Lower-dimensional Bayesian Mallows model for rank-based unsupervised transcriptomic analysis

ISBA 2022

Emilie Eliseussen[1], Thomas Fleischer[2] and Valeria Vitelli[1]

June 30, 2022

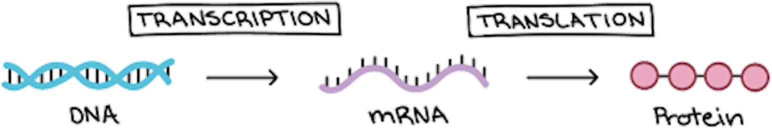
[1] Oslo Centre for Biostatistics and Epidemiology, University of Oslo
[2] Department of Cancer Genetics, Oslo University Hospital, Oslo

Motivation

What is omics data?

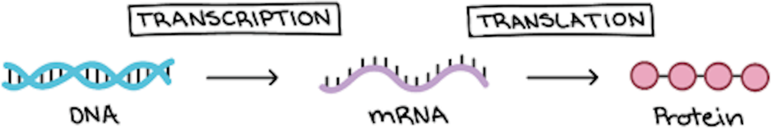
What is omics data?

- The **Central Dogma of Biology**: genes in DNA provide instructions for proteins.



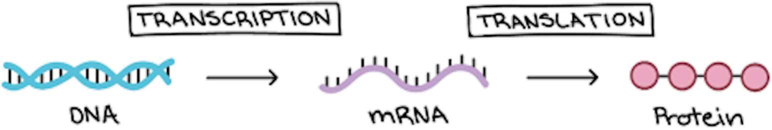
What is omics data?

- The **Central Dogma of Biology**: genes in DNA provide instructions for proteins.
- Omics data is a collective characterization and quantification of biological molecules.



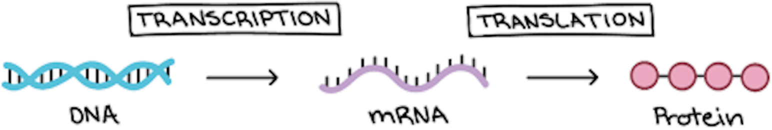
What is omics data?

- The **Central Dogma of Biology**: genes in DNA provide instructions for proteins.
- Omics data is a collective characterization and quantification of biological molecules.
    - Examples of *omics* data: gen*omics* profile DNA, transcript*omics* measure transcripts; prote*omics* and metabol*omics* quantify proteins and metabolites.



What is omics data?

- The **Central Dogma of Biology**: genes in DNA provide instructions for proteins.
- Omics data is a collective characterization and quantification of biological molecules.
    - Examples of *omics* data: gen*omics* profile DNA, transcript*omics* measure transcripts; prote*omics* and metabol*omics* quantify proteins and metabolites.
- Analysis and interpretation of omics data (ideally combined) can lead to a more comprehensive understanding of human health and disease.



Challenges and solutions

Challenges and solutions

Challenges: high-dimensionality ($p >> n$), noise, non-normality, heterogeneity, complex structures, no outcome, …

Challenges and solutions

Challenges: high-dimensionality ($p >> n$), noise, non-normality, heterogeneity, complex structures, no outcome, ...

Possible **solutions:**

Challenges and solutions

Challenges: high-dimensionality ($p >> n$), noise, non-normality, heterogeneity, complex structures, no outcome, ...

Possible **solutions:**

- **Unsupervised** variable selection/dimension reduction.

Challenges and solutions

Challenges: high-dimensionality ($p >> n$), noise, non-normality, heterogeneity, complex structures, no outcome, ...

Possible **solutions:**

- **Unsupervised** variable selection/dimension reduction.
- Transform data into **rankings**: why?

Challenges and solutions

Challenges: high-dimensionality ($p >> n$), noise, non-normality, heterogeneity, complex structures, no outcome, ...

Possible **solutions:**

- **Unsupervised** variable selection/dimension reduction.
- Transform data into **rankings**: why?
    - More robust to noise, outliers, heterogeneity.

Challenges and solutions

Challenges: high-dimensionality ($p >> n$), noise, non-normality, heterogeneity, complex structures, no outcome, ...

Possible **solutions:**

- **Unsupervised** variable selection/dimension reduction.
- Transform data into **rankings**: why?
  - More robust to noise, outliers, heterogeneity.
  - Easier to perform **data integration**: no scaling involved – increase reproducibility.

Main goals

Main goals

Short-term goal: have a rank-based unsupervised variable selection method able to analyze high-dimensional data.

Main goals

Short-term goal: have a rank-based unsupervised variable selection method able to analyze high-dimensional data.

Long-term goal: perform multi-omic data integration to increase statistical power, sample size and improve our understanding of biological systems.

Methodology

Preliminaries (i)

Preliminaries (i)

- $\mathcal{A} = \{A_1, ..., A_n\}$ a finite set of n items ranked by N assessors.
  - In our case: items ← genes, assessors ← patients.

