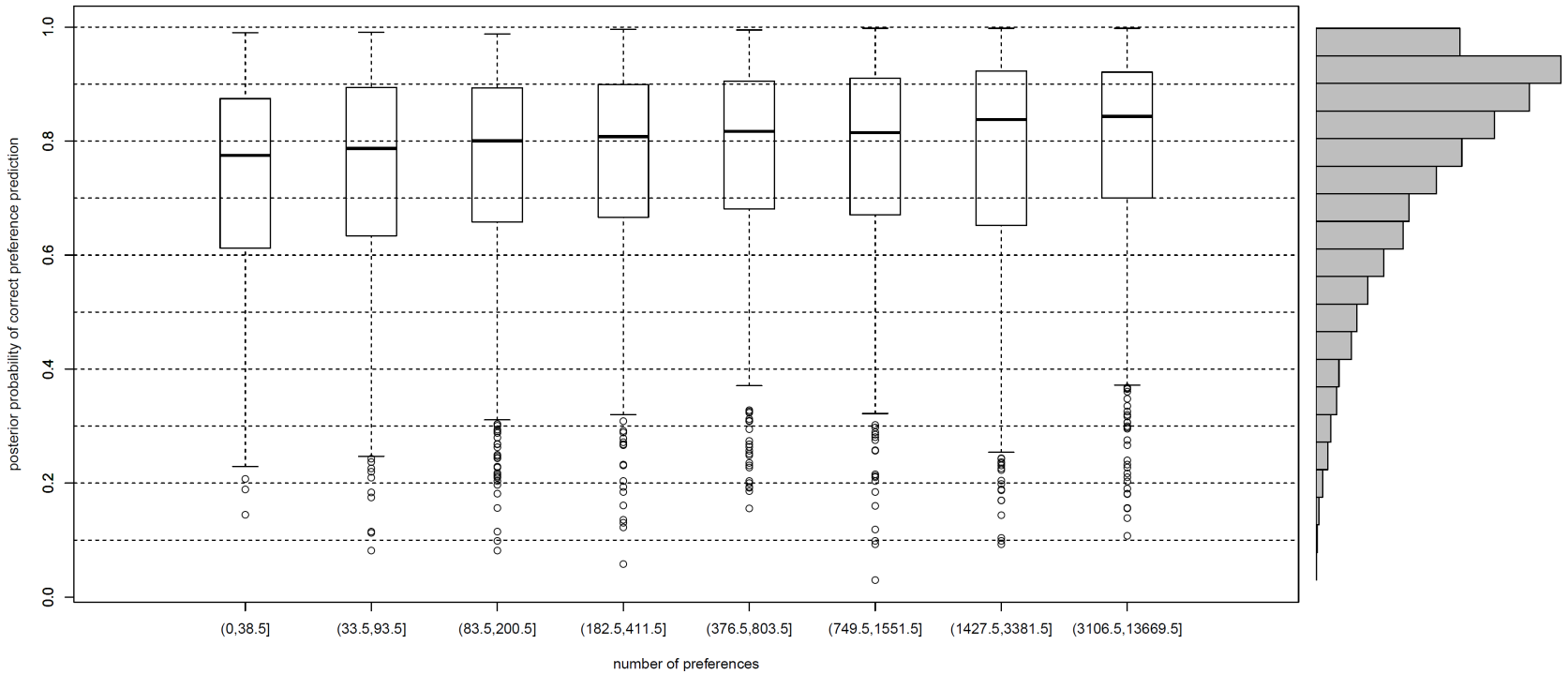
# Preference prediction for the model with $C = 9$ clusters

- inspect the posterior predictive probabilities $P(\tilde{\mathbf{R}}_j|\text{data})$ for each assessor
- compute the posterior probability of guessing the discarded preference **right**
- plot these posterior probabilities for all assessors $\rightarrow$
  the median across assessors is 0.809 for the model with 9 groups; moreover 89 % of these probabilities were higher than 0.5



# Preference prediction for the model with $C = 9$ clusters

We can also inspect the same posterior predictive probabilities stratifying on **how many preferences the assessor was giving**



# Meta-analysis of differential gene expression

**Context:**

- Studies of differential gene expression between two conditions produce a list of genes, ranked according to their level of differential expression as measured by some test statistics.

# Meta-analysis of differential gene expression

**Context:**

- Studies of differential gene expression between two conditions produce a list of genes, ranked according to their level of differential expression as measured by some test statistics.
- Little agreement among gene lists found by independent studies comparing the same conditions leads to difficulties in finding a consensus list over all available studies. This situation raises the question of whether a consensus top list over all available studies can be found.

# Meta-analysis of differential gene expression

**Context:**

- Studies of differential gene expression between two conditions produce a list of genes, ranked according to their level of differential expression as measured by some test statistics.
- Little agreement among gene lists found by independent studies comparing the same conditions leads to difficulties in finding a consensus list over all available studies. This situation raises the question of whether a consensus top list over all available studies can be found.
- Biologists are often concerned with the few most relevant genes in the specific context of the pathology, to set in place further more detailed lab experiments.

# Meta-analysis of differential gene expression

**Benchmark for meta-analysis:** five studies comparing **prostate cancer patients** with healthy controls (Dhanasekaran et al. 2001; Luo et al. 2001; Singh et al. 2002; True et al. 2006; Welsh et al. 2001). The top-25 lists from each study contained 89 genes in total.

