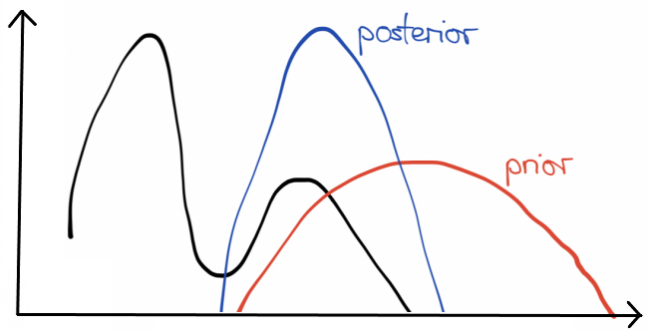


Bayesian methods for rank and preference data from recommender systems to cancer genomics

Valeria Vitelli

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

RLadies Meetup
Teknologihuset – March 19, 2019

Outline

Key collaborators

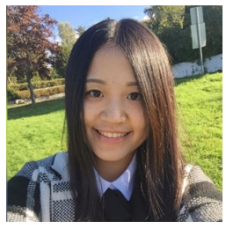
Original method:

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

Cancer Genomics:

Manuela Zucknick,
OCBE - UiO

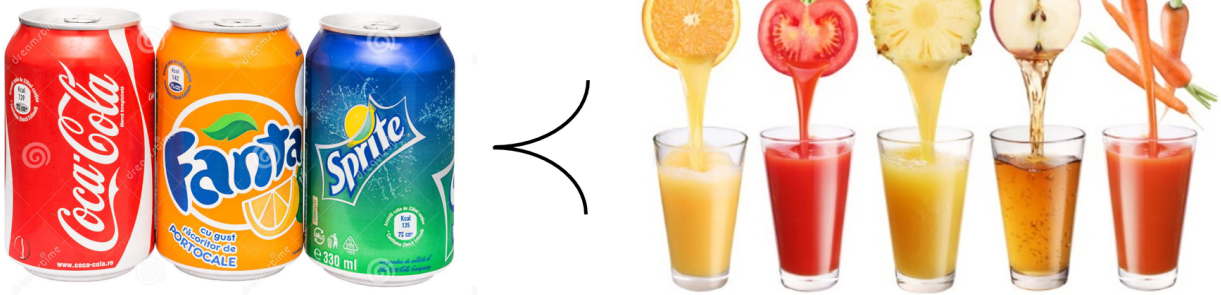
Vessela Kristensen, UiO & OUS

Thomas Fleischer, OUS

Marta Crispino,
INRIA, Grenoble, France

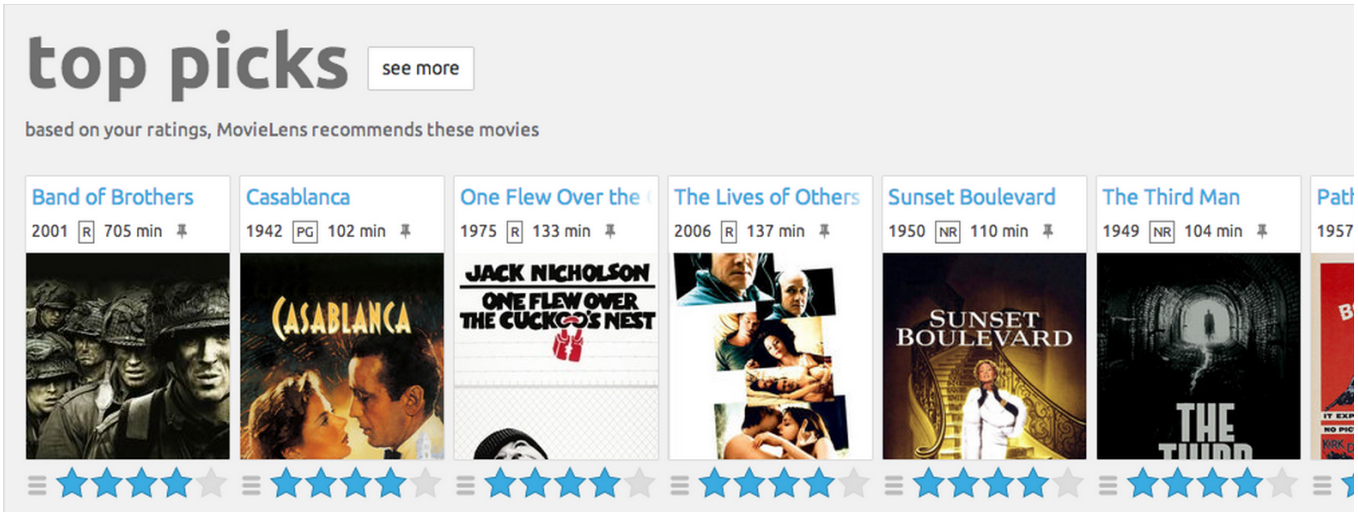
Why preference learning matters?

- customers express preferences about products and services;



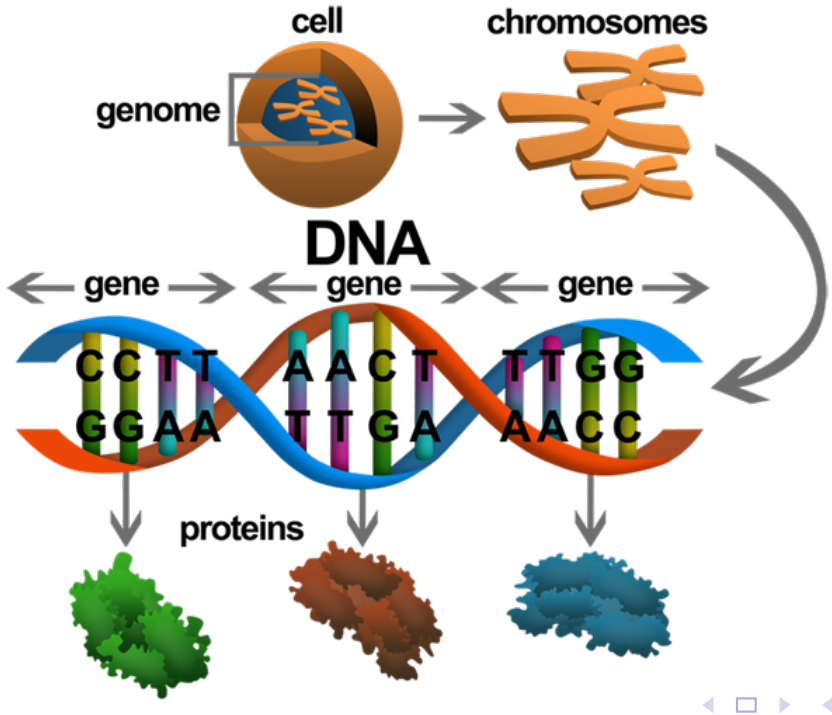
Why preference learning matters?

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



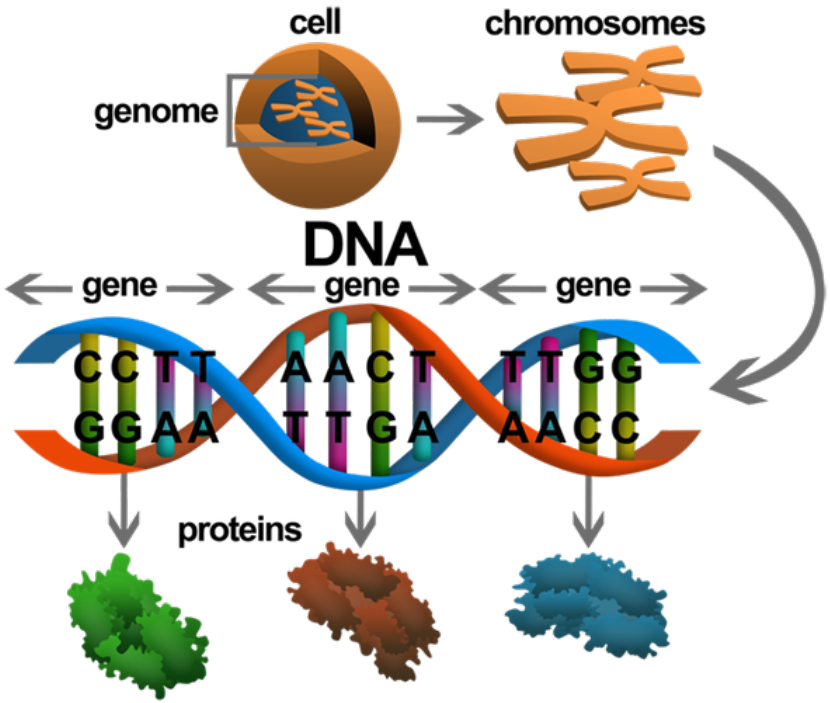
Why preference learning matters?

