Identifying microbiota sub-communities using topic models

Microbiome Data Analysis Workshop 2025

Laura Symul
laura.symul@uclouvain.be
UCLouvain, LIDAM, ISBA

2024-12-17

github.com/lasy/t-2425-Microbiota-UHasselt

Outline

Hello 👋👋

Introduction

Topic models

Choosing K, the number of topics

Conclusions & references

🎓 I am an assistant professor (👶) in “non-clinical biostatistics” at UCLouvain

🔬 My *interdisciplinary* research agenda revolves around “Statistics for Women’s Health”

🤝 I am part of the Vaginal Microbiome Research Consortium (VMRC)

🤓 Controversy? The microbiome is *NOT* compositional in essence!

🧗 Besides work, I love climbing (and talking (too much) about it)

Code and data availability

These slides are made with Quarto, and all analyses presented today are executed within the slides.

▶ Code

You can find the source for this slidedeck on GitHub: lasy/t-2425-Microbiota-UHasselt

Data used here are publicly available and linked in the slides.

Outline

Hello 👋👋

Introduction

Topic models

Choosing K, the number of topics

Conclusions & references

Running Examples

We'll use two microbiota datasets:

- One vaginal microbiota dataset: The ISALA dataset (Lebeer et al. 2023)
- One gut microbiota dataset (LLD study) from the `curatedMetagenomicsData` Bioconductor package (Pasolli et al. 2017)

Microbiology paper (Lebeer et al. 2023).

De-identified data for over 3000 samples is available on the LebeerLab GitHub repository as `tsv` files.

We'll use the data aggregated at the Genus level (but Species level for *Lactobacillus* spp.) for this example.

▶ Code

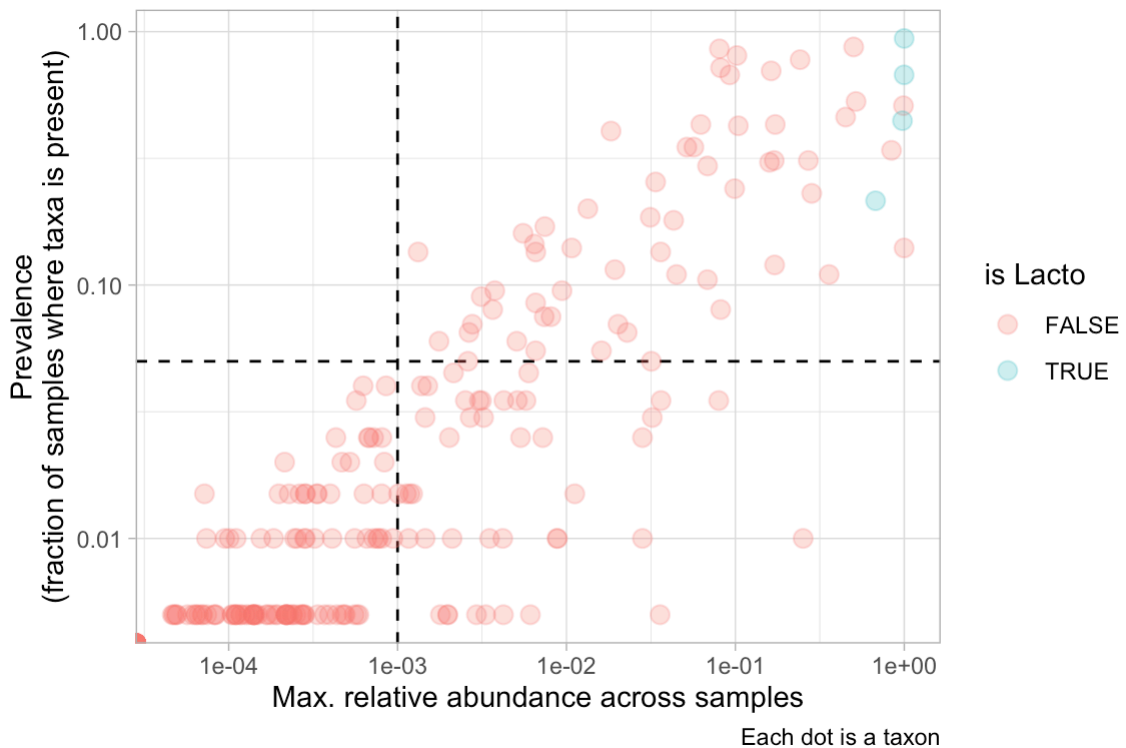
▶ Code

```
# A SummarizedExperiment-tibble abstraction: 1,661,478 × 38
# Features=546 | Samples=3043 | Assays=counts, rel_ab
   .feature           .sample  counts  rel_ab general_age health_bmi health_antibiotic_3m…¹
sexual_intercourse_2…² technical_run_202010…³ technical_run_202010…⁴
   <chr>              <chr>     <dbl>   <dbl>       <dbl>      <dbl> <lgl>                  <lgl>
<lgl>                  <lgl>
 1 Methanobrevibacter SAMEA13…      0 0                29       18.9 TRUE                   FALSE
TRUE                   FALSE
```

Vaginal microbiota dataset: ISALA

We take a random subset of 200 samples (out of) to make it easier to work with.

We then also filter out “rare taxa”:



This reduces the number of taxa from 546 to 60.

Gut microbiota dataset: `curatedMetagenomics`

We'll use data from the `curatedMetagenomicData` Bioconductor package.

Gut microbiota dataset: `curatedMetagenomics`

and more specifically, data from the `LifeLinesDeep 2016` study.