# Meta-analysis of differential gene expression

**Benchmark for meta-analysis:** five studies comparing **prostate cancer patients** with healthy controls (Dhanasekaran et al. 2001; Luo et al. 2001; Singh et al. 2002; True et al. 2006; Welsh et al. 2001). The top-25 lists from each study contained 89 genes in total.

| Rank | MAP | $P(\rho \leq i)$ | $P(\rho \leq 10)$ | $P(\rho \leq 25)$ |
|---|---|---|---|---|
| 1 | HPN | 0.58 | 0.72 | 0.84 |
| 2 | AMACR | 0.59 | 0.69 | 0.8 |
| 3 | NME2 | 0.26 | 0.56 | 0.64 |
| 4 | GDF15 | 0.32 | 0.67 | 0.79 |
| 5 | FASN | 0.61 | 0.65 | 0.76 |
| 6 | SLC25A6 | 0.19 | 0.63 | 0.71 |
| 7 | OACT2 | 0.61 | 0.63 | 0.71 |
| 8 | UAP1 | 0.62 | 0.64 | 0.74 |
| 9 | KRT18 | 0.6 | 0.61 | 0.72 |
| 10 | EEF2 | 0.64 | 0.64 | 0.75 |

# Meta-analysis of differential gene expression

**Benchmark for meta-analysis:** five studies comparing **prostate cancer patients** with healthy controls (Dhanasekaran et al. 2001; Luo et al. 2001; Singh et al. 2002; True et al. 2006; Welsh et al. 2001). The top-25 lists from each study contained 89 genes in total.

| Rank | MAP | $P(\rho \leq i)$ | $P(\rho \leq 10)$ | $P(\rho \leq 25)$ |
|---|---|---|---|---|
| 1 | HPN | 0.58 | 0.72 | 0.84 |
| 2 | AMACR | 0.59 | 0.69 | 0.8 |
| 3 | NME2 | 0.26 | 0.56 | 0.64 |
| 4 | GDF15 | 0.32 | 0.67 | 0.79 |
| 5 | FASN | 0.61 | 0.65 | 0.76 |
| 6 | SLC25A6 | 0.19 | 0.63 | 0.71 |
| 7 | OACT2 | 0.61 | 0.63 | 0.71 |
| 8 | UAP1 | 0.62 | 0.64 | 0.74 |
| 9 | KRT18 | 0.6 | 0.61 | 0.72 |
| 10 | EEF2 | 0.64 | 0.64 | 0.75 |

- Like De Conde et al. (2006), Deng et al. (2014), and Lin and Ding (2009), HPN and AMACR are ranked first and second in the MAP consensus ranking.

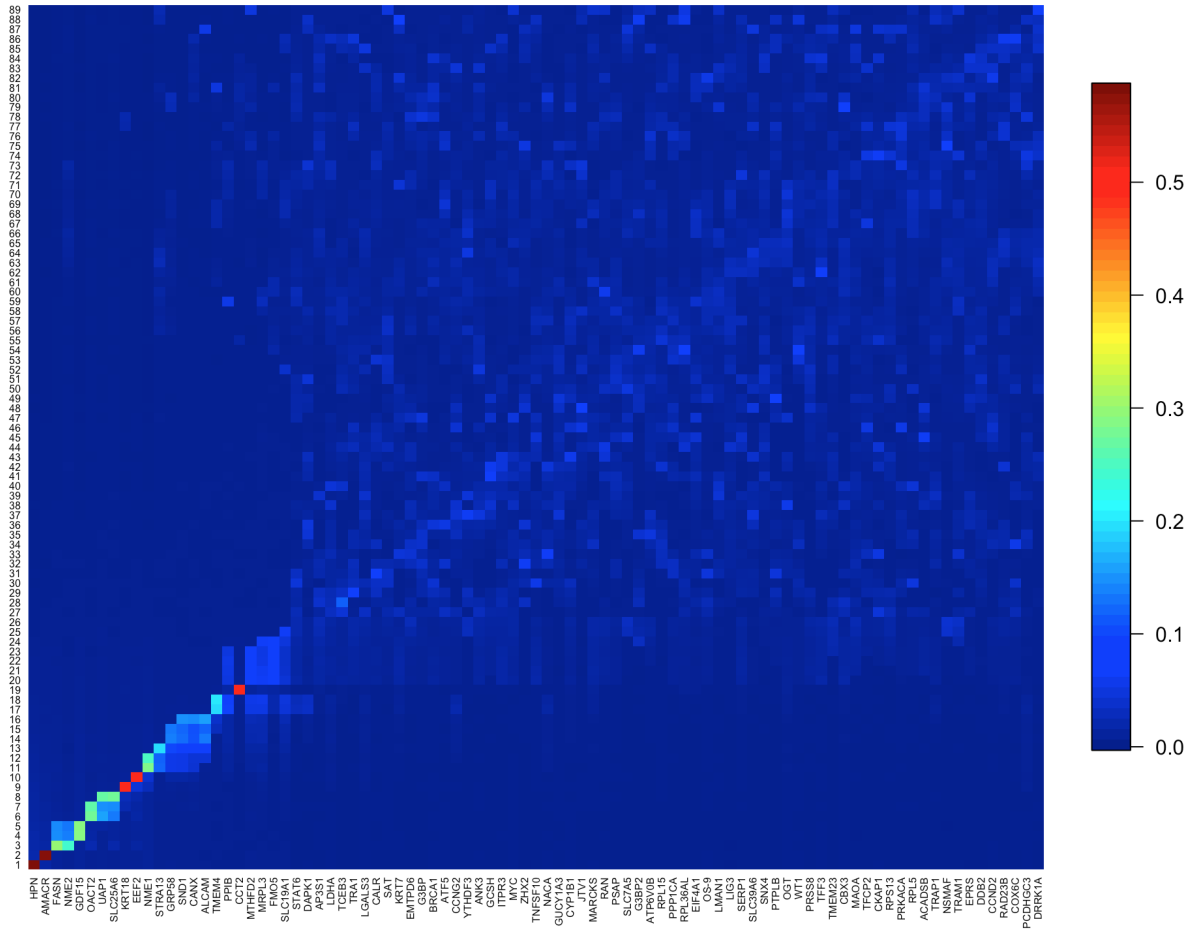
# Meta-analysis of differential gene expression