- 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.



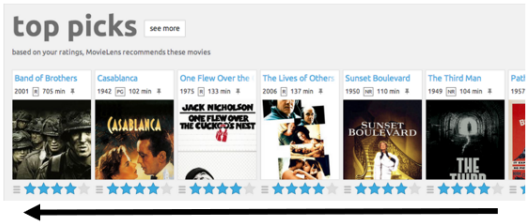
Because preference data is everywhere!

- 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.

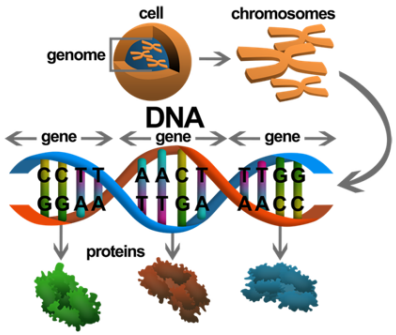


Building blocks of a preference learning framework

sugar intake / taste / healthy



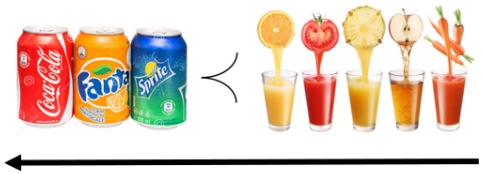
personal taste / time / popularity



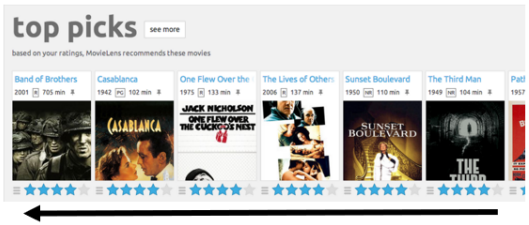
biological process / other factors

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

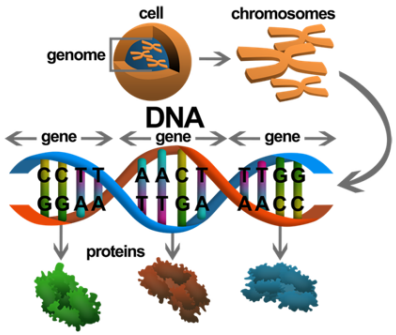
Building blocks of a preference learning framework



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? From single-cancer to pan-cancer

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

Where to use this? From single-cancer to pan-cancer

- 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.

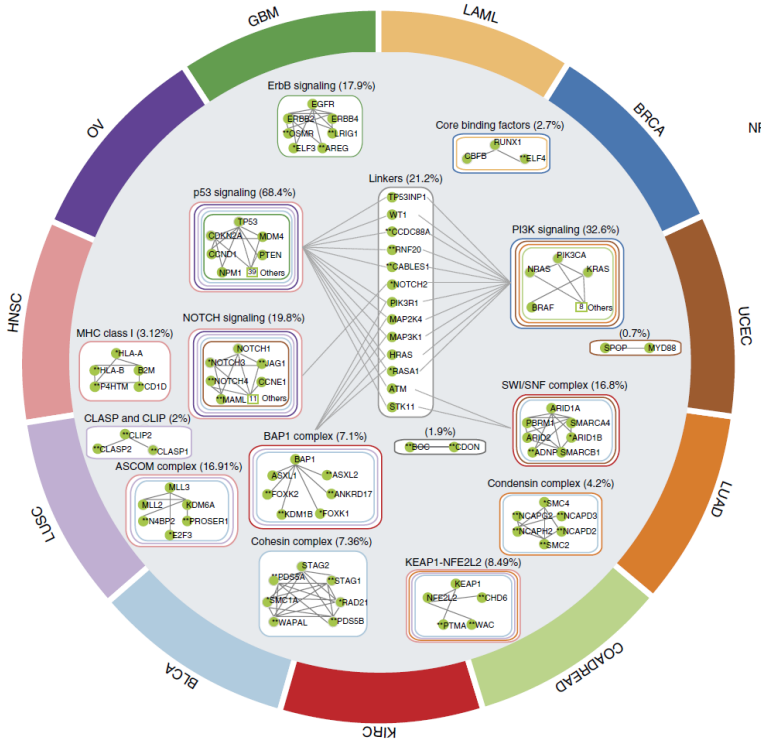
Where to use this? From single-cancer to pan-cancer

- 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

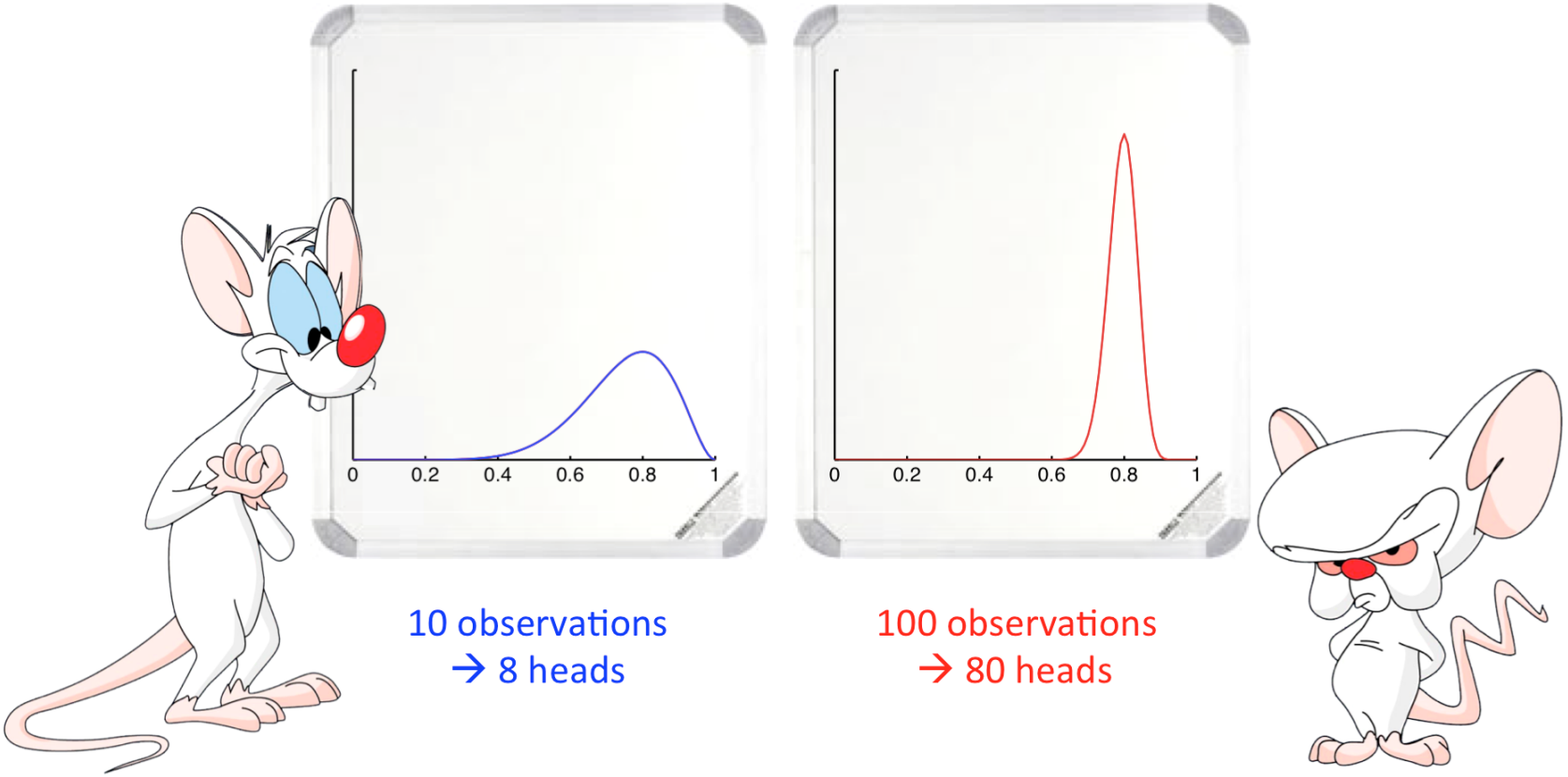
For the future:

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

Why does uncertainty matter?