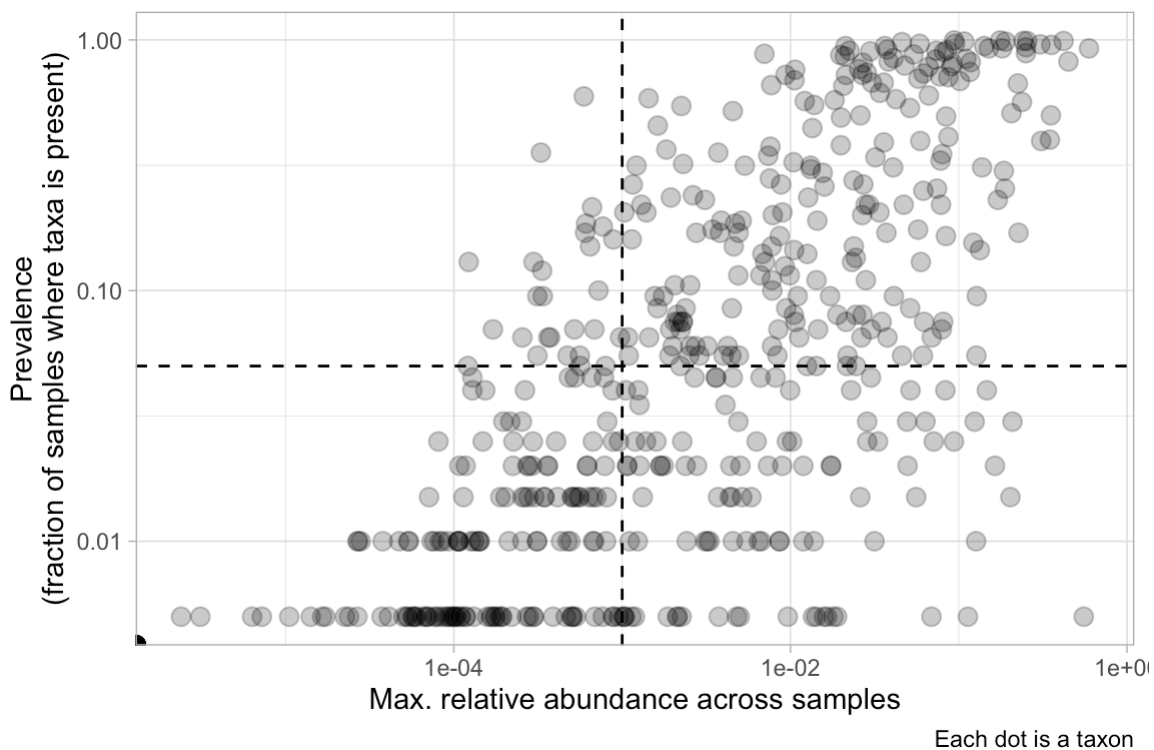
As for the ISALA data, we create a `SummarizedExperiment` object from the package data.

```
# A SummarizedExperiment-tibble abstraction: 722,995 × 32
# Features=637 | Samples=1135 | Assays=counts, rel_ab
   .feature              .sample counts  rel_ab study_name subject_id body_site antibiotics_current_…¹
study_condition disease   age age_category gender country
   <chr>                 <chr>    <dbl>   <dbl> <chr>      <chr>      <chr>     <chr>                  <chr>
<chr>   <int> <chr>         <chr>  <chr>
 1 [Bacteroides] pectin… EGAR00…      0 0       LifeLines… sub_90020… stool     no                     control
healthy    35 adult         female NLD
 2 [Butyribacterium] me… EGAR00…      0 0       LifeLines… sub_90020… stool     no                     control
healthy    35 adult         female NLD
 3 [Clostridium] hylemo… EGAR00…      0 0       LifeLines… sub_90020… stool     no                     control
healthy    35 adult         female NLD
 4 [Clostridium] innocu… EGAR00…      0 0       LifeLines… sub_90020… stool     no                     control
healthy    35 adult         female NLD
 5 [Clostridium] leptum  EGAR00…   9031 2.88e-4 LifeLines… sub_90020… stool     no                     control
healthy    35 adult         female NLD
 6 [Clostridium] methyl… EGAR00…      0 0       LifeLines… sub_90020… stool     no                     control
healthy    35 adult         female NLD
 7 [Clostridium] scinde… EGAR00…      0 0       LifeLines… sub_90020… stool     no                     control
healthy    35 adult         female NLD
 8 [Clostridium] spirof… EGAR00…      0 0       LifeLines… sub_90020… stool     no                     control
healthy    35 adult         female NLD
```

Gut microbiota dataset: `curatedMetagenomics`

As for the ISALA data, we take a random subset of 200 samples (out of 1135) to make it easier to work with.

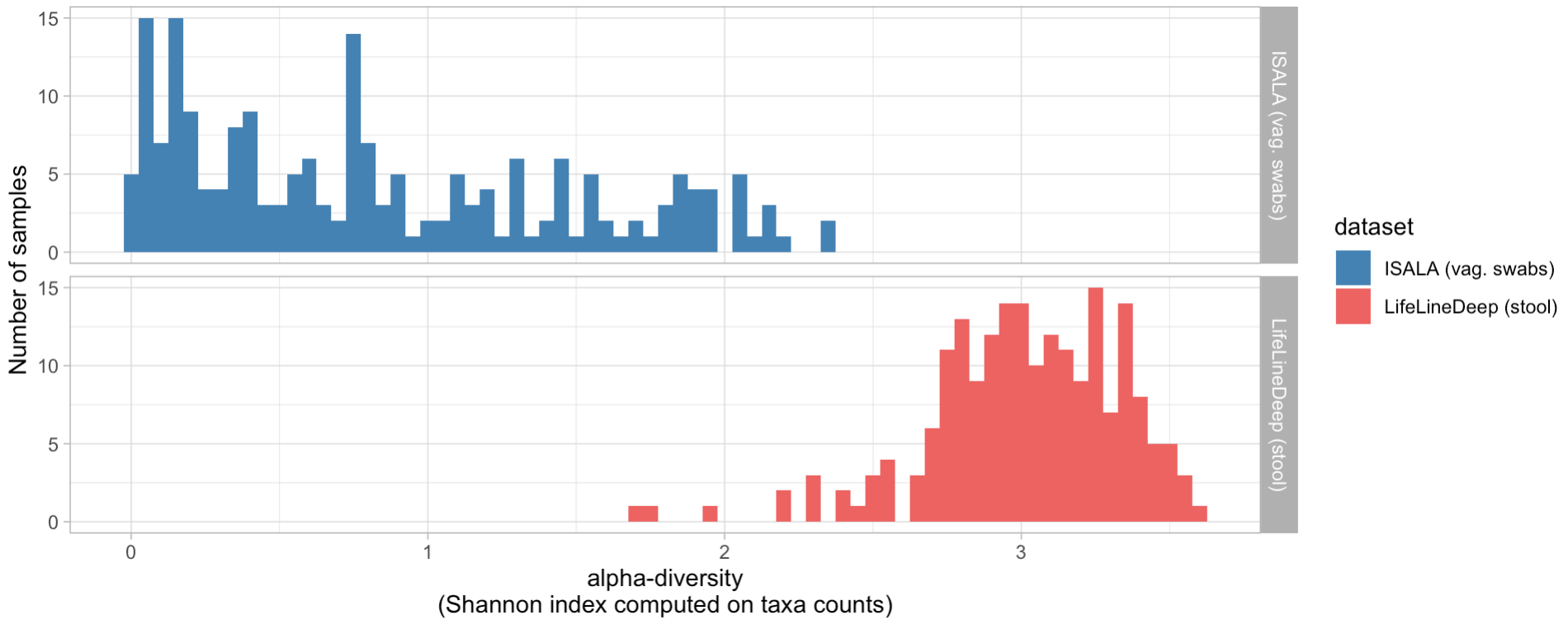
We then also filter out “rare taxa”:



github.com/lasy/t-2425-Microbiota-UHasselt

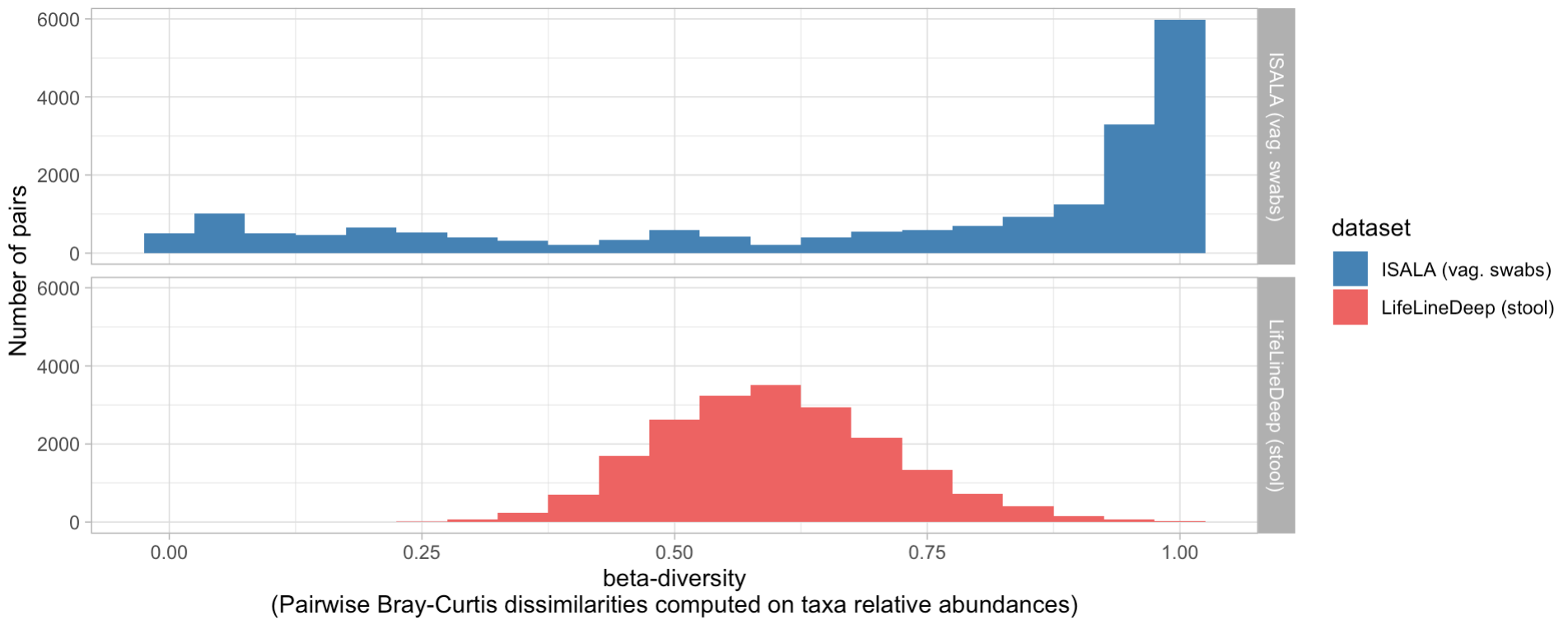
Quick exploration and comparison of the data: Microbiota diversity

α-diversity



➡️ Vaginal swabs α-diversity is lower than stool samples α-diversity.

β-diversity



➡️ Vaginal swabs β-diversity has a wide distribution: some samples are very similar (complete overlap), others are very different (no overlap).

Any typical groups or types?

Any typical microbiota groups or types?

ISALA (vaginal swabs)

LLD (stool) - aggregated at the Family level

github.com/lasy/t-2425-Microbiota-UHasselt

“Subcommunities” in microbiotas

⚠️ Note: bacterial “*subcommunity*” is not a well defined term in microbiology.

But it aims to reflect the consequences of well establish concepts (*i.e.*, bacterial interactions).

Bacterial interactions

Many bacteria interact with each other in many ways, mostly through metabolic exchanges.

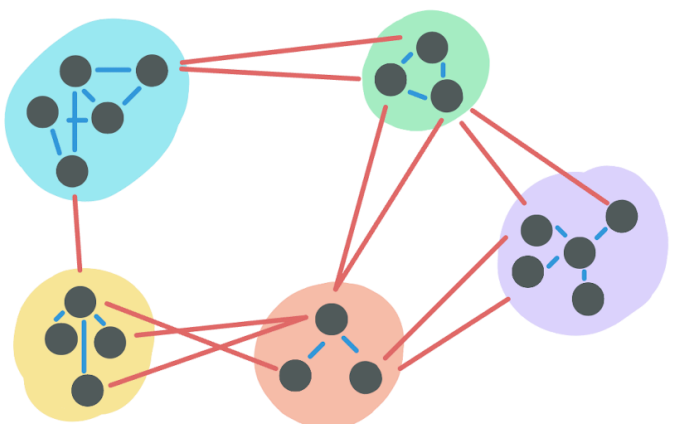
Bacterial interactions

In microbiotas, this leads to complex networks of interactions.

Subcommunities

Bacteria-bacteria interactions lead to non-random co-occurence patterns.

In the context of microbiotas, *subcommunities* are groups of bacteria that frequently co-occur.



- Subcommunities can co-exist in the same microbiota
- There can be *transitions* between subcommunity dominance

Duality: “types” ⇔ subcommunities

If there are subcommunities, we expect microbiota “types” to be associated with them.

And if there are microbiota “types”, we expect subcommunities membership to match the composition of prototypical “type” samples.

Outline

Hello 👋👋

Introduction

Topic models

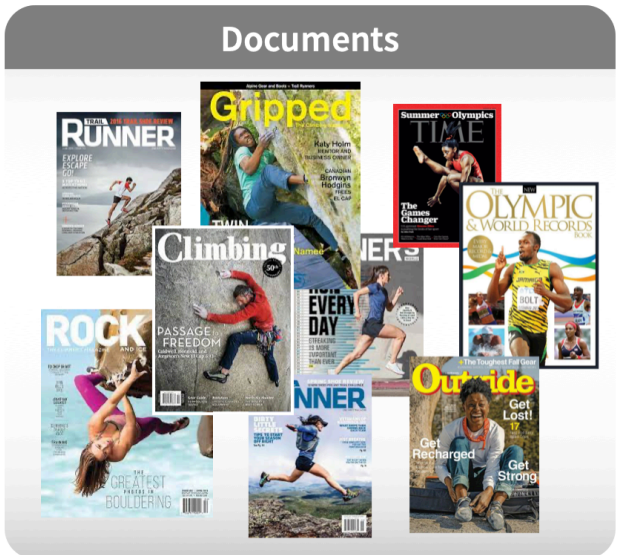
Choosing K, the number of topics

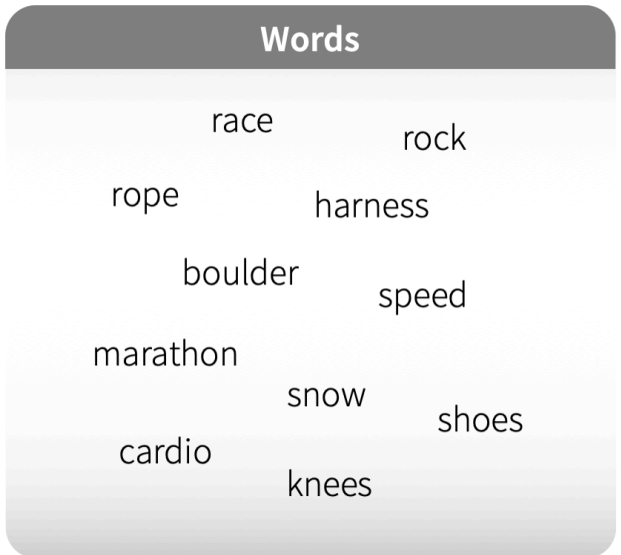
Conclusions & references

What are topic models?