Preliminaries (i)

- $\mathcal{A} = \{A_1, ..., A_n\}$ a finite set of n items ranked by N assessors.
    - In our case: items $\leftarrow$ genes, assessors $\leftarrow$ patients.
- We assume the data are complete rankings $\boldsymbol{R}_j \sim \text{Mallows}(\boldsymbol{\rho}, \alpha)$, $j = 1, ..., N$.

Preliminaries (i)

- $\mathcal{A} = \{A_1, ..., A_n\}$ a finite set of n items ranked by N assessors.
    - In our case: items ← genes, assessors ← patients.
- We assume the data are complete rankings $\boldsymbol{R}_j \sim \text{Mallows}(\boldsymbol{\rho}, \alpha)$, $j = 1, ..., N$.
- The Mallows model [Mallows, 1957] is a probabilistic model for a ranking $\boldsymbol{R}$ defined on the space $\mathcal{P}_n$ of permutations of dimension n:

$$P(\boldsymbol{R}|\alpha, \boldsymbol{\rho}) = \frac{1}{Z_n(\alpha, \boldsymbol{\rho})} \exp\left\{-\frac{\alpha}{n} d(\boldsymbol{R}, \boldsymbol{\rho})\right\} 1_{\mathcal{P}_n}(\boldsymbol{R})$$

where α is a scale parameter, $\boldsymbol{\rho}$ is the consensus ranking, $Z_n(\alpha, \boldsymbol{\rho})$ is the normalizing function and $d(\cdot, \cdot)$ is a distance measure between rankings.

Preliminaries (ii)

The Bayesian Mallows model (BMM) described in [Vitelli et al., 2018], however, does not scale to typical -omics dimensions.

Preliminaries (ii)

The Bayesian Mallows model (BMM) described in [Vitelli et al., 2018], however, does not scale to typical -omics dimensions.

The lower-dimensional Bayesian Mallows model (lowBMM) [Eliseussen et al., 2022]:

Preliminaries (ii)

The Bayesian Mallows model (BMM) described in [Vitelli et al., 2018], however, does not scale to typical -omics dimensions.

The lower-dimensional Bayesian Mallows model (lowBMM) [Eliseussen et al., 2022]:

- $\mathcal{A}^* = \{A_{i_1}, ..., A_{i_{n*}}\}$ is an n^*-dimensional reduced set of items, with $n^* << n$, $\mathcal{A}^* \subset \mathcal{A}$, **$n$* is fixed.**

Preliminaries (ii)

The Bayesian Mallows model (BMM) described in [Vitelli et al., 2018], however, does not scale to typical -omics dimensions.

The lower-dimensional Bayesian Mallows model (lowBMM) [Eliseussen et al., 2022]:

- $\mathcal{A}^* = \{A_{i_1}, ..., A_{i_{n*}}\}$ is an n^*-dimensional reduced set of items, with $n^* << n$, $\mathcal{A}^* \subset \mathcal{A}$, **$n$* is fixed.**
- We assume the data are complete rankings, however only a subset follows the Mallows model while the rest of the data are assumed to be unranked.

Preliminaries (ii)

The Bayesian Mallows model (BMM) described in [Vitelli et al., 2018], however, does not scale to typical -omics dimensions.

The lower-dimensional Bayesian Mallows model (lowBMM) [Eliseussen et al., 2022]:

- $\mathcal{A}^* = \{A_{i_1}, ..., A_{i_{n*}}\}$ is an n^*-dimensional reduced set of items, with $n^* << n$, $\mathcal{A}^* \subset \mathcal{A}$, **$n$* is fixed.**
- We assume the data are complete rankings, however only a subset follows the Mallows model while the rest of the data are assumed to be unranked.
- Scale parameter α fixed.

Preliminaries (ii)

The Bayesian Mallows model (BMM) described in [Vitelli et al., 2018], however, does not scale to typical -omics dimensions.

The lower-dimensional Bayesian Mallows model (lowBMM) [Eliseussen et al., 2022]:

- $\mathcal{A}^* = \{A_{i_1}, ..., A_{i_{n*}}\}$ is an n^*-dimensional reduced set of items, with $n^* << n$, $\mathcal{A}^* \subset \mathcal{A}$, **$n^*$ is fixed.**
- We assume the data are complete rankings, however only a subset follows the Mallows model while the rest of the data are assumed to be unranked.
- Scale parameter α fixed.
- We assume assessors are homogeneous (no mixtures).

The lower-dimensional Bayesian Mallows model (lowBMM)

$$\begin{aligned}
\mathbf{R}_j|_{\mathcal{A}^*} &\sim \text{Mallows}(\boldsymbol{\rho}, \alpha), \qquad && j = 1, \ldots, N \\
\mathbf{R}_j|_{\mathcal{A} \setminus \mathcal{A}^*} &\sim \mathcal{U}(\mathcal{P}_{n-n^*}), && j = 1, \ldots, N \\
\boldsymbol{\rho}|\mathcal{A}^* &\sim \mathcal{U}(\mathcal{P}_{n^*}) \\
\mathcal{A}^* &\sim \mathcal{U}(\mathcal{P}_{\mathcal{C}})
\end{aligned}$$

$\mathcal{C}$: collection of all $\binom{n}{n^*}$ possible sets.