One (possible) example → who are you going to believe more?

Note: point estimates are not going to help you here!



Bayesian modeling in a nutshell

Bayesian modeling in a nutshell $\rightarrow$ really fast!!

1. model the data with a probability distribution: $P_\theta(\text{data})$

Bayesian modeling in a nutshell $\rightarrow$ really fast!!

1. model the data with a probability distribution: $P_\theta(\text{data})$
   this is called likelihood, because it should be a "reasonably good" mathematical description of the data generating process;

Bayesian modeling in a nutshell $\rightarrow$ really fast!!

1. model the data with a probability distribution: $P_\theta(\text{data})$
   this is called likelihood, because it should be a "reasonably good" mathematical description of the data generating process;
   θ are the parameters in the data distribution, they are our **purpose**

Bayesian modeling in a nutshell $\rightarrow$ really fast!!

1. model the data with a probability distribution: $P_\theta(\text{data})$
   this is called likelihood, because it should be a "reasonably good" mathematical description of the data generating process;
   θ are the parameters in the data distribution, they are our **purpose**
2. decide a reasonable probabilistic distribution for θ: $P(\theta)$

Bayesian modeling in a nutshell $\rightarrow$ really fast!!

1. model the data with a probability distribution: $P_\theta(\text{data})$
   this is called likelihood, because it should be a "reasonably good" mathematical description of the data generating process;
   θ are the parameters in the data distribution, they are our **purpose**
2. decide a reasonable probabilistic distribution for θ: $P(\theta)$
   this is called prior, because it is our initial guess about parameters;
   priors can be non-informative

Bayesian modeling in a nutshell $\rightarrow$ really fast!!

1. model the data with a probability distribution: $P_\theta(\text{data})$
   this is called likelihood, because it should be a "reasonably good" mathematical description of the data generating process;
   θ are the parameters in the data distribution, they are our **purpose**
2. decide a reasonable probabilistic distribution for θ: $P(\theta)$
   this is called prior, because it is our initial guess about parameters;
   priors can be non-informative
3. **Bayes' Theorem** is our **tool** to perform inference about θ based on their *a posteriori* (i.e., *after the data*) distribution

$$P(\theta|\text{data}) \propto P_\theta(\text{data}) \cdot P(\theta)$$

Bayesian modeling in a nutshell $\rightarrow$ really fast!!

1. model the data with a probability distribution: $P_\theta(\text{data})$
   this is called likelihood, because it should be a "reasonably good" mathematical description of the data generating process;
   θ are the parameters in the data distribution, they are our **purpose**
2. decide a reasonable probabilistic distribution for θ: $P(\theta)$
   this is called prior, because it is our initial guess about parameters;
   priors can be non-informative
3. **Bayes' Theorem** is our **tool** to perform inference about θ based on their *a posteriori* (i.e., *after the data*) distribution

$$P(\theta|\text{data}) \propto P_\theta(\text{data}) \cdot P(\theta)$$

Remarks

- **Colors matter:** posterior is (really!) a combination of prior and likelihood (likelihood counts more in big data problems), and

Bayesian modeling in a nutshell $\rightarrow$ really fast!!

1. model the data with a probability distribution: $P_\theta(\text{data})$
   this is called likelihood, because it should be a "reasonably good" mathematical description of the data generating process;
   θ are the parameters in the data distribution, they are our **purpose**
2. decide a reasonable probabilistic distribution for θ: $P(\theta)$
   this is called prior, because it is our initial guess about parameters;
   priors can be non-informative
3. **Bayes' Theorem** is our **tool** to perform inference about θ based on their *a posteriori* (i.e., *after the data*) distribution

$$P(\theta|\text{data}) \propto P_\theta(\text{data}) \cdot P(\theta)$$

Remarks

- **Colors matter:** posterior is (really!) a combination of prior and likelihood (likelihood counts more in big data problems), and
- prior and posterior are **dynamic concepts:**
  "the posterior of today is the prior of tomorrow" (cit. myself).

Ingredients for our specific Bayesian model

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

Ingredients for our specific Bayesian model

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

2 **1** **3**

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

Inferential tasks

1. estimate the **consensus ranking** across the assessors, to 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

?
?
?

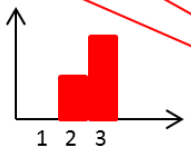
Inferential tasks

2 in case of **incomplete data** (partial rankings, preferences, ...)

Inferential tasks

2 in case of **incomplete data** (partial rankings, preferences, ...)
predict the ranks of the missing items (individually for each assessor)

1	1	1	3	2	3	3
2	3	?	1	1	?	1
3	2	?	2	3	?	2



?
?
?

uncertainty quantification

Inferential tasks

3 partition the assessors into **classes**, each sharing a consensus ranking of the items,

Inferential tasks

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

1	1	1	3	2	3	3
2	3	2	1	1	2	1
3	2	3	2	3	1	2

Minimal notation

- N is the number of assessors

Minimal notation

- N is the number of assessors and n is the number of items

Minimal notation

- N is the number of assessors and n is the number of items
- let's focus on **complete data:**

Minimal notation

- N is the number of assessors and n is the number of items
- let's focus on **complete data:** this means that each assessor provides a ranking of all the items (very unrealistic!!)

Minimal notation

- N is the number of assessors and n is the number of items
- let's focus on **complete data:** this means that each assessor provides a ranking of all the items (very unrealistic!!)
- call $\mathbf{R}_j$ the ranking given by the j-th assessor to the full set of n items;

Minimal notation

- N is the number of assessors and n is the number of items
- let's focus on **complete data:** this means that each assessor provides a ranking of all the items (very unrealistic!!)
- call $\mathbf{R}_j$ the ranking given by the j-th assessor to the full set of n items; this means that $\mathbf{R}_j$ is a permutation vector of the first n integers $(1, \ldots, n)$

Minimal notation

- N is the number of assessors and n is the number of items
- let's focus on **complete data:** this means that each assessor provides a ranking of all the items (very unrealistic!!)
- call $\mathbf{R}_j$ the ranking given by the j-th assessor to the full set of n items; this means that $\mathbf{R}_j$ is a permutation vector of the first n integers $(1, \ldots, n)$
- in this setting, the only possible task is **rank aggregation,** i.e. finding a shared consensus ranking (and associated variability)

Minimal notation

- N is the number of assessors and n is the number of items
- let's focus on **complete data:** this means that each assessor provides a ranking of all the items (very unrealistic!!)
- call $\mathbf{R}_j$ the ranking given by the j-th assessor to the full set of n items; this means that $\mathbf{R}_j$ is a permutation vector of the first n 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)

Gives a distribution for a ranking $\mathbf{R}$ as

$$P_{\alpha,\boldsymbol{\rho}}(\mathbf{R}) = \frac{1}{Z_n(\alpha,\boldsymbol{\rho})} \exp\{-(\alpha/n)d(\mathbf{R},\boldsymbol{\rho})\},$$

where:

Mallows model (Mallows 1957)

Gives a distribution for a ranking $\mathbf{R}$ as

$$P_{\alpha,\boldsymbol{\rho}}(\mathbf{R}) = \frac{1}{Z_n(\alpha,\boldsymbol{\rho})} \exp\{-(\alpha/n)d(\mathbf{R},\boldsymbol{\rho})\},$$

where:

- $\boldsymbol{\rho}$ is the consensus ranking,

Mallows model (Mallows 1957)