Topic models

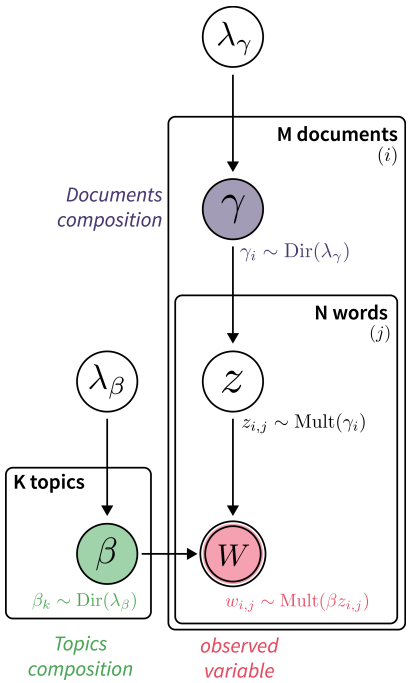
Topic models are statistical models for identifying "*topics*" in corpus of text documents.

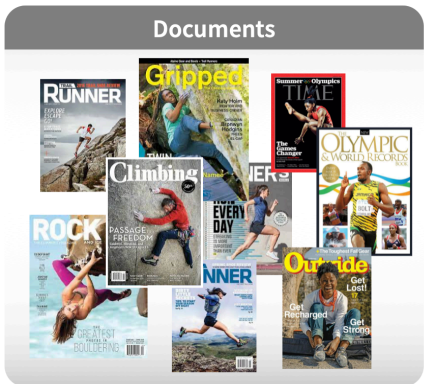


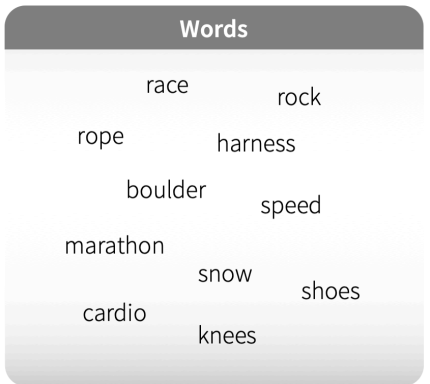


Several Bayesian and non-Bayesian models have been proposed:

LDA







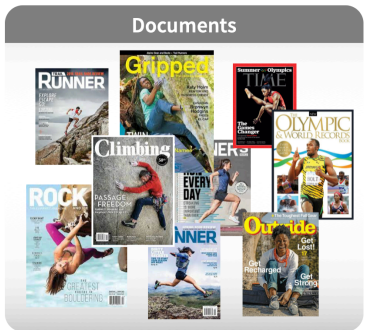
(Blei, Ng, and Jordan 2003)

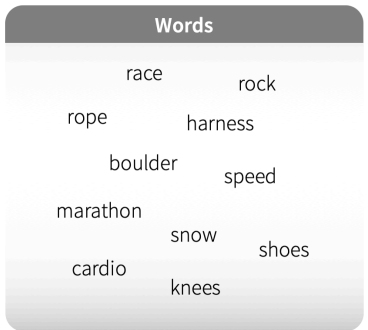
Parameter *inference* is typically done using Gibbs sampling or variational inference.

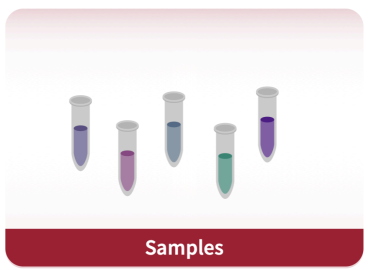
Several LDA implementations are available in R; among the best is the `topicmodels` package (Grün and Hornik 2011).

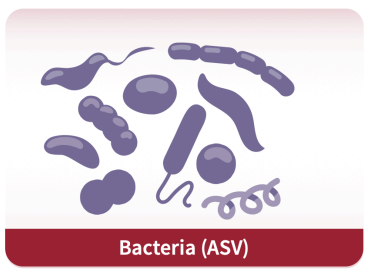
LDA for microbiota data?

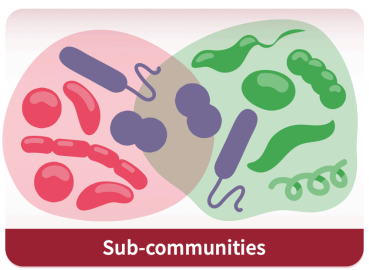
LDA for microbiota data











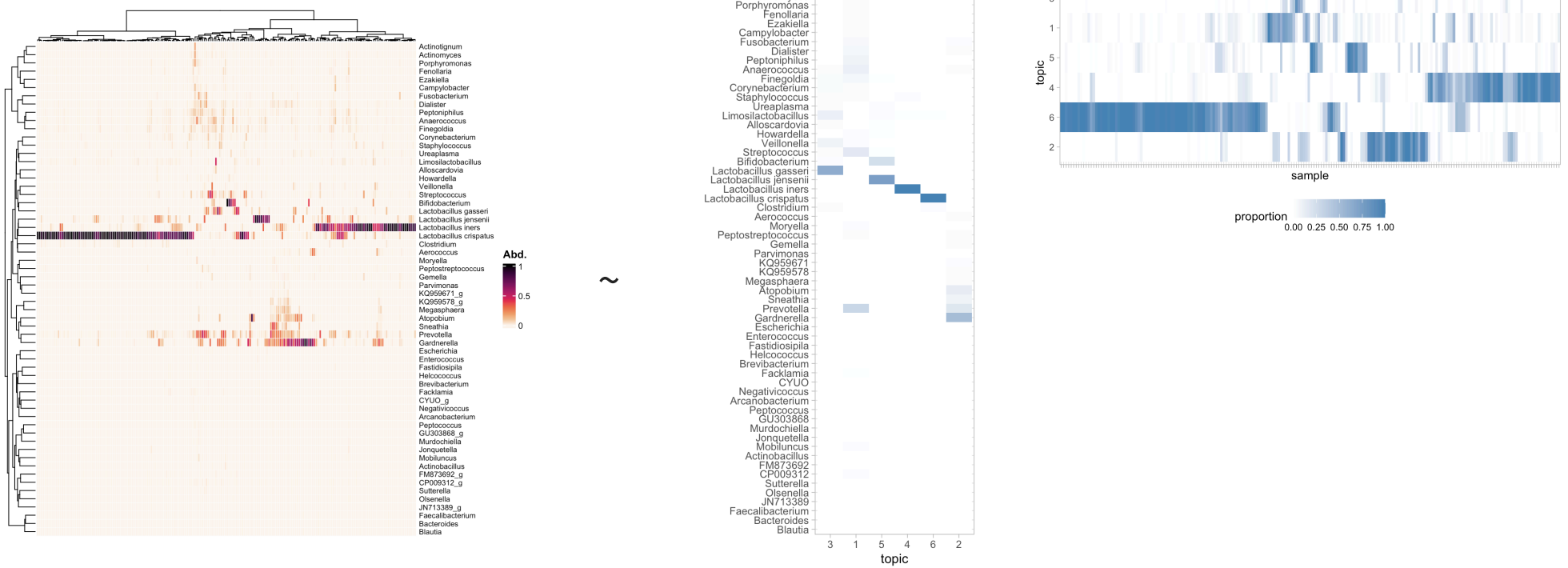
NLP	Microbiota analyses
Documents	samples
Words	taxa (ASV, species, genus, …)
Topics	sub-communities