**Benchmark for meta-analysis:** five studies comparing **prostate cancer patients** with healthy controls (Dhanasekaran et al. 2001; Luo et al. 2001; Singh et al. 2002; True et al. 2006; Welsh et al. 2001). The top-25 lists from each study contained 89 genes in total.

| Rank | MAP | $P(\rho \leq i)$ | $P(\rho \leq 10)$ | $P(\rho \leq 25)$ |
|---|---|---|---|---|
| 1 | HPN | 0.58 | 0.72 | 0.84 |
| 2 | AMACR | 0.59 | 0.69 | 0.8 |
| 3 | NME2 | 0.26 | 0.56 | 0.64 |
| 4 | GDF15 | 0.32 | 0.67 | 0.79 |
| 5 | FASN | 0.61 | 0.65 | 0.76 |
| 6 | SLC25A6 | 0.19 | 0.63 | 0.71 |
| 7 | OACT2 | 0.61 | 0.63 | 0.71 |
| 8 | UAP1 | 0.62 | 0.64 | 0.74 |
| 9 | KRT18 | 0.6 | 0.61 | 0.72 |
| 10 | EEF2 | 0.64 | 0.64 | 0.75 |

- Like De Conde et al. (2006), Deng et al. (2014), and Lin and Ding (2009), HPN and AMACR are ranked first and second in the MAP consensus ranking.
- If comparing with `RankAggreg` (Pihur et al. 2009), and the aggregation methods implemented in `TopKLists` (Schimek et al. 2015), the same genes are nearly always shared between different top$-k$ lists.

# Meta-analysis of differential gene expression



# Pan-cancer analysis of TCGA data

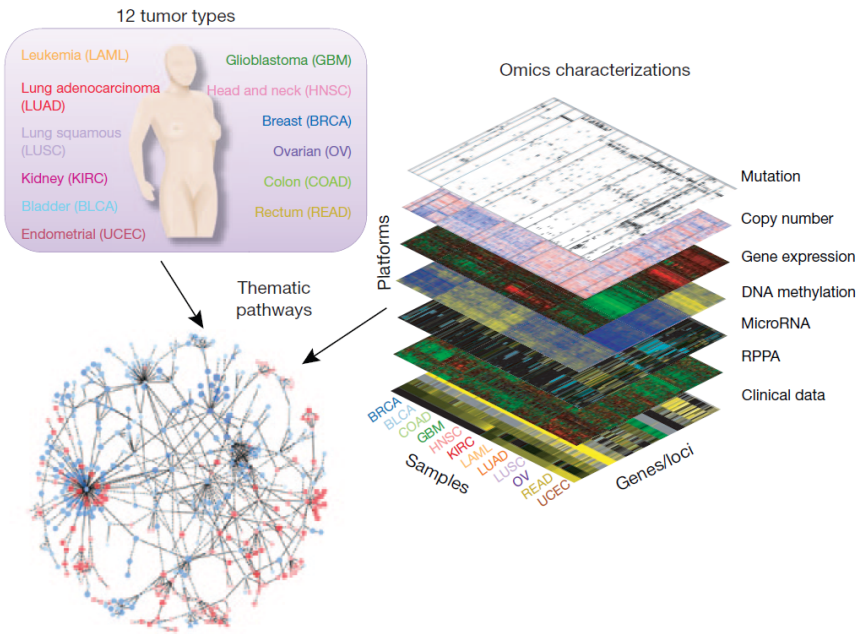
**Context:**

- High-throughput technologies allow a systematic exploration of the genetic and epigenetic basis of cancer.

# Pan-cancer analysis of TCGA data

**Context:**

- High-throughput technologies allow a systematic exploration of the genetic and epigenetic basis of cancer.
- The Cancer Genome Atlas (TCGA) started in 2006: profiling 10,000 tumour samples from 20 tumour types (`http://www.nature.com/tcga/`): multiple technical platforms; increasingly complete picture of molecular alterations in cancer.



# Pan-cancer analysis of TCGA data

**Context:**

- Hoadley et al. (2014) proposed a pan-cancer analysis on $N = 2617$ samples across 12 tumors; they perform single-platform analyses, and then combine clustering results via cluster-of-clusters method.

# Pan-cancer analysis of TCGA data

**Context:**

- Hoadley et al. (2014) proposed a pan-cancer analysis on $N = 2617$ samples across 12 tumors; they perform single-platform analyses, and then combine clustering results via cluster-of-clusters method. We use the same samples and use gene expression data (RNA-seq).

# Pan-cancer analysis of TCGA data

**Context:**

- Hoadley et al. (2014) proposed a pan-cancer analysis on $N = 2617$ samples across 12 tumors; they perform single-platform analyses, and then combine clustering results via cluster-of-clusters method. We use the same samples and use gene expression data (RNA-seq).
- Ciriello et al. (2013) also performed a pan-cancer analysis where they selected 479 functional events (based on copy numbers, mutations and methylation).