Gives a distribution for a ranking $\mathbf{R}$ as

$$P_{\alpha,\boldsymbol{\rho}}(\mathbf{R}) = \frac{1}{Z_n(\alpha,\boldsymbol{\rho})} \exp\{-(\alpha/n) d(\mathbf{R},\boldsymbol{\rho})\},$$

where:

- $\boldsymbol{\rho}$ is the consensus ranking,
- α is a positive parameter describing the data variability,

Mallows model (Mallows 1957)

Gives a distribution for a ranking $\mathbf{R}$ as

$$P_{\alpha,\boldsymbol{\rho}}(\mathbf{R}) = \frac{1}{Z_n(\alpha,\boldsymbol{\rho})} \exp\{-(\alpha/n) d(\mathbf{R},\boldsymbol{\rho})\},$$

where:

- $\boldsymbol{\rho}$ is the consensus ranking,
- α is a positive parameter describing the data variability,
- $d(\cdot,\cdot)$ is a distance measure between rankings,

Mallows model (Mallows 1957)

Gives a distribution for a ranking $\mathbf{R}$ as

$$P_{\alpha,\boldsymbol{\rho}}(\mathbf{R}) = \frac{1}{Z_n(\alpha,\boldsymbol{\rho})} \exp\{-(\alpha/n) d(\mathbf{R},\boldsymbol{\rho})\},$$

where:

- $\boldsymbol{\rho}$ is the consensus ranking,
- α is a positive parameter describing the data variability,
- $d(\cdot,\cdot)$ is a distance measure between rankings,
- $Z_n(\alpha,\boldsymbol{\rho})$ is the normalizing constant of the model.

Mallows model (Mallows 1957)

Gives a distribution for a ranking $\mathbf{R}$ as

$$P_{\alpha,\boldsymbol{\rho}}(\mathbf{R}) = \frac{1}{Z_n(\alpha,\boldsymbol{\rho})} \exp\{-(\alpha/n) d(\mathbf{R},\boldsymbol{\rho})\},$$

$d(\cdot,\cdot)$ deserves some more words

Mallows model (Mallows 1957)

Gives a distribution for a ranking $\mathbf{R}$ as

$$P_{\alpha,\boldsymbol{\rho}}(\mathbf{R}) = \frac{1}{Z_n(\alpha,\boldsymbol{\rho})} \exp\{-(\alpha/n) d(\mathbf{R},\boldsymbol{\rho})\},$$

$d(\cdot,\cdot)$ deserves some more words

- it gives the model much flexibility in describing reality;

Mallows model (Mallows 1957)

Gives a distribution for a ranking $\mathbf{R}$ as

$$P_{\alpha,\boldsymbol{\rho}}(\mathbf{R}) = \frac{1}{Z_n(\alpha,\boldsymbol{\rho})} \exp\{-(\alpha/n) d(\mathbf{R},\boldsymbol{\rho})\},$$

$d(\cdot,\cdot)$ deserves some more words

- it gives the model much flexibility in describing reality;
- it is crucial for the analysis, since it influences results;

Mallows model (Mallows 1957)

Gives a distribution for a ranking $\mathbf{R}$ as

$$P_{\alpha,\boldsymbol{\rho}}(\mathbf{R}) = \frac{1}{Z_n(\alpha,\boldsymbol{\rho})} \exp\{-(\alpha/n) d(\mathbf{R},\boldsymbol{\rho})\},$$

$d(\cdot,\cdot)$ deserves some more words

- it gives the model much flexibility in describing reality;
- it is crucial for the analysis, since it influences results;
- it has an impact on computation effciency, too;

Mallows model (Mallows 1957)

Gives a distribution for a ranking $\mathbf{R}$ as

$$P_{\alpha,\boldsymbol{\rho}}(\mathbf{R}) = \frac{1}{Z_n(\alpha,\boldsymbol{\rho})} \exp\{-(\alpha/n) d(\mathbf{R},\boldsymbol{\rho})\},$$

$d(\cdot,\cdot)$ deserves some more words

- it gives the model much flexibility in describing reality;
- it is crucial for the analysis, since it influences results;
- it has an impact on computation effciency, too;
- many struggle with it: good challenge for research!

A Bayesian model for recommender systems

Remember that our data are N complete rankings $\mathbf{R}_1, \ldots, \mathbf{R}_N$ (describing the assessors' preferences about the items)

A Bayesian model for recommender systems

Remember that our data are N complete rankings $\mathbf{R}_1, \ldots, \mathbf{R}_N$ (describing the assessors' preferences about the items)

- Assume that the data follow the Mallows model. Then, the likelihood is

$$P_{\alpha,\boldsymbol{\rho}}(\mathbf{R}_1, \ldots, \mathbf{R}_N) = Z_n(\alpha)^{-N} \exp\left\{ \frac{-\alpha}{n} \sum_{j=1}^{N} d(\mathbf{R}_j, \boldsymbol{\rho}) \right\}$$

A Bayesian model for recommender systems

Remember that our data are N complete rankings $\mathbf{R}_1, \ldots, \mathbf{R}_N$ (describing the assessors' preferences about the items)

- Assume that the data follow the Mallows model. Then, the likelihood is

$$P_{\alpha,\boldsymbol{\rho}}(\mathbf{R}_1, \ldots, \mathbf{R}_N) = Z_n(\alpha)^{-N} \exp\left\{ \frac{-\alpha}{n} \sum_{j=1}^{N} d(\mathbf{R}_j, \boldsymbol{\rho}) \right\}$$

- Priors: $\boldsymbol{\rho}$ is a priori uniform, since often no prior information is available;

A Bayesian model for recommender systems

Remember that our data are N complete rankings $\mathbf{R}_1, \ldots, \mathbf{R}_N$ (describing the assessors' preferences about the items)

- Assume that the data follow the Mallows model. Then, the likelihood is

$$P_{\alpha,\boldsymbol{\rho}}(\mathbf{R}_1, \ldots, \mathbf{R}_N) = Z_n(\alpha)^{-N} \exp\left\{ \frac{-\alpha}{n} \sum_{j=1}^{N} d(\mathbf{R}_j, \boldsymbol{\rho}) \right\}$$

- Priors: $\boldsymbol{\rho}$ is a priori uniform, since often no prior information is available; α gets a truncated exponential prior, since this is a reasonable choice to a priori ensure a good dispersion around the consensus ranking.

A Bayesian model for recommender systems

Remember that our data are N complete rankings $\mathbf{R}_1, \ldots, \mathbf{R}_N$ (describing the assessors' preferences about the items)

- Assume that the data follow the Mallows model. Then, the likelihood is

$$P_{\alpha,\boldsymbol{\rho}}(\mathbf{R}_1, \ldots, \mathbf{R}_N) = Z_n(\alpha)^{-N} \exp\left\{ \frac{-\alpha}{n} \sum_{j=1}^{N} d(\mathbf{R}_j, \boldsymbol{\rho}) \right\}$$

- Priors: $\boldsymbol{\rho}$ is a priori uniform, since often no prior information is available; α gets a truncated exponential prior, since this is a reasonable choice to a priori ensure a good dispersion around the consensus ranking.
- Probabilistic statements about $\boldsymbol{\rho}$ and α are based on the posterior distribution below, which can be targeted via MCMC algorithms

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

This framework was first described in (Vitelli et al. 2018).

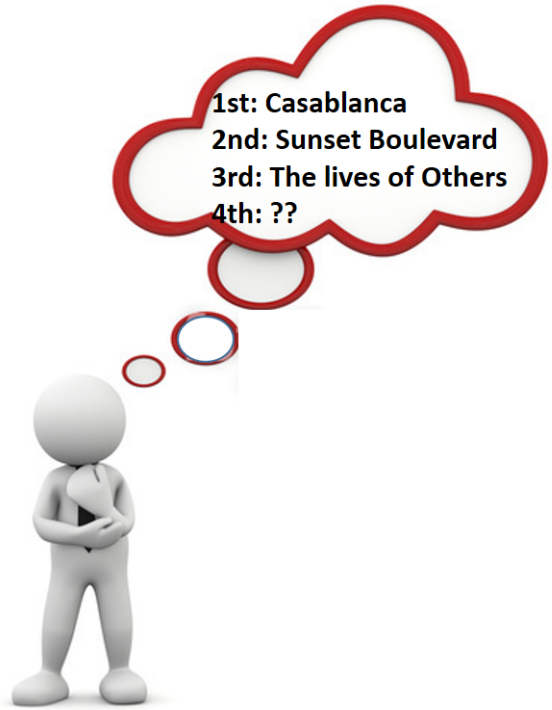
Incomplete data

This is a very common situation in the applications, especially if n is large. Possible situations:

Incomplete data

This is a very common situation in the applications, especially if n is large. Possible situations:

1. **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.