(Sankaran and Holmes 2019; Symul et al. 2023)

LDA on microbiota data

ISALA (vag. swabs) | LifeLines DEEP (stool)

- Code



github.com/lasy/t-2425-Microbiota-UHasselt

LDA *vs* clustering

LDA *vs* clustering

LDA *vs* clustering

- In longitudinal data, clustering can't capture the *rate* of transitions
- Mixed memberships (topic models) are more powerful than categorical memberships (clustering) to find associations with responses.
- "Transition" clusters may group samples that are on very different trajectories.

Outline

Hello 👋👋

Introduction

Topic models

Choosing K, the number of topics

Conclusions & references

Perplexity-based methods

Log-likelihood and Perplexity

The log-likelihood of observing counts $\mathbf{X}$ $(\mathrm{I} \times \mathrm{J})$ given a model M with parameters β and γ is

$$\square(\mathbf{X}|\mathrm{M}) = \square(\mathbf{X}|\beta,\gamma) = \sum_{\mathrm{i}=1}^{\mathrm{I}} \log \mathrm{p}(\mathbf{x}_{\mathrm{i}\cdot}|\beta,\gamma) = \sum_{\mathrm{i}=1}^{\mathrm{I}} \sum_{\mathrm{j}=1}^{\mathrm{J}} \mathrm{x}_{\mathrm{ij}} \log(\beta\ \gamma)_{\mathrm{ij}}$$

The perplexity PP is then defined as $\mathrm{PP}(\mathbf{X}) = \exp\left(-\frac{\square(\mathbf{X}|\beta,\gamma)}{\sum_{\mathrm{i}=1}^{\mathrm{I}} \sum_{\mathrm{j}=1}^{\mathrm{J}} \mathrm{x}_{\mathrm{ij}}}\right)$

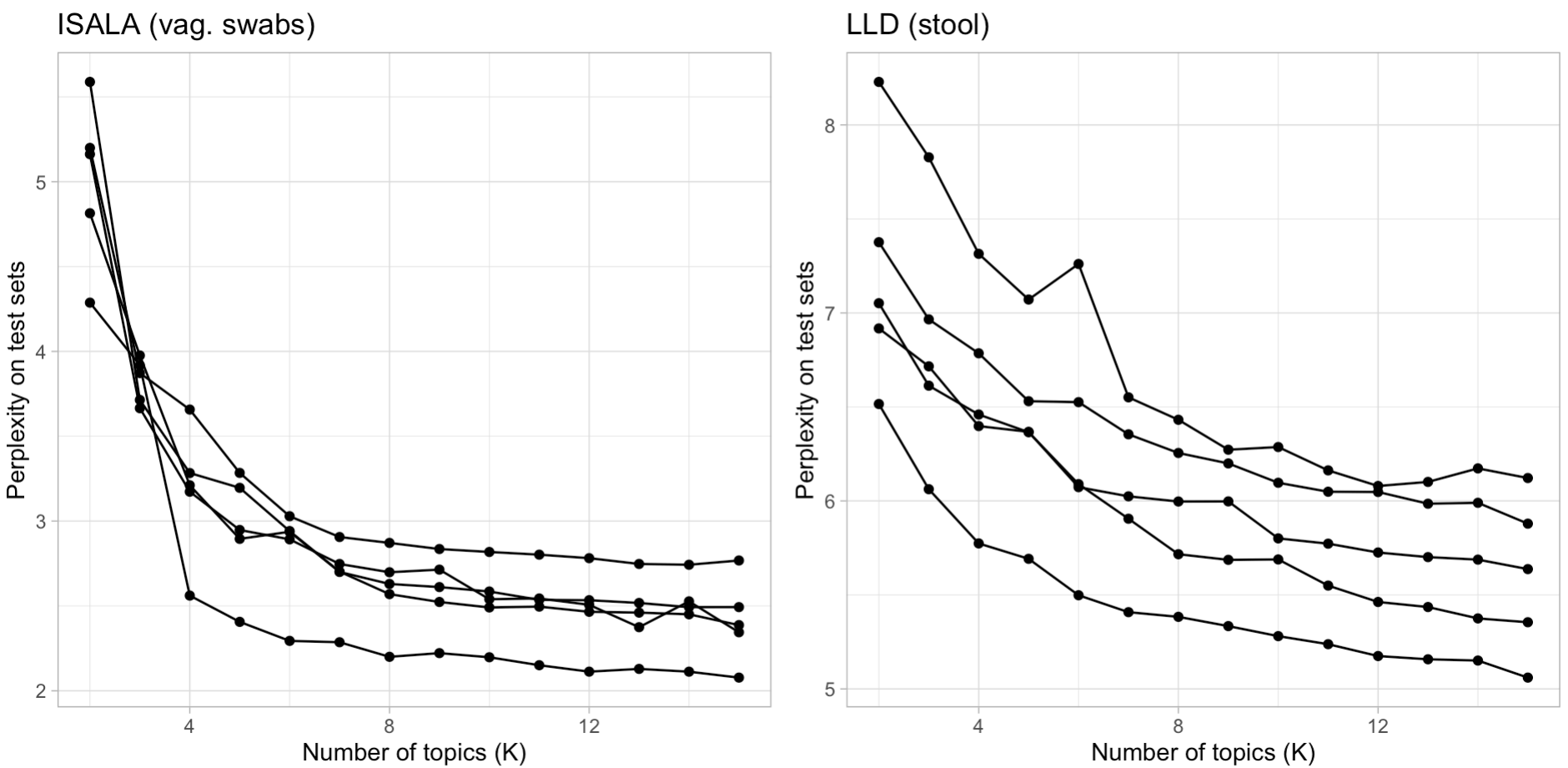
See (Grün and Hornik 2011) for details.

The perplexity is expected to decrease as the number of topics increases, so we look for an elbow.

▶ Code

Perplexity-based cross-validation

We can also choose K based on the perplexity in cross-validation.



github.com/lasy/t-2425-Microbiota-UHasselt

Alignement across K

(Fukuyama, Sankaran, and Symul 2023)

General idea

If some true K^* sub-communities exist, we expect the corresponding topics to be robustly identified by LDA at any $\mathrm{K} \geq \mathrm{K}^*$.