The lower-dimensional Bayesian Mallows model (lowBMM)

$$\begin{aligned} R_j|_{\mathcal{A}^*} &\sim \text{Mallows}(\boldsymbol{\rho}, \alpha), \qquad && j = 1, \ldots, N \\ R_j|_{\mathcal{A}\setminus\mathcal{A}^*} &\sim \mathcal{U}(\mathcal{P}_{n-n^*}), && j = 1, \ldots, N \\ \boldsymbol{\rho}|\mathcal{A}^* &\sim \mathcal{U}(\mathcal{P}_{n^*}) \\ \mathcal{A}^* &\sim \mathcal{U}(\mathcal{P}_{\mathcal{C}}) \end{aligned}$$

$\mathcal{C}$: collection of all $\binom{n}{n^*}$ possible sets.

Posterior distribution:

$$P(\boldsymbol{\rho}, \mathcal{A}^*|\mathrm{R}_1, ..., \mathrm{R}_N) \propto \exp\left\{-\frac{\alpha}{n^*}\sum_{j=1}^{N} d_{\mathcal{A}^*}(\mathrm{R}_j, \boldsymbol{\rho})\right\} 1_{\mathcal{P}_{n^*}}(\boldsymbol{\rho}) 1_{\mathcal{C}}(\mathcal{A}^*).$$

MH-MCMC for lowBMM

We iterate over the following steps:

MH-MCMC for lowBMM

We iterate over the following steps:

1. MH-step: update $\boldsymbol{\rho}$
   Sample: $\rho_{\text{prop}} \sim \text{LS}(\rho_{m-1}, l)$, where LS is the ”leap-and-shift” function from [Vitelli et al., 2018].

MH-MCMC for lowBMM

We iterate over the following steps:

1. MH-step: update $\boldsymbol{\rho}$
   Sample: $\rho_{\text{prop}} \sim \text{LS}(\rho_{m-1}, l)$, where LS is the ”leap-and-shift” function from [Vitelli et al., 2018].
2. MH-step: update $\mathcal{A}^*$
   Sample: $\mathcal{A}^*_{\text{prop}} \sim q(\mathcal{A}^*_{\text{prop}}|\mathcal{A}^*_{m-1})$, where q is described in [Eliseussen et al., 2022].

MH-MCMC for lowBMM

We iterate over the following steps:

1. MH-step: update $\boldsymbol{\rho}$
   Sample: $\rho_{\text{prop}} \sim \text{LS}(\rho_{m-1}, l)$, where LS is the ”leap-and-shift” function from [Vitelli et al., 2018].
2. MH-step: update $\mathcal{A}^*$
   Sample: $\mathcal{A}^*_{\text{prop}} \sim q(\mathcal{A}^*_{\text{prop}} | \mathcal{A}^*_{m-1})$, where q is described in [Eliseussen et al., 2022].

Two tuning parameters involved:

MH-MCMC for lowBMM

We iterate over the following steps:

1. MH-step: update $\boldsymbol{\rho}$
   Sample: $\rho_{\text{prop}} \sim \text{LS}(\rho_{m-1}, l)$, where LS is the ”leap-and-shift” function from [Vitelli et al., 2018].
2. MH-step: update $\mathcal{A}^*$
   Sample: $\mathcal{A}^*_{\text{prop}} \sim q(\mathcal{A}^*_{\text{prop}} | \mathcal{A}^*_{m-1})$, where q is described in [Eliseussen et al., 2022].

Two tuning parameters involved:

- l: ”leap size” for $\boldsymbol{\rho}$: l ∼20% n^*, from previous empirical studies.

MH-MCMC for lowBMM

We iterate over the following steps:

1. MH-step: update $\boldsymbol{\rho}$
   Sample: $\rho_{\text{prop}} \sim \text{LS}(\rho_{m-1}, l)$, where LS is the ”leap-and-shift” function from [Vitelli et al., 2018].
2. MH-step: update $\mathcal{A}^*$
   Sample: $\mathcal{A}^*_{\text{prop}} \sim q(\mathcal{A}^*_{\text{prop}} | \mathcal{A}^*_{m-1})$, where q is described in [Eliseussen et al., 2022].

Two tuning parameters involved:

- l: ”leap size” for $\boldsymbol{\rho}$: l ∼20% n^*, from previous empirical studies.
- L: ”swap size” for $\mathcal{A}^*$: keep low.

Data examples

Simulation study: set-up

Simulation study: set-up

- Two data dimensions: toy example ($n = 20$) and more realistic dimension ($n = 1000$).