Incomplete data

This is a very common situation in the applications, especially if n is large. Possible situations:

1. **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.
2. 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).

Incomplete data

This is a very common situation in the applications, especially if n is large. Possible situations:

1. **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.
2. 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).

Incomplete data

Bayesian modeling provides an easy solution: **data augmentation.**

Incomplete data

Bayesian modeling provides an easy solution: **data augmentation.**

Skecth of the MCMC algorithm in the complete/incomplete data case:

1. **complete data MCMC:** estimate $\boldsymbol{\rho} \leftrightarrow$ estimate α;

Incomplete data

Bayesian modeling provides an easy solution: **data augmentation.**

Skecth of the MCMC algorithm in the complete/incomplete data case:

1. **complete data MCMC:** estimate $\boldsymbol{\rho} \leftrightarrow$ estimate α;
2. **incomplete data MCMC:**

   complete data MCMC $\leftrightarrow$ data augmentation

Incomplete data

Bayesian modeling provides an easy solution: **data augmentation.**

Skecth of the MCMC algorithm in the complete/incomplete data case:

1. **complete data MCMC:** estimate $\boldsymbol{\rho} \leftrightarrow$ estimate α;
2. **incomplete data MCMC:**

   complete data MCMC $\leftrightarrow$ data augmentation

Note: this procedure can be iterated

Incomplete data

Bayesian modeling provides an easy solution: **data augmentation.**

Skecth of the MCMC algorithm in the complete/incomplete data case:

1. **complete data MCMC:** estimate $\boldsymbol{\rho} \leftrightarrow$ estimate α;
2. **incomplete data MCMC:**

   complete data MCMC $\leftrightarrow$ data augmentation

Note: this procedure can be iterated

Example: skecth of a new MCMC algorithm for items selection:

3. **new MCMC:**

   incomplete data MCMC $\leftrightarrow$ new modeling tools

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.

1. We use a **mixture of Mallows models** to cluster the N assessors according to how they rank the n items.

$$P_{\boldsymbol{\rho}_c,\alpha_c,\mathbf{z}}(\mathbf{R}_1,\ldots,\mathbf{R}_N) = \prod_{j=1}^{N} Z_n(\alpha_{z_j}) \exp\left\{\frac{-\alpha_{z_j}}{n} d(\mathbf{R}_j,\boldsymbol{\rho}_{z_j})\right\}$$

Clustering

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

1. We use a **mixture of Mallows models** to cluster the N assessors according to how they rank the n items.
2. We estimate a latent ranking of the items ρ_c with its dispersion parameter α_c for each cluster of assessors.

$$P_{\rho_c,\alpha_c,\mathbf{z}}(\mathbf{R}_1,\ldots,\mathbf{R}_N) = \prod_{j=1}^{N} Z_n(\alpha_{z_j}) \exp\left\{\frac{-\alpha_{z_j}}{n} d(\mathbf{R}_j,\rho_{z_j})\right\}$$

Clustering

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

1. We use a **mixture of Mallows models** to cluster the N assessors according to how they rank the n items.
2. We estimate a latent ranking of the items $\boldsymbol{\rho}_c$ with its dispersion parameter α_c for each cluster of assessors.
3. The latent augmented variables $\mathbf{z} = (z_1, \ldots, z_N)$ assign each assessor to one of the clusters.

$$P_{\boldsymbol{\rho}_c, \alpha_c, \mathbf{z}}(\mathbf{R}_1, \ldots, \mathbf{R}_N) = \prod_{j=1}^{N} Z_n(\alpha_{z_j}) \exp\left\{ \frac{-\alpha_{z_j}}{n} d(\mathbf{R}_j, \boldsymbol{\rho}_{z_j}) \right\}$$

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 also 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

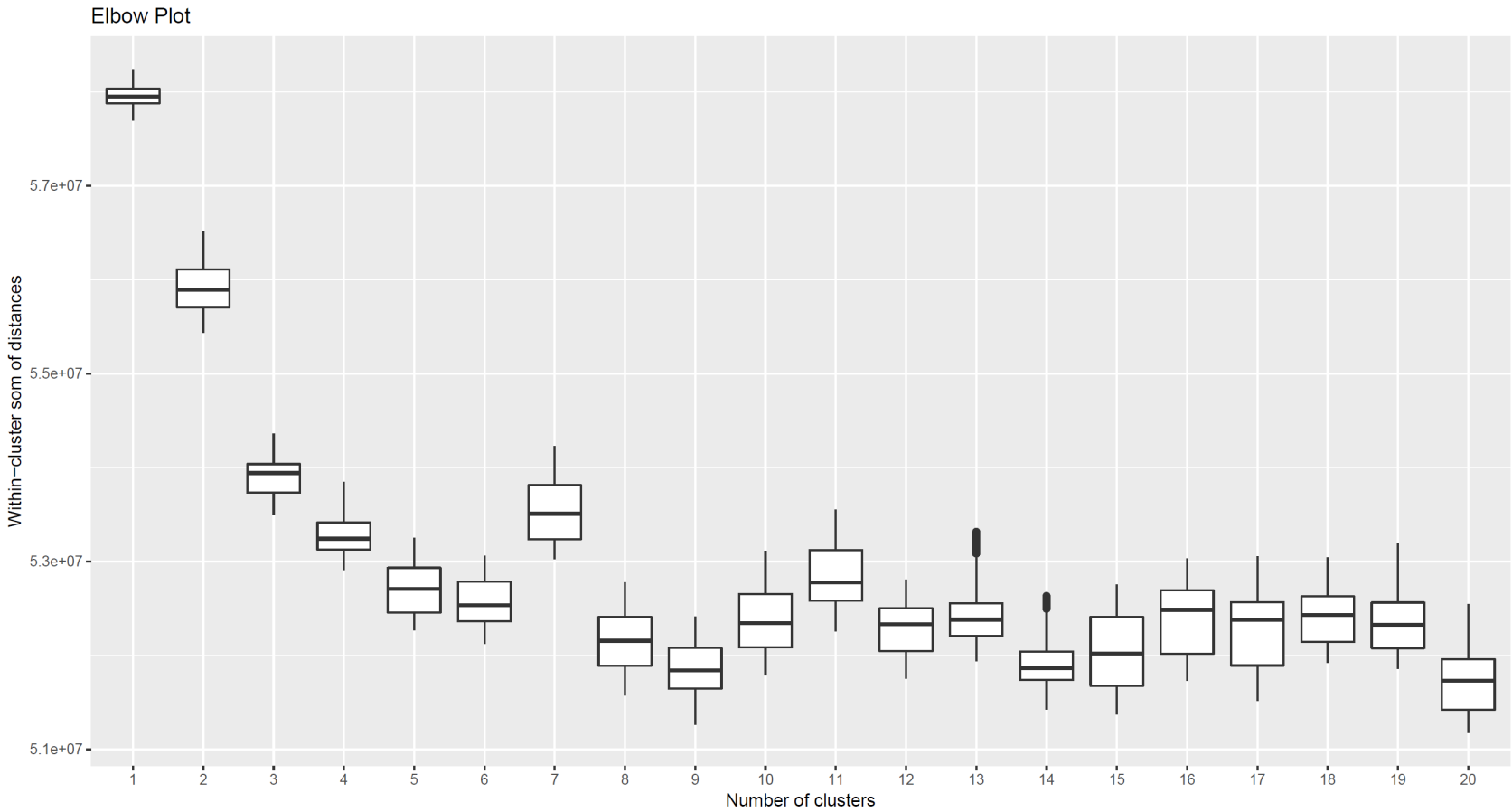
Data characteristics:

- $n = 200$ 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 ρ_c



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

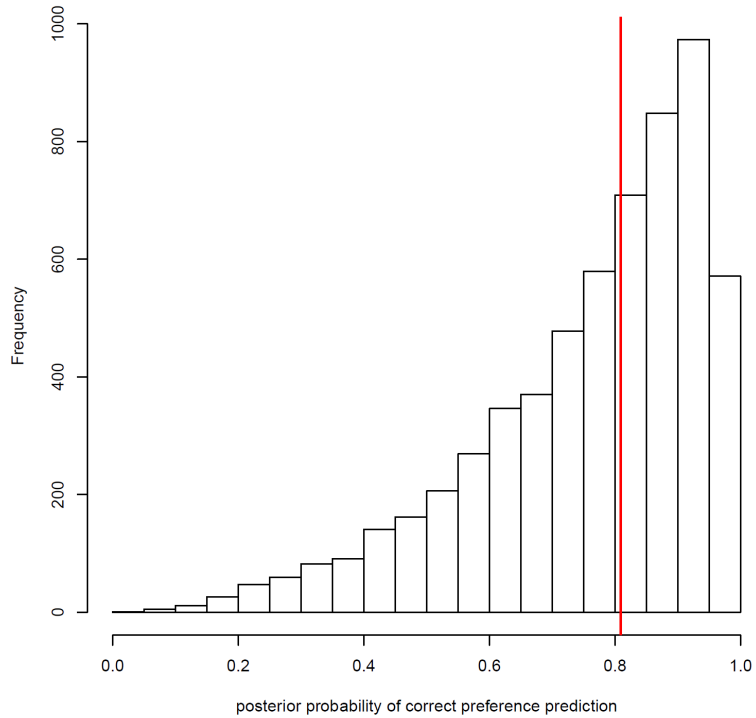
- inspect the posterior predictive probabilities $P(\tilde{\mathbf{R}}_j|\text{data})$ for each assessor

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**

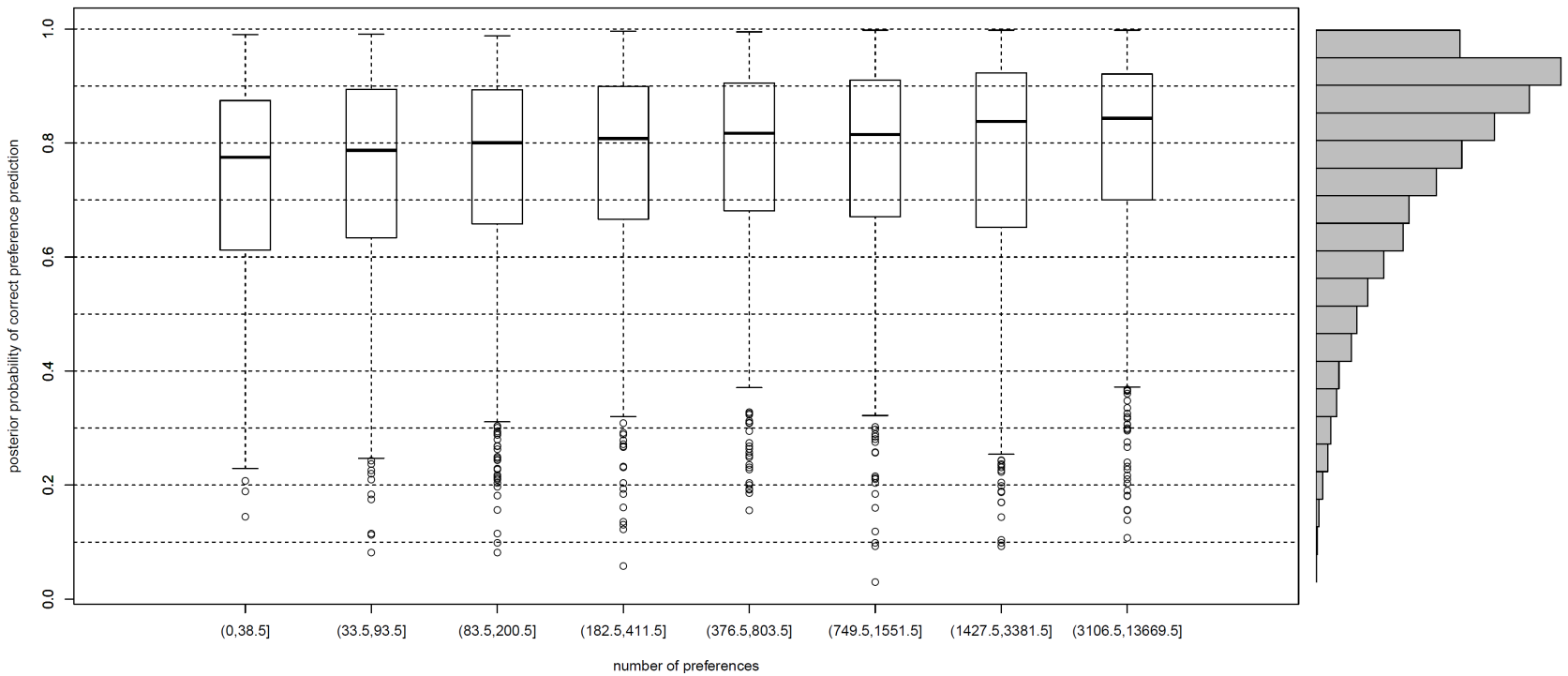
Preference prediction for the model with $C = 9$ clusters

- 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 in Genomics

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 in Genomics

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 in Genomics

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 in Genomics: a benchmark dataset

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.

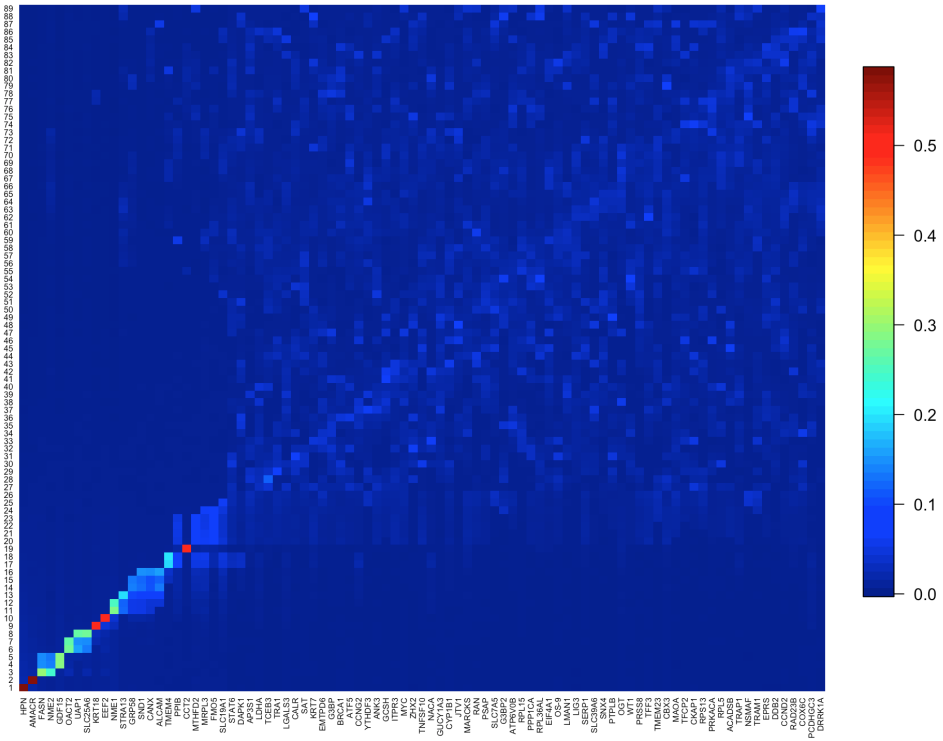
Meta-analysis in Genomics: a benchmark dataset

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	Luo et al. (2001)	Welsh et al. (2001)	Dhanasekaran et al. (2001)	True et al. (2006)	Singh et al. (2002)
1	HPN	HPN	OGT	AMACR	HPN
2	AMACR	AMACR	AMACR	HPN	SLC25A6
3	CYP1B1	OACT2	FASN	NME2	EEF2
4	ATF5	GDF15	HPN	CBX3	SAT
5	BRCA1	FASN	UAP1	GDF15	NME2
6	LGALS3	ANK3	GUCY1A3	MTHFD2	LDHA
7	MYC	KRT18	OACT2	MRPL3	CANX
8	PCDHGC3	UAP1	SLC19A1	SLC25A6	NACA
9	WT1	GRP58	KRT18	NME1	FASN
10	TFF3	PPIB	EEF2	COX6C	SND1