# Pan-cancer analysis of TCGA data

**Context:**

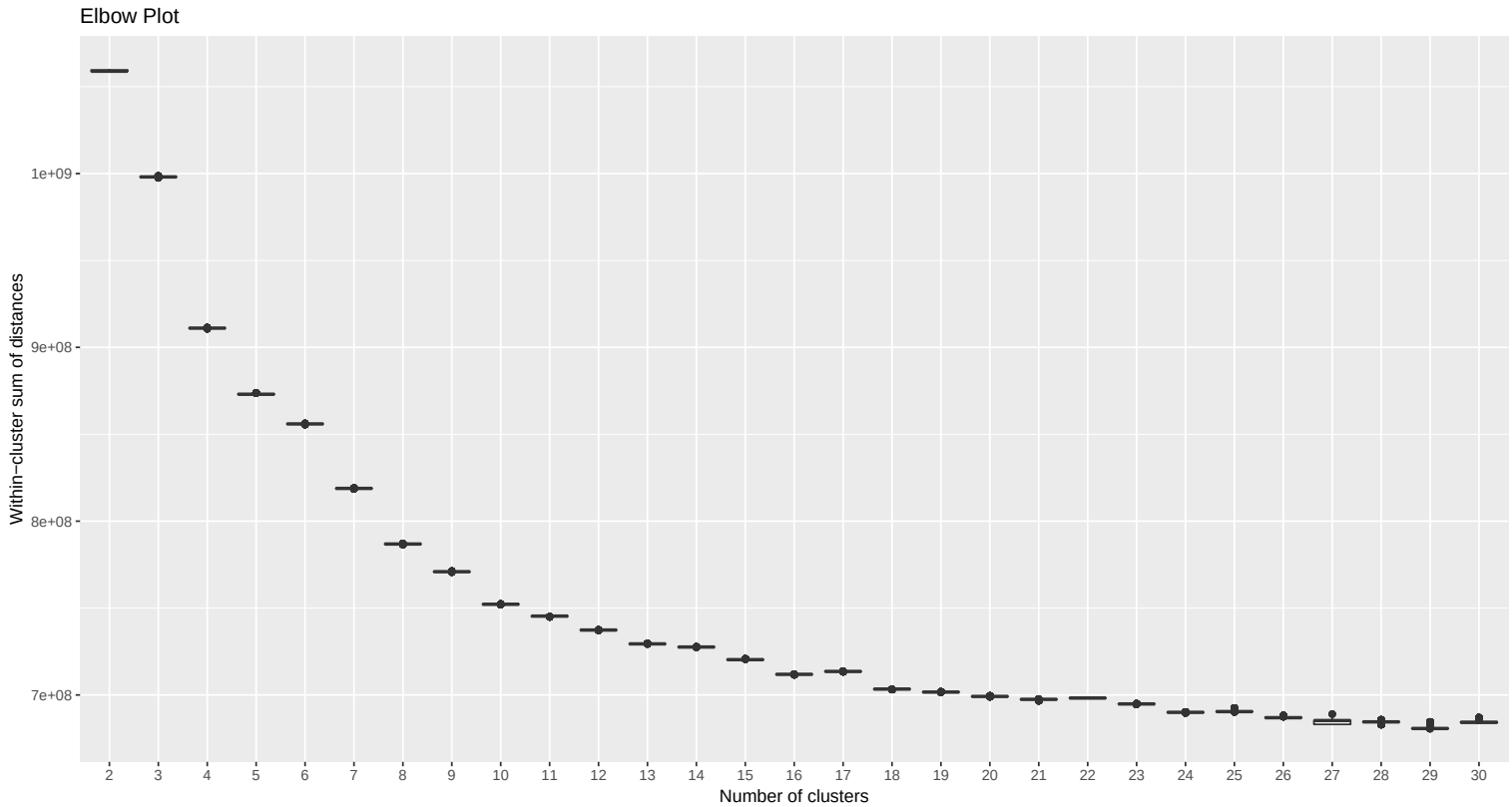
- Hoadley et al. (2014) proposed a pan-cancer analysis on $N = 2617$ samples across 12 tumors; they perform single-platform analyses, and then combine clustering results via cluster-of-clusters method. We use the same samples and use gene expression data (RNA-seq).
- Ciriello et al. (2013) also performed a pan-cancer analysis where they selected 479 functional events (based on copy numbers, mutations and methylation). We use the same selection, which covers $n = 1247$ genes.