Simulation study: set-up

- Two data dimensions: toy example ($n = 20$) and more realistic dimension ($n = 1000$).
- Data simulated in a top-rank scenario, i.e. items in $\mathcal{A}^*$ are top ranked.

Simulation study: set-up

- Two data dimensions: toy example ($n = 20$) and more realistic dimension ($n = 1000$).
- Data simulated in a top-rank scenario, i.e. items in $\mathcal{A}^*$ are top ranked.
- Posterior summaries computed after burn-in: $\hat{\boldsymbol{\rho}}_{\hat{\mathcal{A}}^*}$, $\hat{\mathcal{A}}^*$ (details in [Eliseussen et al., 2022]).

Simulation study: set-up

- Two data dimensions: toy example ($n = 20$) and more realistic dimension ($n = 1000$).
- Data simulated in a top-rank scenario, i.e. items in $\mathcal{A}^*$ are top ranked.
- Posterior summaries computed after burn-in: $\hat{\boldsymbol{\rho}}_{\hat{\mathcal{A}}^*}$, $\hat{\mathcal{A}}^*$ (details in [Eliseussen et al., 2022]).
- Performance measures:

Simulation study: set-up

- Two data dimensions: toy example ($n = 20$) and more realistic dimension ($n = 1000$).
- Data simulated in a top-rank scenario, i.e. items in $\mathcal{A}^*$ are top ranked.
- Posterior summaries computed after burn-in: $\hat{\boldsymbol{\rho}}_{\hat{\mathcal{A}}^*}$, $\hat{\mathcal{A}}^*$ (details in [Eliseussen et al., 2022]).
- Performance measures:
  - $\hat{p} = n_{\text{corr}}/n^*$, with $n_{\text{corr}} = |\mathcal{A}^* \bigcap \hat{\mathcal{A}}^*|$.

Simulation study: set-up

- Two data dimensions: toy example ($n = 20$) and more realistic dimension ($n = 1000$).
- Data simulated in a top-rank scenario, i.e. items in $\mathcal{A}^*$ are top ranked.
- Posterior summaries computed after burn-in: $\hat{\boldsymbol{\rho}}_{\hat{\mathcal{A}}^*}$, $\hat{\mathcal{A}}^*$ (details in [Eliseussen et al., 2022]).
- Performance measures:
  - $\hat{p} = n_{\text{corr}}/n^*$, with $n_{\text{corr}} = |\mathcal{A}^* \bigcap \hat{\mathcal{A}}^*|$.
  - $d_{\text{norm}}(\boldsymbol{\rho}_{\mathcal{A}^*}, \hat{\boldsymbol{\rho}}_{\mathcal{A}^*}) = d(\boldsymbol{\rho}_{\mathcal{A}^*}, \hat{\boldsymbol{\rho}}_{\mathcal{A}^*})/n_{\text{corr}}$.

Simulation study: set-up

- Two data dimensions: toy example ($n = 20$) and more realistic dimension ($n = 1000$).
- Data simulated in a top-rank scenario, i.e. items in $\mathcal{A}^*$ are top ranked.
- Posterior summaries computed after burn-in: $\hat{\boldsymbol{\rho}}_{\hat{\mathcal{A}}^*}$, $\hat{\mathcal{A}}^*$ (details in [Eliseussen et al., 2022]).
- Performance measures:
  - $\hat{p} = n_{\text{corr}}/n^*$, with $n_{\text{corr}} = |\mathcal{A}^* \bigcap \hat{\mathcal{A}}^*|$.
  - $d_{\text{norm}}(\boldsymbol{\rho}_{\mathcal{A}^*}, \hat{\boldsymbol{\rho}}_{\mathcal{A}^*}) = d(\boldsymbol{\rho}_{\mathcal{A}^*}, \hat{\boldsymbol{\rho}}_{\mathcal{A}^*})/n_{\text{corr}}$.
  - Computing time

Comparison with other methods

	Mallows-based	Frequentist/ Bayesian	Estimate uncertainty	Variable selection
Mallows model (MM)[1]	✓	Frequentist	✗	✗
Extended Mallows model (EMM)	✓	Frequentist	✓	✗
Partition Mallows model (PAMA)	✓	Bayesian	✓	✓
Bayesian Mallows model (BMM)	✓	Bayesian	✓	✗
Markov chain-based methods MC_1, MC_2, MC_3	✗	Frequentist	✗	✗
Cross Entropy Monte Carlo (CEMC)	✗	Frequentist	✗	✗
BORDA	✗	–	✗	✗

Table 1: Methods used in the comparison with lowBMM.

[1] R package ***PerMallows*** used.