Meta-analysis in Genomics: results

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



Bayesian Integrative Genomics for PanCancer studies

Context:

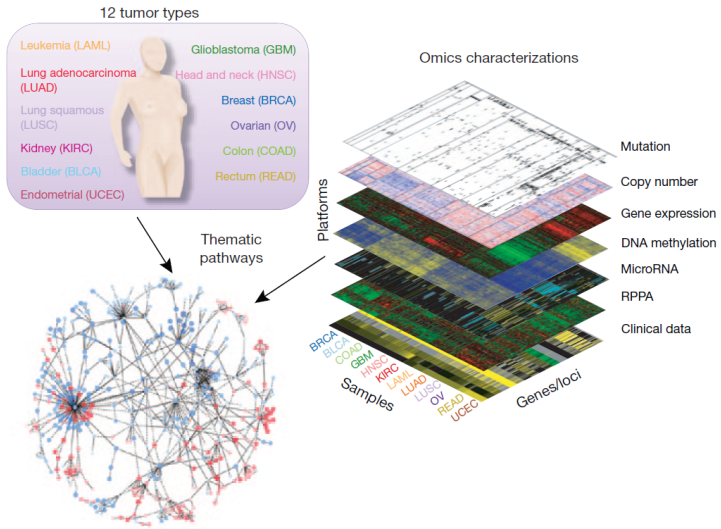
- High-throughput technologies have allowed a systematic exploration of the genetic basis of cancer
- The Cancer Genome Atlas (TCGA) started in 2006: profiling 10,000 tumour samples from 20 tumour types (`http://www.nature.com/tcga/`)

Data:

- $N = 2617$ samples across 12 tumors;
- RNA-seq data, 479 selected functional events which cover $n = 1247$ genes (Ciriello et al. 2013)

Analysis:

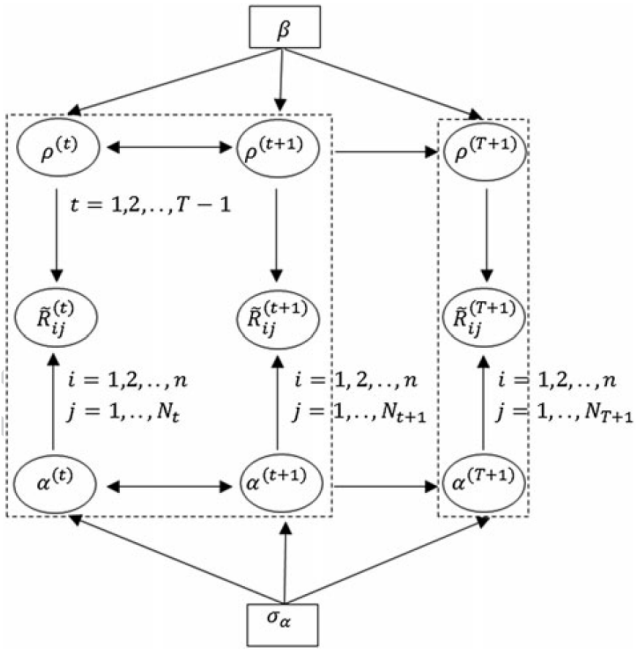
- Bayesian rank-based inference (Vitelli et al. 2018)

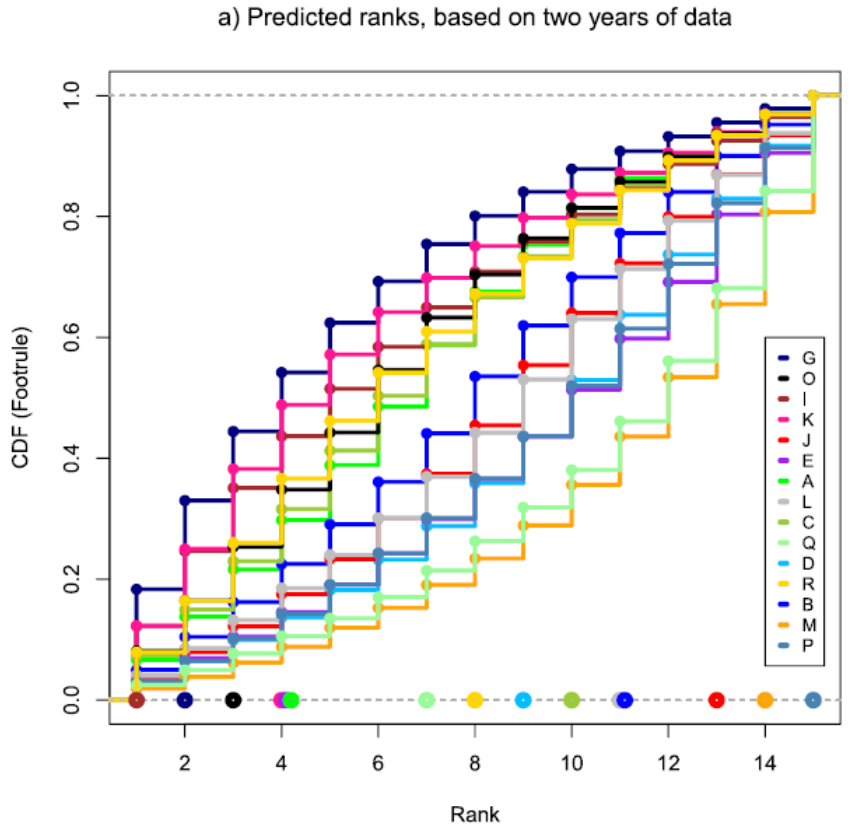


Computations:

- 1.1 mil MCMC iter

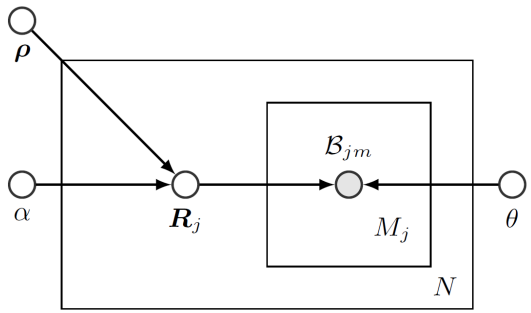
Time dependency: rankings of students along school years