So, topics that are similar across different K are more likely to correspond to these true sub-communities.

And topics that are not similar across different K are more likely to be spurious.

How to identify topics that are similar across different K?

Matching topics across K (“Aligning” them)

How to match topics across different K?

Topics that are present in the same samples are likely to be similar.

ISALA (vag. swabs) | LLD (stool) - family level

Matching topics across K ("Aligning" them)

or…

Topics that have a similar composition are likely to be similar.

ISALA (vag. swabs) | LLD (stool) - family level

Duality: β & γ

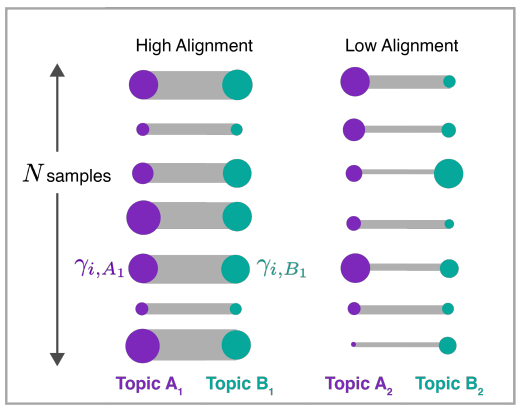
“Topics that are present in the same samples are likely to be similar”

$\longrightarrow$ Alignment based on the γ matrix (proportion of each topic in each sample)

“Topics that have a similar composition are likely to be similar”

$\longrightarrow$ Alignment based on the β matrix (probability of each taxon in each topic)

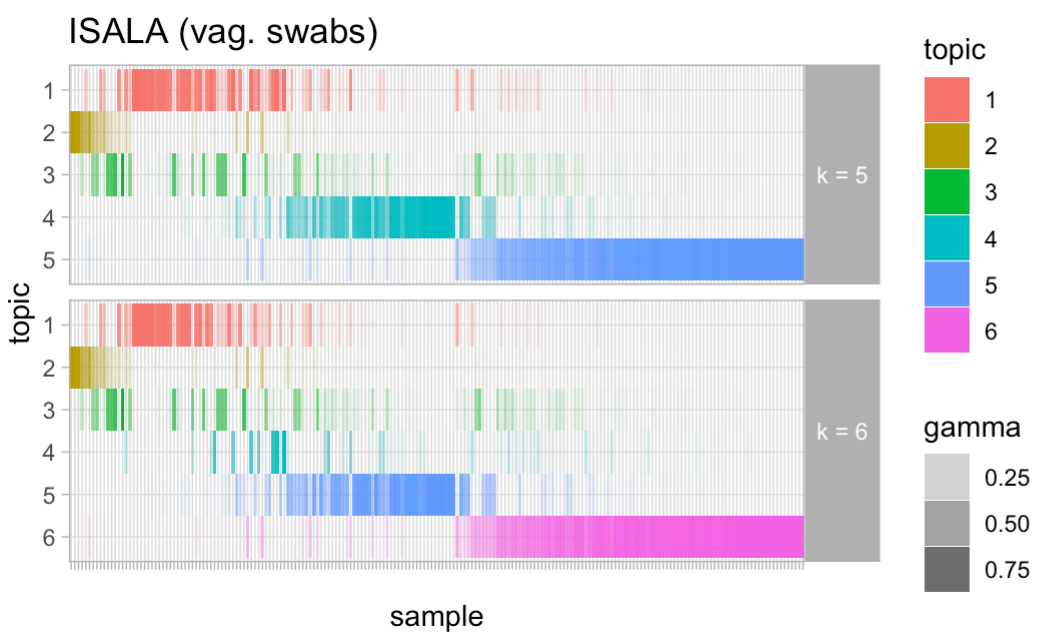
Sample-composition-based alignement (γ)

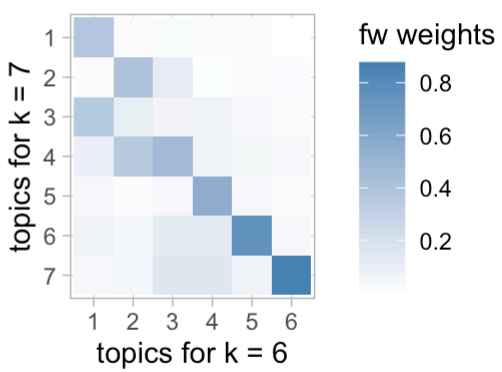


Alignment weights for the product method:

$$w_{jk} = \sum_{i=1}^{N} \gamma_{ij}^{(1)} \gamma_{ik}^{(2)}$$

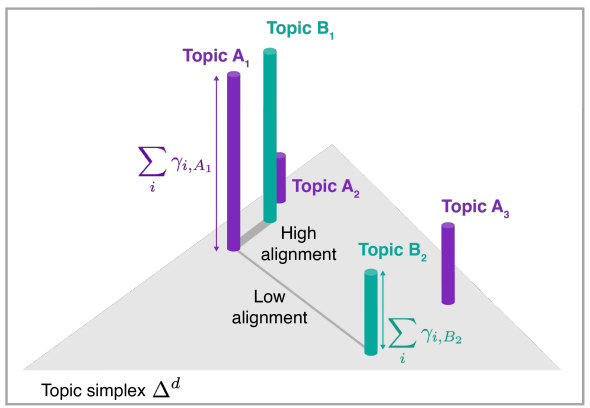
where $\gamma_{ij}^{(1)}$ are the proportions of topic j in sample i for model 1.





github.com/lasy/t-2425-Microbiota-UHasselt

Topic-composition-based alignement (β)