Comparison with other methods: toy example

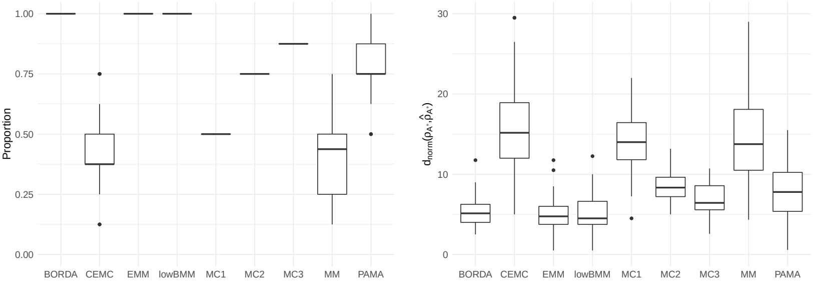


Figure 1: Boxplots of $\hat{p}$ (left) and d_{norm} (right) over 50 repetitions. $n = 20$, $n^* = 8$, $N = 5$, $M = 10^3$, $\alpha = 2$.

Comparison with other methods: toy example

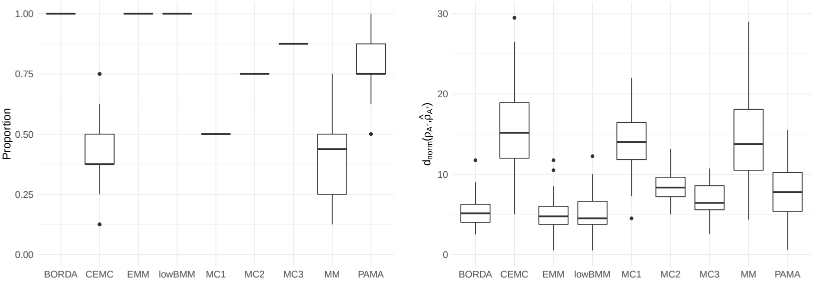


Figure 1: Boxplots of $\hat{p}$ (left) and d_{norm} (right) over 50 repetitions. $n = 20$, $n^* = 8$, $N = 5$, $M = 10^3$, $\alpha = 2$.

Method	BORDA	CEMC	EMM	lowBMM	MC1	MC2	MC3	MM	PAMA
time (sec)	0.02	58.42	0.05	0.54	4.42	4.42	4.42	0.0003	17.58

Table 2: Average computing times over 50 runs.

Comparison with other methods: larger dimension

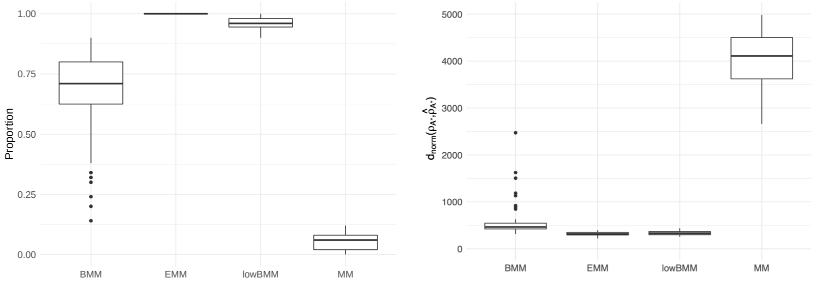


Figure 2: Boxplots of $\hat{p}$ (left) and d_{norm} (right) over 50 repetitions. $n = 1000$, $N = 50$, $n^* = 50$, $M = 7.5 \cdot 10^4$, $\alpha = 5$.

Comparison with other methods: larger dimension

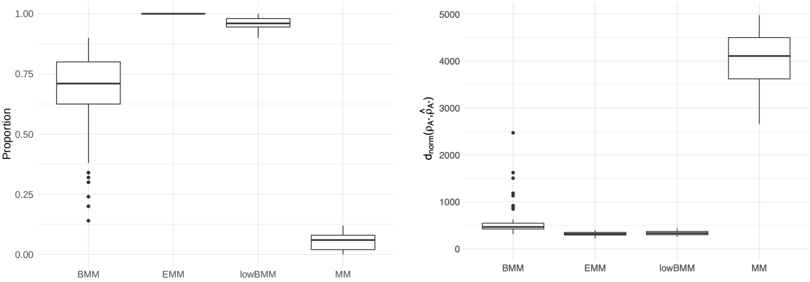


Figure 2: Boxplots of $\hat{p}$ (left) and d_{norm} (right) over 50 repetitions. $n = 1000$, $N = 50$, $n^* = 50$, $M = 7.5 \cdot 10^4$, $\alpha = 5$.

Method	BMM	EMM	lowBMM	MM
time (sec)	116.36	2307.42	236.92	9.48

Table 3: Average computing times over 50 runs.

Real data application: set-up

[2] http://www.nature.com/tcga/

Real data application: set-up

Raw RNAseq data from The Cancer Genome Atlas (TCGA)[2]:

[2] http://www.nature.com/tcga/

Real data application: set-up

Raw RNAseq data from The Cancer Genome Atlas (TCGA)[2]:

- $n = 15348$ genes and $N = 265$ ovarian cancer patients.