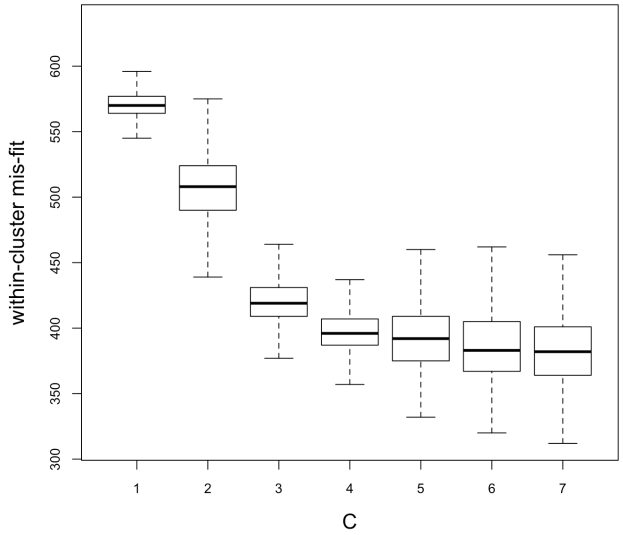


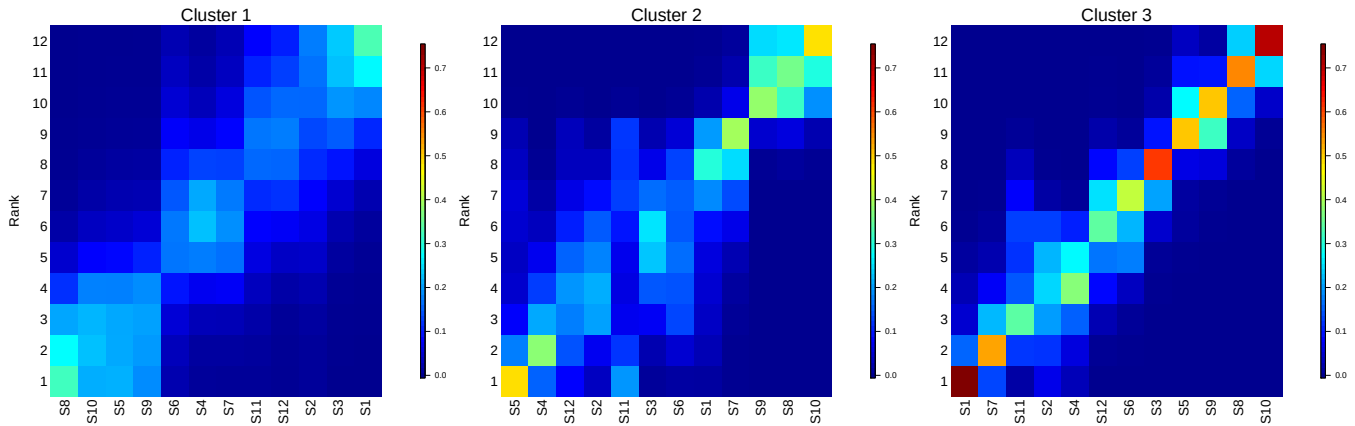


Asfaw et al. (2017)

Inconsistent preferences

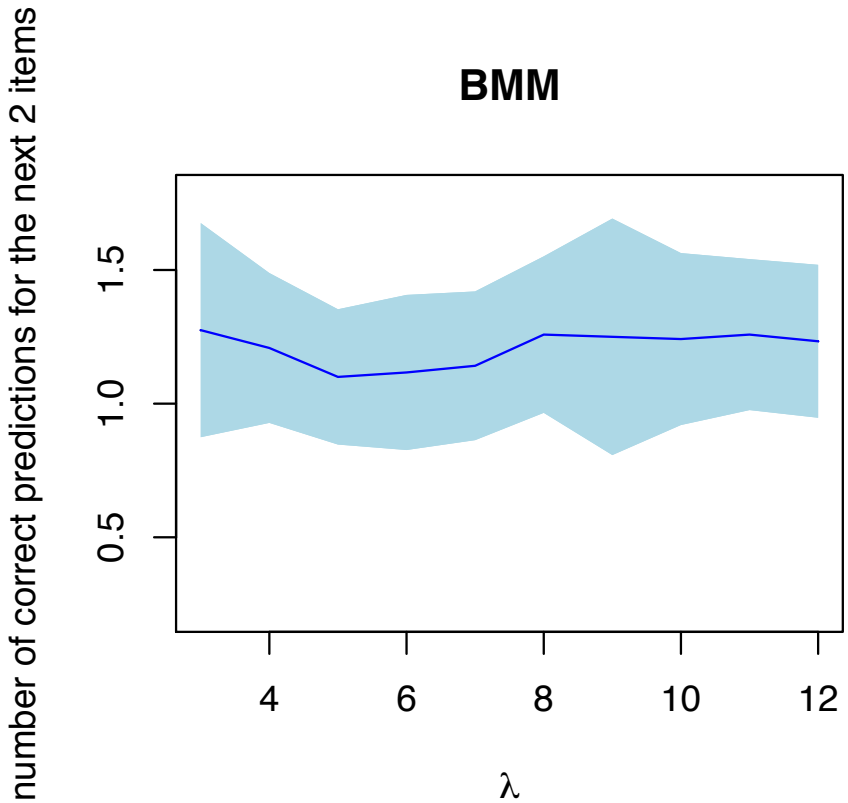


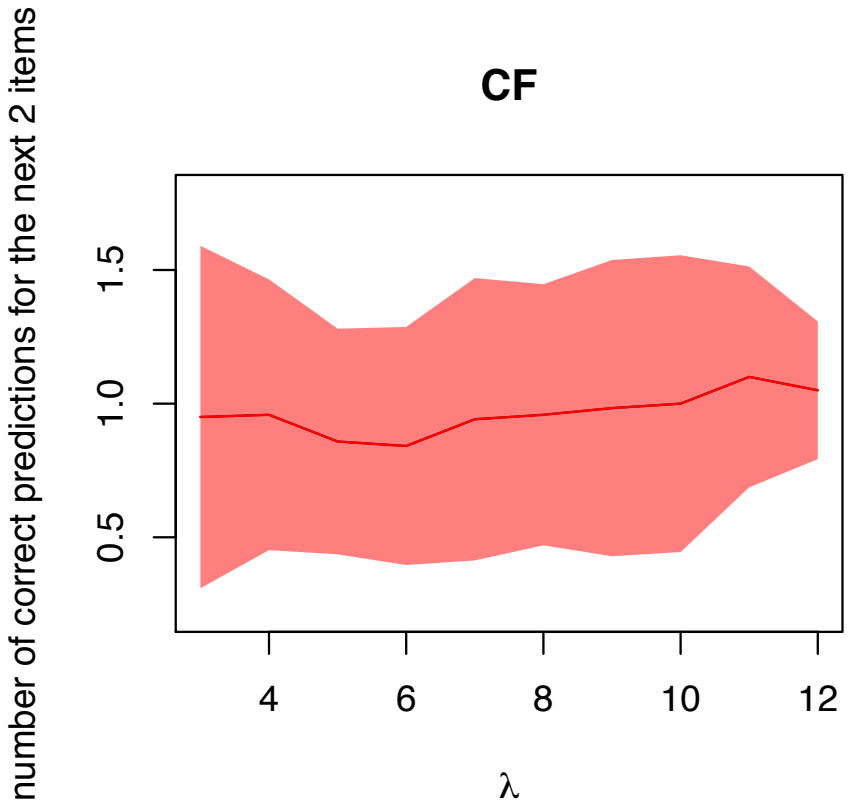




Crispino et al. (2018)

Comparing the BMM with Collaborative Filtering





Liu et al. (2018)

Open points / directions worth exploring

Open points / directions worth exploring

- **crucial model extensions:**
  - variable selection (rank only the items which are worth being ranked),

Open points / directions worth exploring

- **crucial model extensions:**
  - variable selection (rank only the items which are worth being ranked),
  - informative prior for ρ (genomics: include gene pathways info),

Open points / directions worth exploring

- **crucial model extensions:**
  - variable selection (rank only the items which are worth being ranked),
  - informative prior for ρ (genomics: include gene pathways info),
  - covariates (for items & assessors);

Open points / directions worth exploring

- **crucial model extensions:**
  - variable selection (rank only the items which are worth being ranked),
  - informative prior for ρ (genomics: include gene pathways info),
  - covariates (for items & assessors);
- **computational aspects:** scalability, alternatives to MCMC, ...

Open points / directions worth exploring

- **crucial model extensions:**
  - variable selection (rank only the items which are worth being ranked),
  - informative prior for ρ (genomics: include gene pathways info),
  - covariates (for items & assessors);
- **computational aspects:** scalability, alternatives to MCMC, ...
- **other model extensions:** un-equal quality of assessors, infinite mixture (automatically select the number of groups), on-line predictions.

Open points / directions worth exploring

- **crucial model extensions:**
  - variable selection (rank only the items which are worth being ranked),
  - informative prior for ρ (genomics: include gene pathways info),
  - covariates (for items & assessors);
- **computational aspects:** scalability, alternatives to MCMC, ...
- **other model extensions:** un-equal quality of assessors, infinite mixture (automatically select the number of groups), on-line predictions.

Thanks for your attention!
Questions?

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.

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

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 in the *Annals of Applied Statistics*, 2018.

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.

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.

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

Q. Liu*, M. Crispino*, I. Scheel, V. Vitelli and A. Frigessi, "Model-based learning from preference data", *Annuals Review of Statistics and Its Applications*, 2018.

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.

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.

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

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.

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.

J.B. Welsh et al. "Analysis of Gene Expression Identifies Candidate Markers and Pharmacological Targets in Prostate Cancer", *Cancer Research*, **61**(16), 5974–5978, 2001.