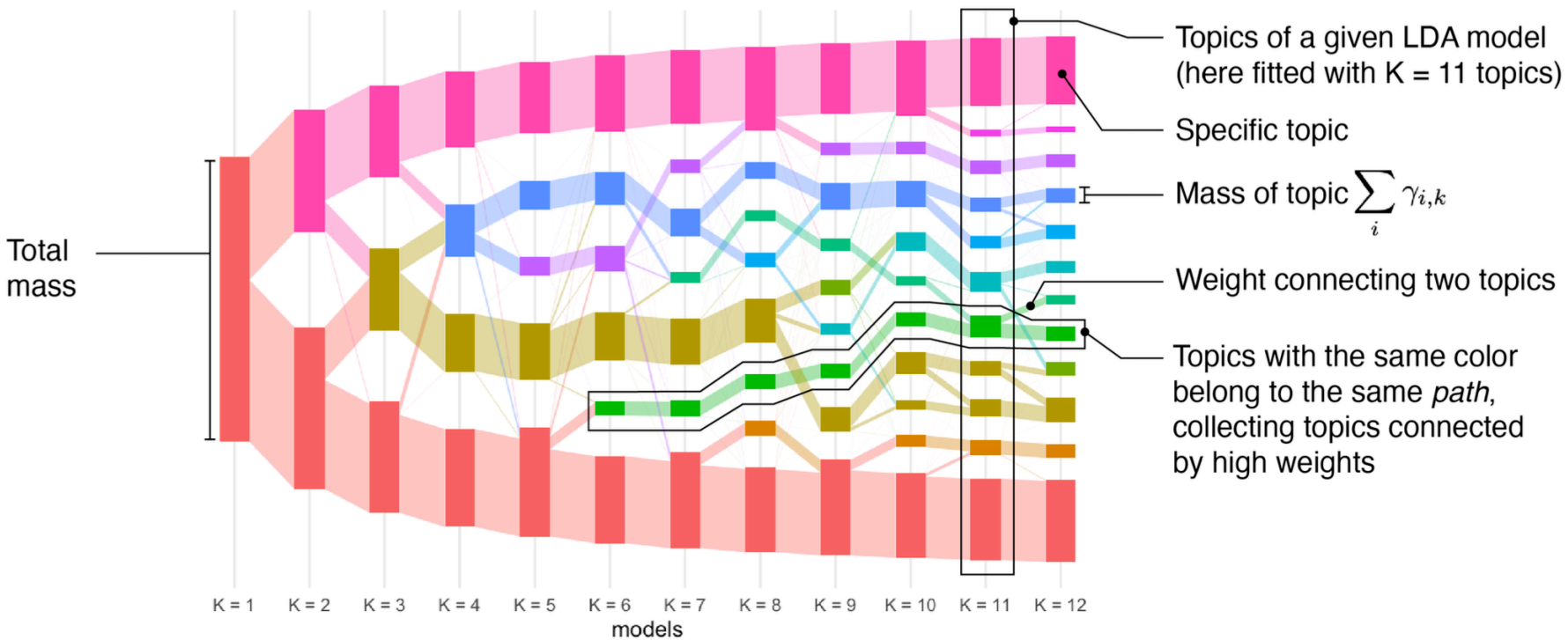
Alignment weights are obtained by solving the optimal transport problem

$$\min_{W \in \square(p,q)} \langle C, W \rangle$$

$$\square(p,q) := \{W \in R_+^{|V_p| \times |V_q|} : W\mathbf{1}_{|V_q|} = p \text{ and } W^T\mathbf{1}_{|V_p|}$$

where $C(v, v') := JSD(\beta(v), \beta(v'))$ is the **cost** of transporting mass between topics v and v'.

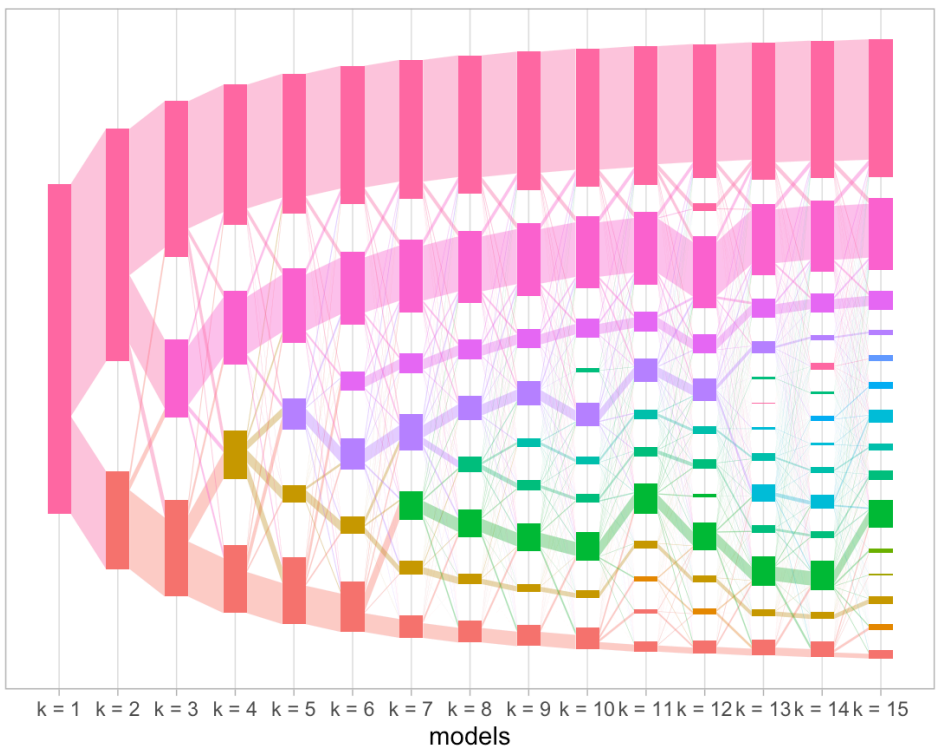
Alignment plots



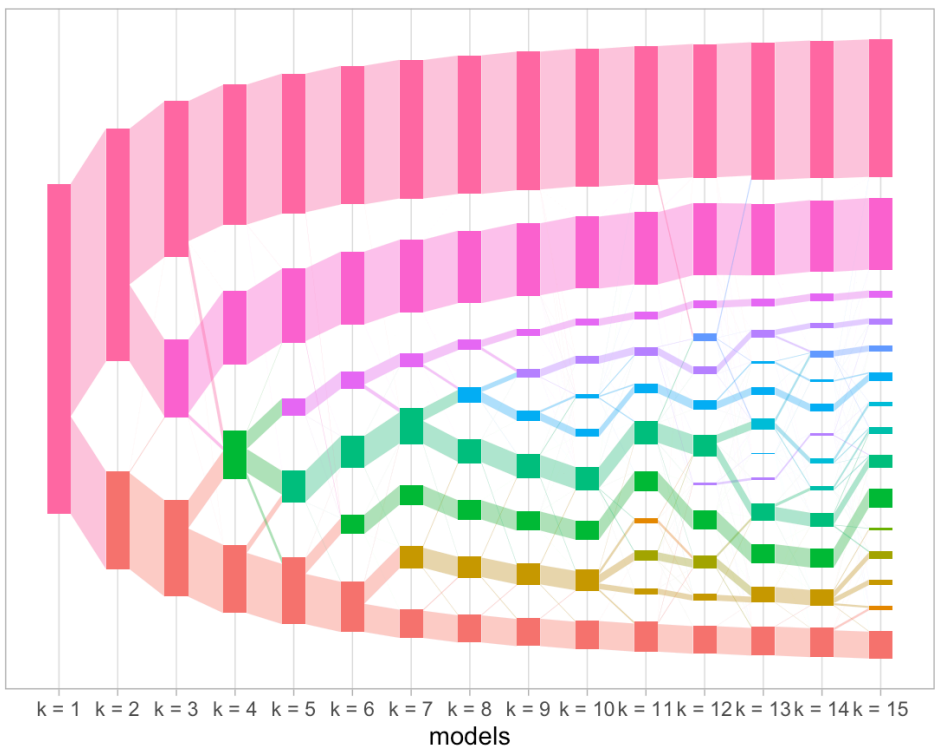
Alignment plots

Isala (vag. swabs) | LLD (stool)