# Pan-cancer analysis of TCGA data

**Context:**

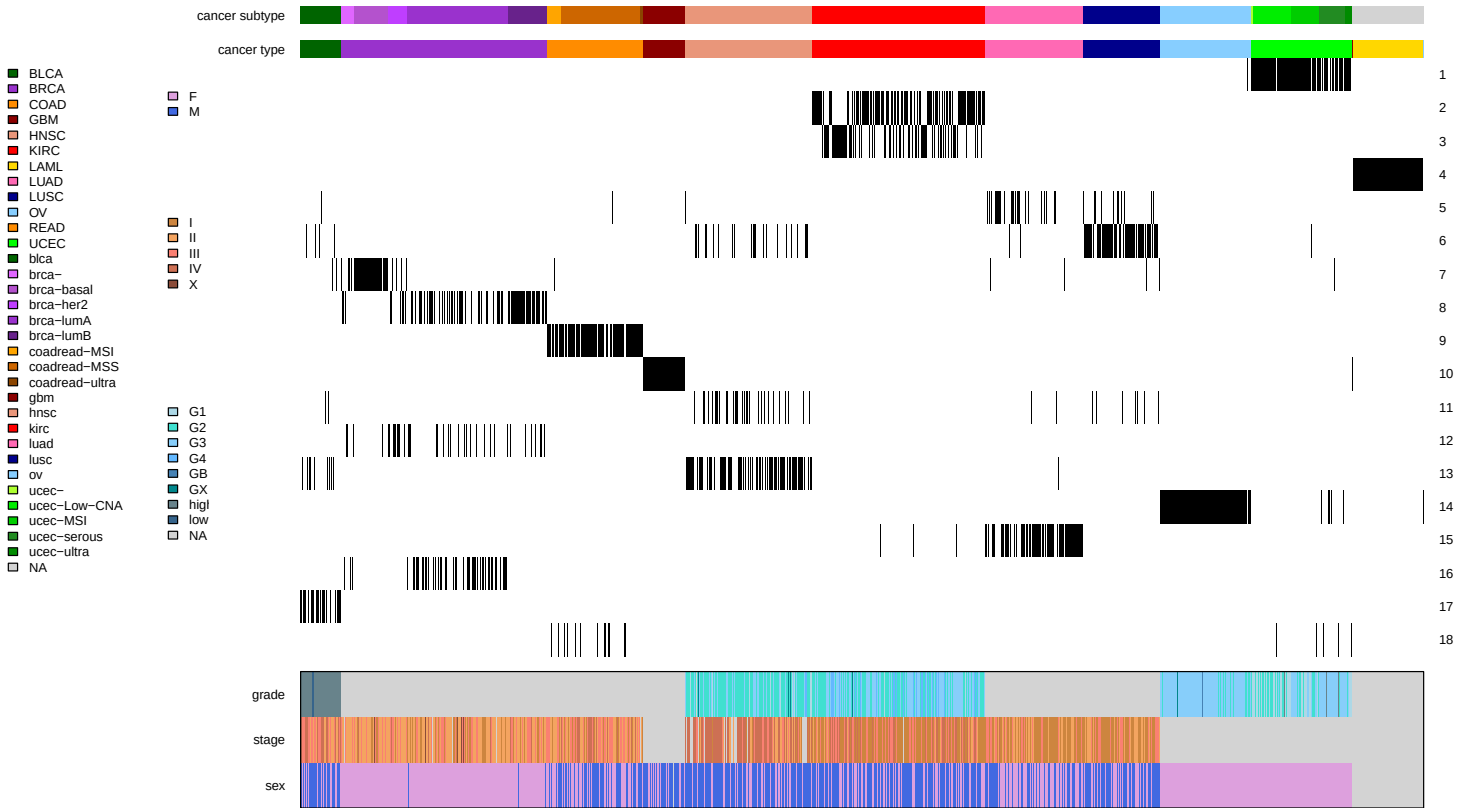
- Hoadley et al. (2014) proposed a pan-cancer analysis on $N = 2617$ samples across 12 tumors; they perform single-platform analyses, and then combine clustering results via cluster-of-clusters method. We use the same samples and use gene expression data (RNA-seq).
- Ciriello et al. (2013) also performed a pan-cancer analysis where they selected 479 functional events (based on copy numbers, mutations and methylation). We use the same selection, which covers $n = 1247$ genes.
- MCMC setting: we run `BayesMallows` for 1.1 million MCMC iterations, with 1 million burn-in. Use $L^1$ metric (footrule). No missing data; aim: clustering.

# Pan-cancer analysis of TCGA data: number of clusters



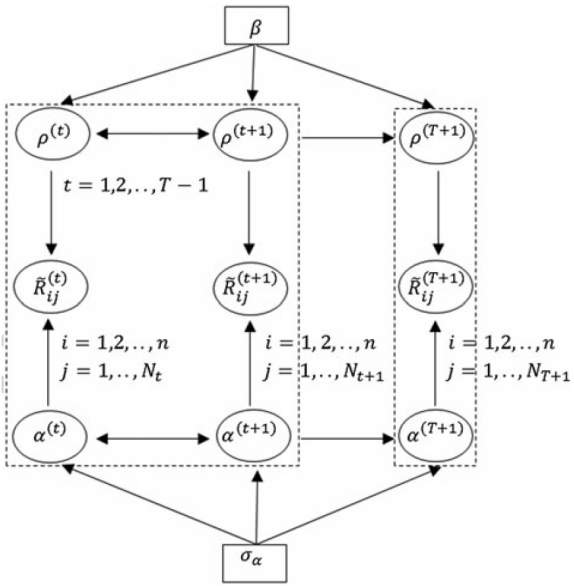

Vitelli et al. (2019)

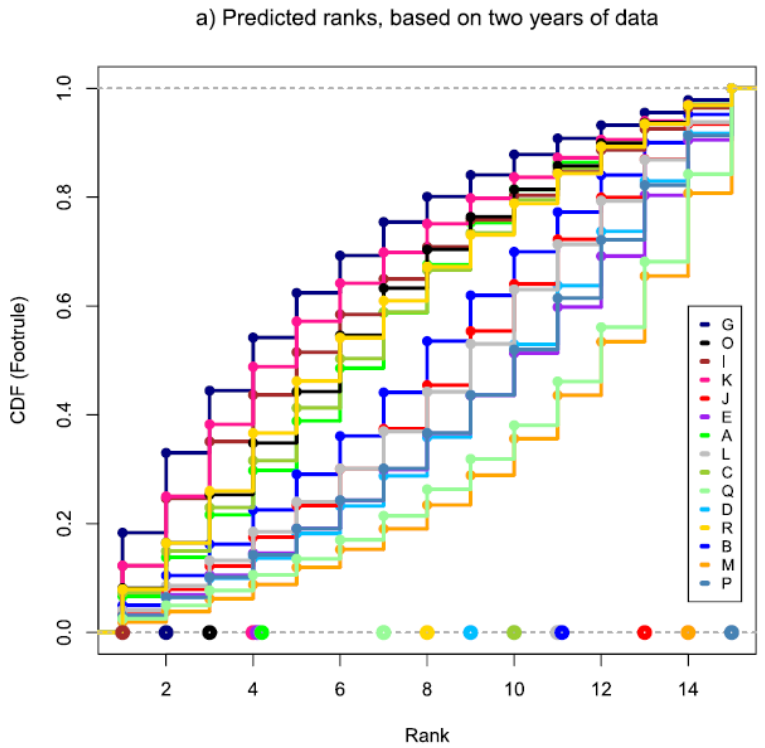
# Pan-cancer analysis of TCGA data: 18 clusters