[2] http://www.nature.com/tcga/

Real data application: set-up

Raw RNAseq data from The Cancer Genome Atlas (TCGA)[2]:

- $n = 15348$ genes and $N = 265$ ovarian cancer patients.
- $n^* = 500$, number of genes to be selected.

[2] http://www.nature.com/tcga/

Real data application: set-up

Raw RNAseq data from The Cancer Genome Atlas (TCGA)[2]:

- $n = 15348$ genes and $N = 265$ ovarian cancer patients.
- $n^* = 500$, number of genes to be selected.
- Tuning parameters: $L = 1$, $l = 100$.

[2] http://www.nature.com/tcga/

Real data application: set-up

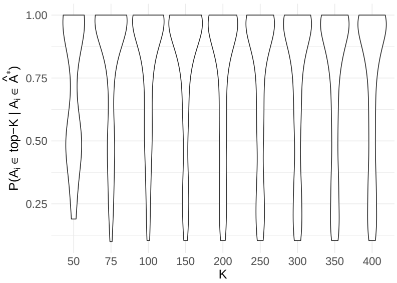
Raw RNAseq data from The Cancer Genome Atlas (TCGA)[2]:

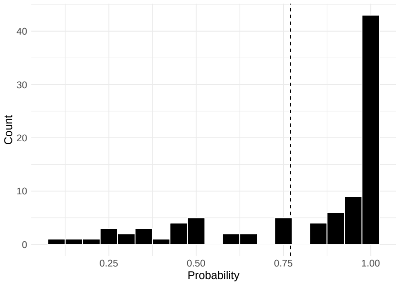
- $n = 15348$ genes and $N = 265$ ovarian cancer patients.
- $n^* = 500$, number of genes to be selected.
- Tuning parameters: $L = 1$, $l = 100$.
- $\alpha = 10$ (estimated off-line).

[2] http://www.nature.com/tcga/

Real data application: post-processing of results

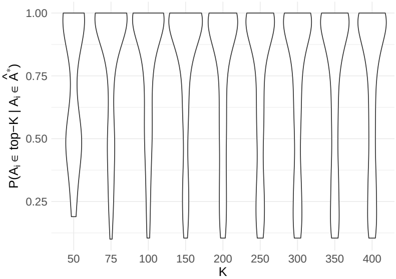
- Post-processing step: compute “top probability selection”: $\hat{\mathcal{A}}^*_{\text{top}} = \{A_i \in \mathcal{A}^* \text{ s.t. } P(A_i \in \text{top-}K, i = 1, \dots, n \mid A_i \in \hat{\mathcal{A}}^*) > c\}$.

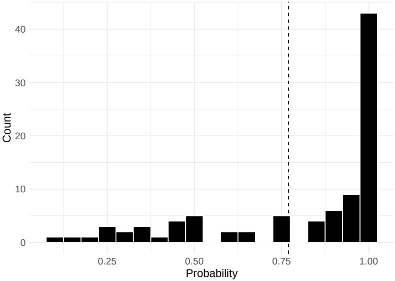




Real data application: post-processing of results

- Post-processing step: compute "top probability selection": $\hat{\mathcal{A}}^*_{\text{top}} = \{A_i \in \mathcal{A}^* \text{ s.t.} \, P(A_i \in \text{top-}K, i = 1, \ldots, n \,|\, A_i \in \hat{\mathcal{A}}^*) > c\}$.
- In our case: $K = 75$ and $c = 0.77$ resulting in $|\hat{\mathcal{A}}^*_{\text{top}}| = 63$ genes.





Interpretation of results: GSEA

Gene Set Name	# Genes in Gene set	# Genes in overlap	p-value
Regulation of cell differentiation (GOBP)	1618	16	2.59E-9
Regulation of multicellular organimsal development (GOBP)	1397	13	2.05E-7
Regulation of anatomical structure morphogenesis (GOBP)	1006	11	4.17E-7
Positive regulation of developmental process (GOBP)	1284	12	6.17E-7
Positive regulation of cell differentiation (GOBP)	844	10	7.2E-7
Response to endogenous stimulus (GOBP)	1624	13	1.12E-6
Cellular response to nitrogen compound (GOBP)	698	9	1.37E-6
Sensory organ development (GOBP)	534	8	1.84E-6
Animal organ morphogenesis (GOBP)	1025	10	4.08E-6
Striated muscle cell differentiation (GOBP)	269	6	4.12E-6

Table 4: Overview of the top-10 gene sets ranked according to the associated p-value from a GSEA performed on the selection $\hat{\mathcal{A}}^*_{\text{top}}$.

Conclusions and way forward

Main takeaways

- Ranks are more robust to outliers, noise, and allows for easier comparisons between multiple data sources.

Main takeaways

- Ranks are more robust to outliers, noise, and allows for easier comparisons between multiple data sources.
- Variable selection is essential in high-dimensional settings such as omics.