Product alignment

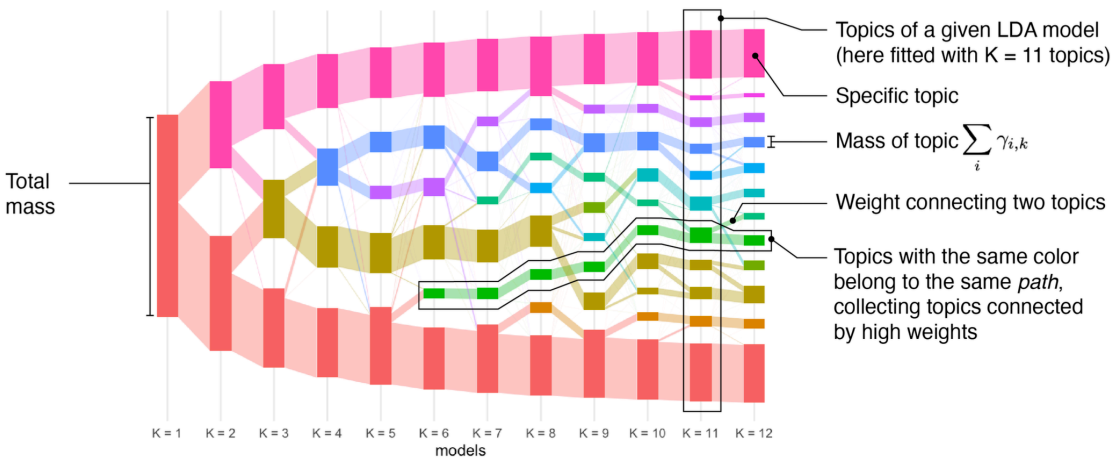


Transport alignment



github.com/lasy/t-2425-Microbiota-UHasselt

“True” *vs.* “Spurious” sub-communities: diagnostic scores



“True” *vs.* “Spurious” sub-communities: diagnostic scores

Isala (vag. swabs) | LLD (stool)

Outline

Hello 👋👋

Introduction

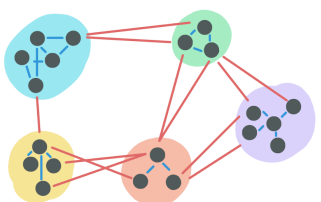
Topic models

Choosing K, the number of topics

Conclusions & references

Summary

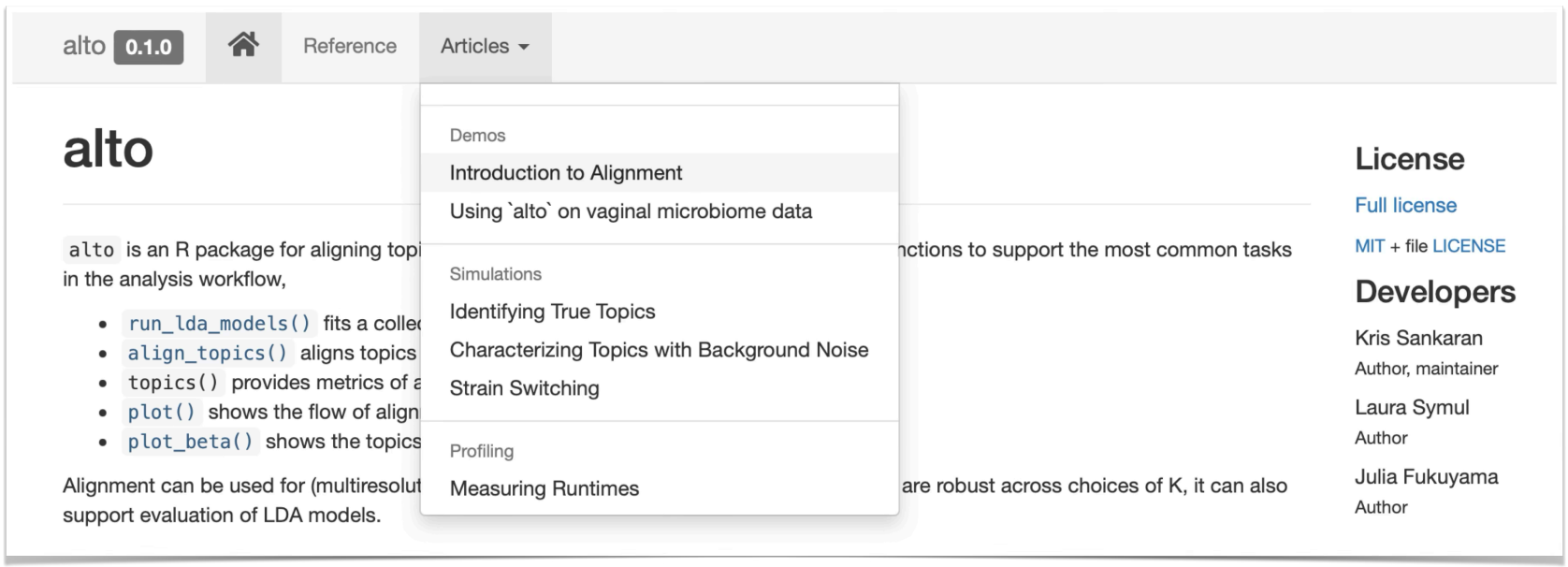
- Bacteria interact with each other in complex ways.
- These positive/negative interactions can lead to the existence of *subcommunities*: groups of bacteria that are more likely to be found together.



- Topic models (*e.g.*, LDA) can be used to identify subcommunities from microbiota count data.
- The number of topics K can be chosen based on (CV) perplexity or through **topic alignment** across K.

`alto` package

`alto` is an R package that implements the "topics alignment" methods:
lasy.github.io/alto



References

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. “Latent Dirichlet Allocation.” *J. Mach. Learn. Res.* 3 (null): 993–1022.

France, Michael T., Bing Ma, Pawel Gajer, Sarah Brown, Michael S. Humphrys, Johanna B. Holm, L. Elaine Waetjen, Rebecca M. Brotman, and Jacques Ravel. 2020. “VALENCIA: A Nearest Centroid Classification Method for Vaginal Microbial Communities Based on Composition.” *Microbiome* 8 (1): 166. https://doi.org/10.1186/s40168-020-00934-6.

Fukuyama, Julia, Kris Sankaran, and Laura Symul. 2023. “Multiscale Analysis of Count Data Through Topic Alignment.” *Biostatistics* 24 (4): 1045–65. https://doi.org/10.1093/biostatistics/kxac018.

Grün, Bettina, and Kurt Hornik. 2011. “Topicmodels: An r Package for Fitting Topic Models.” *Journal of Statistical Software*.

Lebeer, Sarah, Sarah Ahannach, Thies Gehrmann, Stijn Wittouck, Tom Eilers, Eline Oerlemans, Sandra Condori, et al. 2023. “A Citizen-Science-Enabled Catalogue of the Vaginal Microbiome and Associated Factors.” *Nature Microbiology* 8 (11): 2183–95. https://doi.org/10.1038/s41564-023-01500-0.

Thank you!

Questions?

github.com/lasy/t-2425-Microbiota-UHasselt