Vitelli et al. (2019)

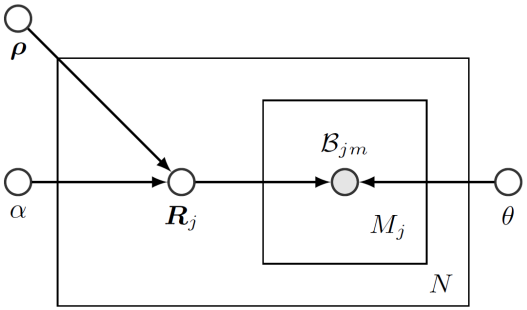
# Time dependency: students along school years

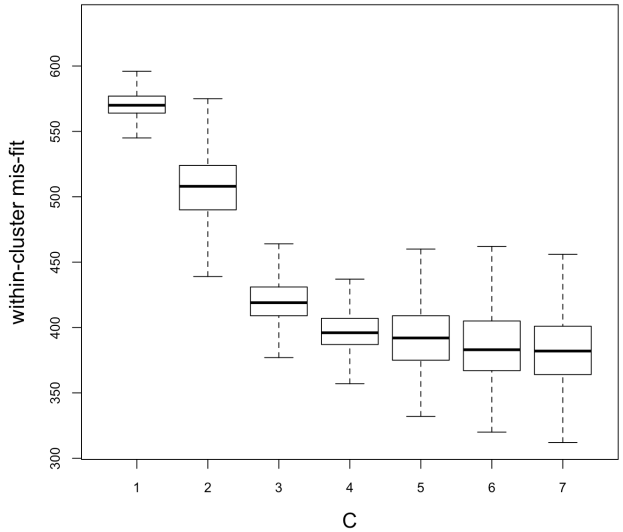


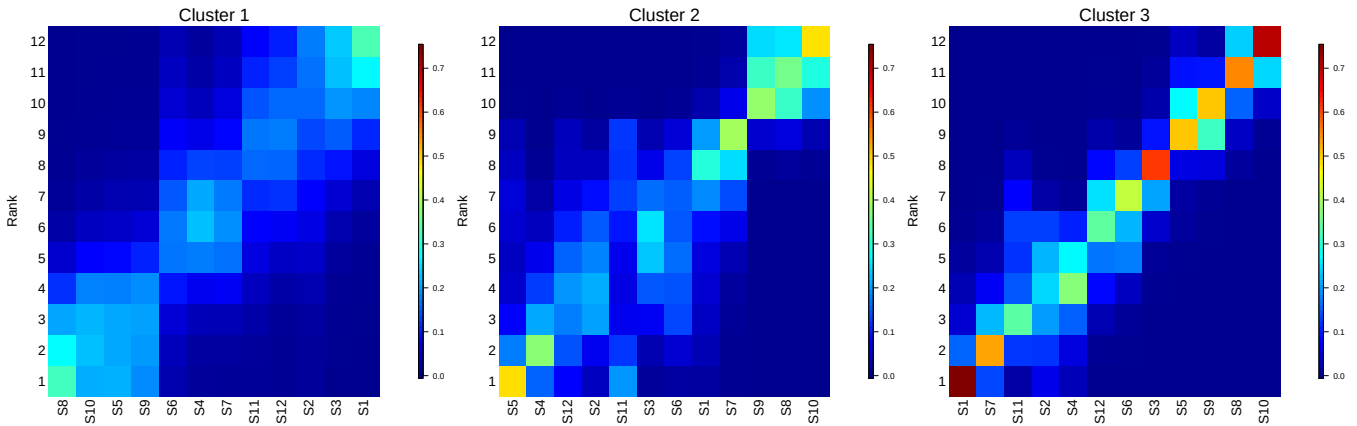




Asfaw et al. (2017)

# Inconsistent preferences: users' contradictions



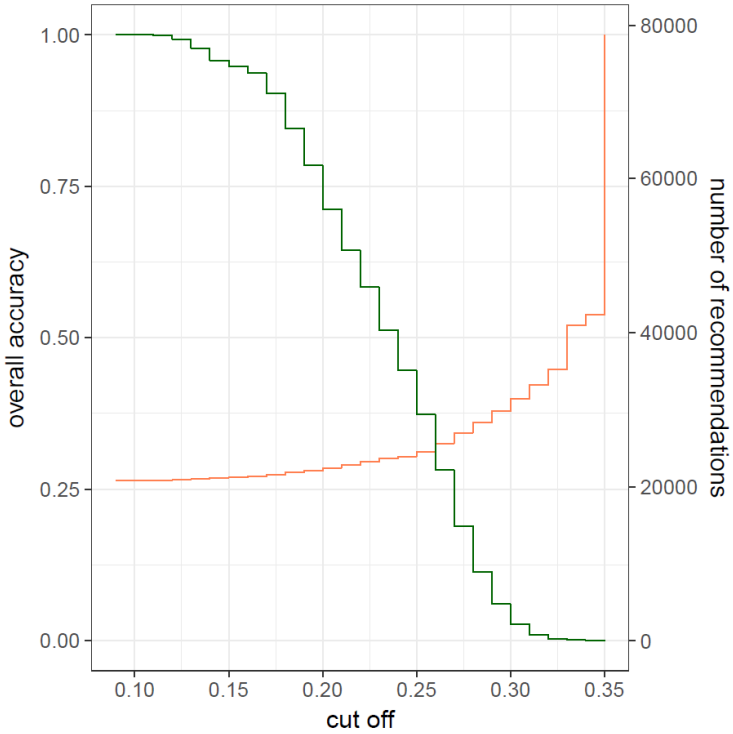




Crispino et al. (2019)