Main takeaways

- Ranks are more robust to outliers, noise, and allows for easier comparisons between multiple data sources.
- Variable selection is essential in high-dimensional settings such as omics.
- Aim of lowBMM: reproducible and robust unsupervised variable selection procedure in a complex high-dimensional setting.

Further directions

- Extensions for lowBMM: clustering, handle missing data, include estimation of α, improved convergence diagnostics.

Further directions

- Extensions for lowBMM: clustering, handle missing data, include estimation of α, improved convergence diagnostics.
- ...once this is in place → multiple data integration.

References

Eliseussen, E., Fleischer, T., and Vitelli, V. (2022).
Rank-based Bayesian variable selection for genome-wide transcriptomic analyses.
Accepted for publication in Statistics in Medicine.

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

Vitelli, V., Sørensen, O., Crispino, M., Frigessi, A., and Arjas, E. (2018).
Probabilistic Preference Learning with the Mallows Rank Model.
Journal of Machine Learning Research, 18(1):5796–5844.

Lower-dimensional Bayesian Mallows model for rank-based unsupervised transcriptomic analysis

ISBA 2022

Emilie Eliseussen[1], Thomas Fleischer[2] and Valeria Vitelli[1]

June 30, 2022

[1] Oslo Centre for Biostatistics and Epidemiology, University of Oslo
[2] Department of Cancer Genetics, Oslo University Hospital, Oslo

Back-up slides

What is ranking data?

A **ranking dataset**: describes a ranking of a set of items according to some specified feature.

Any dataset can be turned into a ranking dataset, e.g.:

	Gene 1	Gene 2	Gene 3	Gene 4
Patient 1	-0.4	1.2	0.9	-21.4
Patient 2	-5.3	0.3	12.1	-1.6

Table 5: RNAseq

	Gene 1	Gene 2	Gene 3	Gene 4
Patient 1	3	1	2	4
Patient 2	4	2	1	3

Table 6: Rankings

The lower-dimensional Bayesian Mallows model (lowBMM)

Likelihood:

$$P(\mathsf{R}_1, ..., \mathsf{R}_N | \boldsymbol{\rho}, \mathcal{A}^*) = \frac{1}{Z_{n^*}(\alpha)^N} \exp\left\{ -\frac{\alpha}{n^*} \sum_{j=1}^{N} d_{\mathcal{A}^*}(\mathsf{R}_j, \boldsymbol{\rho}) \right\} \prod_{j=1}^{N} 1_{\mathcal{P}_{n^*}}(\boldsymbol{\rho}) 1_{\mathcal{C}}(\mathcal{A}^*).$$

Priors: uniform for both parameters, $\pi(\boldsymbol{\rho}|\mathcal{A}^*) = \frac{1}{n^*!} 1_{\mathcal{P}_{n^*}}(\boldsymbol{\rho})$ and $\pi(\mathcal{A}^*) = \frac{1}{|\mathcal{C}|} 1_{\mathcal{C}}(\mathcal{A}^*)$, where $\mathcal{C}$ is the collection of all $\binom{n}{n^*}$ possible sets.

Posterior distribution:

$$P(\boldsymbol{\rho}, \mathcal{A}^* | \mathsf{R}_1, ..., \mathsf{R}_N) \propto \exp\left\{ -\frac{\alpha}{n^*} \sum_{j=1}^{N} d_{\mathcal{A}^*}(\mathsf{R}_j, \boldsymbol{\rho}) \right\} 1_{\mathcal{P}_{n^*}}(\boldsymbol{\rho}) 1_{\mathcal{C}}(\mathcal{A}^*).$$

MCMC: details

In the first step of the algorithm, we propose a new consensus ranking $\boldsymbol{\rho}' \in \mathcal{P}_{n^*}$ using the "leap-and-shift" proposal distribution described in [Vitelli et al., 2018]. The acceptance probability for updating $\boldsymbol{\rho}$ in the MH algorithm is

$$\min\left\{1, \frac{P_l(\boldsymbol{\rho}|\boldsymbol{\rho}')}{P_l(\boldsymbol{\rho}'|\boldsymbol{\rho})} \exp\left[-\frac{\alpha}{n^*}\left(\sum_{j=1}^{N} d_{\mathcal{A}^*}(\mathsf{R}_j, \boldsymbol{\rho}') - \sum_{j=1}^{N} d_{\mathcal{A}^*}(\mathsf{R}_j, \boldsymbol{\rho})\right)\right]\right\}. \tag{1}$$