# Mallows for high-dimensional clicking data

- Clicking dataset from the Norwegian Broadcasting Company (NRK).
- Data characteristics: $n = 200$ (most popular clicks); selected users who have at least 13 clicks; randomly remove 10 clicks per user for prediction purposes.
- Uncertainty in results matters!




Liu et al. (2019b)

# Items selection

Why is this relevant?

- it reduces the dimension of the space where the ranks live (ease of computation);
- it allows for automatically selecting the items that are "relevant enough" to be included in the distance computation (better modeling);
- and it enforces a better distinction between noisy and relevant items (better interpretation).

All these aspects, I feel, have a quite substantial importance in genomics applications.

## Machine learning inspired approach

- find a quick-and-rough pre-ordering of the items (ex. Borda);
- define a sufficiently large "buffer" $L$ of items depending on how many we are interested in (ex. $L = 50$ if we aim at estimating top-20 ranks);
- run subsequent analyses on batches of $2L$ items, by fixing the first $L$ in the pre-ordering and merging them with the next $L$-dimensional batch with increasingly lower rank in the pre-ordering;
- (hopefully) prove that the posterior probability of an item way down in the ordering appearing in the top ranks is decreasing.

## Machine learning inspired approach

- find a quick-and-rough pre-ordering of the items (ex. Borda);
- define a sufficiently large "buffer" $L$ of items depending on how many we are interested in (ex. $L = 50$ if we aim at estimating top-20 ranks);
- run subsequent analyses on batches of $2L$ items, by fixing the first $L$ in the pre-ordering and merging them with the next $L$-dimensional batch with increasingly lower rank in the pre-ordering;
- (hopefully) prove that the posterior probability of an item way down in the ordering appearing in the top ranks is decreasing.

## Model-based approach

Include a variable selection step directly in the model definition, by directly defining a variable including/excluding items.

# Multi-platform analysis

Consider rankings from various genomic layers: gene expressions, copy number variation, protein levels, miRNAs, methylation, and so on.

# Multi-platform analysis

Consider rankings from various genomic layers: gene expressions, copy number variation, protein levels, miRNAs, methylation, and so on.

**Possible ways of approaching data integration in the Mallows model:**

1. estimate $\alpha$ and $\boldsymbol{\rho}$ separately for each layer, and use the unique cluster labels to link the different layers. Advantage: ease of implementation; disadvantage: weak link among the layers (sparse data).

# Multi-platform analysis

Consider rankings from various genomic layers: gene expressions, copy number variation, protein levels, miRNAs, methylation, and so on.

**Possible ways of approaching data integration in the Mallows model:**

1. estimate $\alpha$ and $\boldsymbol{\rho}$ separately for each layer, and use the unique cluster labels to link the different layers. Advantage: ease of implementation; disadvantage: weak link among the layers (sparse data).
2. use the prior to enforce a link among the layers, possibly following the underlying biology. Advantage: can follow the dogma model; disadvantage: implementation less straightforward.

# Multi-platform analysis

Consider rankings from various genomic layers: gene expressions, copy number variation, protein levels, miRNAs, methylation, and so on.
**Possible ways of approaching data integration in the Mallows model:**

1. estimate $\alpha$ and $\boldsymbol{\rho}$ separately for each layer, and use the unique cluster labels to link the different layers. Advantage: ease of implementation; disadvantage: weak link among the layers (sparse data).
2. use the prior to enforce a link among the layers, possibly following the underlying biology. Advantage: can follow the dogma model; disadvantage: implementation less straightforward.
3. identify one common ranking combining the information across layers, and then fit the Mallows model as if we had only one layer. Possibility: rank on the basis of a higher dimensional distance metric, one dimension for each layer, differently weighting the layers to model the "biological importance". Question: results sensitivity to the choice of the metric?

# Open points & directions worth exploring

- **computational aspects:** scalability, MCMC efficiency (mixing, stickiness), alternatives to MCMC, ...

# Open points & directions worth exploring

- **computational aspects:** scalability, MCMC efficiency (mixing, stickiness), alternatives to MCMC, ...
- **crucial model extensions:** infinite mixture (automatic model selection), informative prior for $\rho$ (for ex. in genomics, could include gene pathways information);

# Open points & directions worth exploring

- **computational aspects:** scalability, MCMC efficiency (mixing, stickiness), alternatives to MCMC, ...
- **crucial model extensions:** infinite mixture (automatic model selection), informative prior for $\rho$ (for ex. in genomics, could include gene pathways information);
- **other model extensions:** un-equal quality of assessors, covariates (for items & assessors), non-right invariant distances.

# Open points & directions worth exploring

- **computational aspects:** scalability, MCMC efficiency (mixing, stickiness), alternatives to MCMC, ...
- **crucial model extensions:** infinite mixture (automatic model selection), informative prior for $\rho$ (for ex. in genomics, could include gene pathways information);
- **other model extensions:** un-equal quality of assessors, covariates (for items & assessors), non-right invariant distances.

# Thanks for your attention!

# Selected References

## Original model:

Ø. Sørensen, M. Crispino, Liu, Q. and V. Vitelli, "BayesMallows: An R Package for the Bayesian Mallows Model", *arXiv:1902.08432*, 2019.

V. Vitelli*, Ø. Sørensen*, M. Crispino, A. Frigessi and E. Arjas, "Probabilistic Preference Learning for the Mallows Rank Model", *Journal of Machine Learning Research*, **18**(158), 1–49, 2018.

## Extensions:

D. Asfaw, V. Vitelli, Ø. Sørensen, E. Arjas and A. Frigessi, "Time-varying rankings with the Bayesian Mallows model", *Stat*, **6**(1), 14–30, 2017.

M. Crispino, E. Arjas, N. Barrett, V. Vitelli and A. Frigessi, "A Bayesian Mallows approach to non-transitive pair comparison data: how human are sounds?", *Accepted for publication on the Annals of Applied Statistics*, 2019.

Q. Liu*, M. Crispino*, I. Scheel, V. Vitelli and A. Frigessi, "Model-based learning from preference data", *Annual Review of Statistics and Its Application*, **6**, 329–354, 2019.

Q. Liu, A.H. Reiner, A. Frigessi and I. Scheel, "Diverse personalized recommendations with uncertainty from implicit preference data with the Bayesian Mallows Model", *arXiv:1904.03099*, 2019.

V. Vitelli, T. Fleischer, E. Arjas, V. Kristensen, A. Frigessi and M. Zucknick, "A rank-based Bayesian approach to combining genomic studies for pan-cancer applications". Manuscript in preparation, 2019.

# Other References

Andor, N., Graham, T. A., Jansen, M., Xia, L. C., Aktipis, C. A., Petritsch, C., [...] & Maley, C. C., "Pan-cancer analysis of the extent and consequences of intratumor heterogeneity". *Nature medicine*, **22**(1), 105, 2016.

Aran, D., Sirota, M., & Butte, A. J., "Systematic pan-cancer analysis of tumour purity". *Nature communications*, **6**, 8971, 2015.

Ciriello et al. "Emerging landscape of oncogenic signatures across human cancers", *Nature Genetics*, **45**, 1127–1133, 2013.

R. P. DeConde, S. Hawley, S. Falcon, N. Clegg, B. Knudsen, and R. Etzioni, "Combining results of microarray experiments: A rank aggregation approach", *Statistical Applications in Genetics and Molecular Biology*, **5**(1), Article 12, 2006.

K. Deng, S. Han, K. J. Li, and J. S. Liu, "Bayesian aggregation of order-based rank data", *Journal of the American Statistical Association*, **109**(507), 1023–1039, 2014.

S. M. Dhanasekaran, T. R. Barrette, D. Ghosh, R. Shah, S. Varambally, K. Kurachi, K. J. Pienta, M. A. Rubin, and A. M. Chinnaiyan, "Delineation of prognostic biomarkers in prostate cancer", *Nature* **412**, 822–826, 2001.

Hoadley et al. "Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin", *Cell*, **158**, 929–944, 2014.

Leiserson, M. D., Vandin, F., Wu, H. T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., [...] & Lawrence, M. S., "Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes". *Nature genetics*, **47**(2), 106, 2015.

S. Lin and J. Ding, "Integration of ranked lists via cross entropy Monte Carlo with applications to mRNA and microRNA studies", *Biometrics* **65**(1), 9–18, 2009.

# Other References

J. Luo, D.J. Duggan, Y. Chen, J. Sauvageot, C.M. Ewing, M.L. Bittner, J.M. Trent, W.B. Isaacs, "Human Prostate Cancer and Benign Prostatic Hyperplasia: Molecular Dissection by Gene Expression Profiling", *Cancer Research* **61**(12), 4683–4688, 2001.

C. L. Mallows, "Non-null ranking models", *Biometrika*, **44**(1-2), 114–130, 1957.

J.I. Marden, "Analyzing and Modeling Rank Data", *Monographs on Statistics and Applied Probability*, Chapman & Hall, 1995.

S. Mukherjee, "Estimation in exponential families on permutations", *Annals of Statistics*, 2016.

Pihur, V., Datta, S., and Datta, S., "RankAggreg, an R package for weighted rank aggregation", *BMC bioinformatics*, **10**(1), 2009.

Schimek, M., Budinská, E., Kugler, K., Švendová, V., Ding, J. and Lin, S., "TopKLists: a comprehensive R package for statistical inference, stochastic aggregation, and visualization of multiple omics ranked lists", *Statistical Applications in Genetics and Molecular Biology*, **14**(3), 311–316, 2015.

D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, W.R. Sellers, "Gene expression correlates of clinical prostate cancer behavior", *Cancer Cell*, **1**(2), 203–209, 2002.

L.L. Thurstone "A law of comparative judgment", *Psychological review*, **34**273, 1927.

L. True, I. Coleman, S. Hawley, C.Y. Huang, D. Gifford, R. Coleman, T.M. Beer, E. Gelmann, M. Datta, E. Mostaghel, B. Knudsen, P. Lange, R. Vessella, D. Lin, L. Hood, P.S. Nelson, "A molecular correlate to the Gleason grading system for prostate adenocarcinoma", *Proceedings of the National Academy of Sciences*, **103**(29), 10991–10996, 2006.

J.B. Welsh, L.M. Sapinoso, A.I. Su, S.G. Kern, J. Wang-Rodriguez, C.A. Moskaluk, H.F. Frierson, G.M. Hampton, "Analysis of Gene Expression Identifies Candidate Markers and Pharmacological Targets in Prostate Cancer", *Cancer Research*, **61**(16), 5974–5978, 2001.