We propose a new set $\mathcal{A}^*_{\text{prop}}$ by perturbing $L \in \{1, ..., n^*\}$ elements in the current $\mathcal{A}^*$, selected with uniform probability. The L items are swapped with L items from the set $\mathcal{A} \setminus \mathcal{A}^*$, again uniformly. The move from $\mathcal{A}^*$ to $\mathcal{A}^*_{\text{prop}}$ is accepted with probability:

$$\min\left\{1, \exp\left[-\frac{\alpha}{n^*}\left(\sum_{j=1}^{N} d_{\mathcal{A}^*_{\text{prop}}}(\mathsf{R}_j, \boldsymbol{\rho}) - \sum_{j=1}^{N} d_{\mathcal{A}^*}(\mathsf{R}_j, \boldsymbol{\rho})\right)\right]\right\}. \tag{2}$$

lowBMM: MCMC algorithm

Algorithm 1: MCMC scheme for inference in lowBMM

```
input: $\mathrm{R}_1, .., \mathrm{R}_N$, $\alpha$, $d(\cdot,\cdot)$, $l$, $L$, $M$
output: posterior distributions of $\rho$ and $\mathcal{A}^*$
Initialization: randomly generate $\rho_0$ and $\mathcal{A}^*_0$
for $m \leftarrow 1$ to $M$ do
    M-H step: update $\rho$
    sample: $\rho' \sim \mathrm{LS}(\rho_{m-1}, l)$ restricted on $\mathcal{A}^*_{m-1}$ and $u \sim \mathcal{U}(0,1)$
    compute: $ratio \leftarrow$ equation(1) with $\rho \leftarrow \rho_{m-1}$, $\mathcal{A}^* \leftarrow \mathcal{A}^*_{m-1}$
    if $u < ratio$ then
        $\rho_m \leftarrow \rho'$
    else
        $\rho_m \leftarrow \rho_{m-1}$
    end
    M-H step: update $\mathcal{A}^*$
    sample: $L$ elements in $\mathcal{A}^*_{m-1}$ to get $\mathcal{A}^*_{\text{prop}}$, and $u \sim \mathcal{U}(0,1)$
    compute: $ratio \leftarrow$ equation(2) with $\rho \leftarrow \rho_{m-1}$, $\mathcal{A}^* \leftarrow \mathcal{A}^*_{m-1}$
    if $u < ratio$ then
        $\mathcal{A}^*_m \leftarrow \mathcal{A}^*_{\text{prop}}$
    else
        $\mathcal{A}^*_m \leftarrow \mathcal{A}^*_{m-1}$
    end
end
```

MH-MCMC for lowBMM: postprocessing

Posterior summaries: $\hat{\boldsymbol{\rho}}_{\mathcal{A}^*}$ and $\hat{\mathcal{A}}^*$ are computed in the following way:

1. Suppose M posterior samples are obtained: $\{\boldsymbol{\rho}_m, \mathcal{A}_m^*\}_{m=1}^M$ with $\boldsymbol{\rho}_m = \{\rho_{mi_1^m}, ..., \rho_{mi_{n*}^m}\}$ and $\mathcal{A}_m^* = \{A_{mi_1^m}, ..., A_{mi_{n*}^m}\}$.
2. Given the samples $\{\mathcal{A}_1^*, ..., \mathcal{A}_M^*\}$, let $W \in \mathbb{R}^{M \times n}$ be such that $W_{mi} = 1_{\mathcal{A}_m^*}(A_i)$ for each item A_i, $i = 1, ..., n$.
3. Let $\mathcal{A}'$ be the the "Highest Probability Set" of $\mathcal{A}^*$ (more details in [Eliseussen et al., 2022]). Based on $\mathcal{A}'$ we compute $\bar{\mathsf{x}} \in \mathbb{R}^{|\mathcal{A}'|}$, $\bar{x}_i = \frac{\sum_{m=1}^M \boldsymbol{\rho}_{mi} 1_{\mathcal{A}_m^*}(A_i)}{\sum_{m=1}^M 1_{\mathcal{A}_m^*}(A_i)}$ for all $A_i \in \mathcal{A}'$.
4. We quantify the two posterior summaries of $\boldsymbol{\rho}$ and $\mathcal{A}^*$ as follows:

$$\hat{\mathcal{A}}^* = \{A_i \in \mathcal{A}' \mid rank(\bar{\mathsf{x}}) \leq n^*\}, \qquad \hat{\boldsymbol{\rho}}_{\mathcal{A}^*} = rank(\bar{\mathsf{x}})|_{\hat{\mathcal{A}}^*} \quad (3)$$

Comparison with other methods

	Mallows-based	Frequentist/ Bayesian	Estimate uncertainty	Several distances[3]	Variable selection
Mallows model (MM)[4]	✓	Frequentist	✗	✓	✗
Extended Mallows model (EMM)	✓	Frequentist	✓	✗	✗
Partition Mallows model (PAMA)	✓	Bayesian	✓	✗	✓
Bayesian Mallows model (BMM)	✓	Bayesian	✓	✓	✗
Markov chain-based methods MC$_1$, MC$_2$, MC$_3$	✗	Frequentist	✗	–	✗
Cross Entropy Monte Carlo (CEMC)	✗	Frequentist	✗	–	✗
BORDA	✗	–	✗	–	✗

Table 7: Methods used in the comparison with lowBMM.

[3] Given that the method is Mallows-based.

[4] R package ***PerMallows*** used.