Integration of complex omics data through Multiscale Gaussian Graphical Models

Edmond Sanou

Laboratoire de Mathématiques et Modélisation d'Evry,
Université Paris-Saclay, CNRS, Univ Evry.

September 8th, 2023

PhD Supervisors

C. Ambroise

G. Robin

Climate change, trees and epigenetics (1)


Menu Libération Politique International CheckNews Culture Idées et Débats Société Enquêtes Environnement Ec

Environnement

En forêt de Vierzon, «les arbres meurent de soif»

Article réservé aux abonnés

Sciences-Po, le feuilleton d'une succession dossier

Dans le Cher, le massif domanial de Vierzon-Vouzeron peine à s'adapter au dérèglement climatique. Touchés par la sécheresse, chênes, hêtres, pins et épicéas dépérissent les uns après les autres... sous le regard quasi impuissant des forestiers.

Forest die-off

- Due to drought and heatwave.
- Need to understand the impacts of climate change on forests.
- Key issue: Trees' adaptability.

Climate change, trees and epigenetics (1)


Menu Libération Politique International CheckNews Culture Idées et Débats Société Enquêtes Environnement Ec

Environnement

En forêt de Vierzon, «les arbres meurent de soif»

Article réservé aux abonnés

Sciences-Po, le feuilleton d'une succession dossier

Dans le Cher, le massif domanial de Vierzon-Vouzeron peine à s'adapter au dérèglement climatique. Touchés par la sécheresse, chênes, hêtres, pins et épicéas dépérissent les uns après les autres... sous le regard quasi impuissant des forestiers.

Forest die-off

- Due to drought and heatwave.
- Need to understand the impacts of climate change on forests.
- Key issue: Trees' adaptability.

Climate change, trees and epigenetics (1)

Le Monde Environment | The Earth's Pulses Adaptation PFAS Sand Dealers

ENVIRONMENT • DROUGHT

Adapt or die: Plant life in France feels the effects of continuing drought

The soil moisture index reached an all-time low this month, as heat waves test the durability of the country's plant cover and scientists say some species could be in danger of disappearing.

By Adonis Leroyer

Published on August 19, 2022, at 2:02 pm (Paris), updated on August 19, 2022, at 2:02 pm · 4 min. · Lire en français

Forest die-off

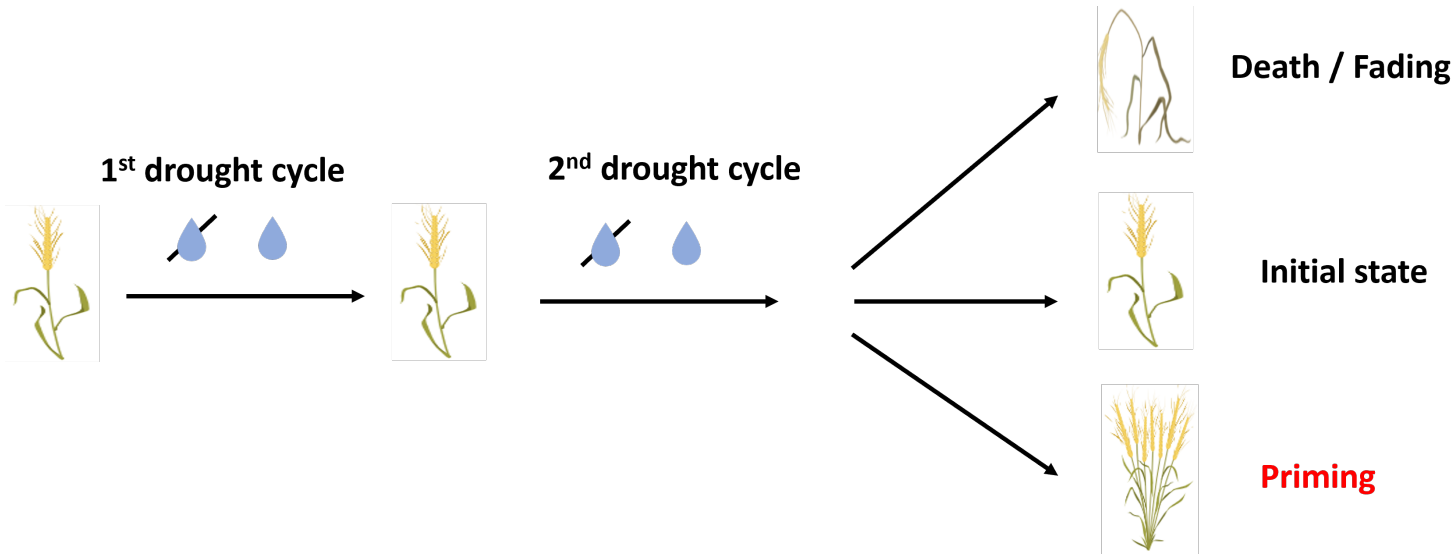
- Due to drought and heatwave.
- Need to understand the impacts of climate change on forests.
- Key issue: Trees' adaptability.

Climate change, trees and epigenetics (2)

Tree's adaptability study

- Genetics: DNA sequence and genetic evolution.
- Epigenetics: phenotypic plasticity and environment.

Example: Epigenetics and priming in plants



What about trees?

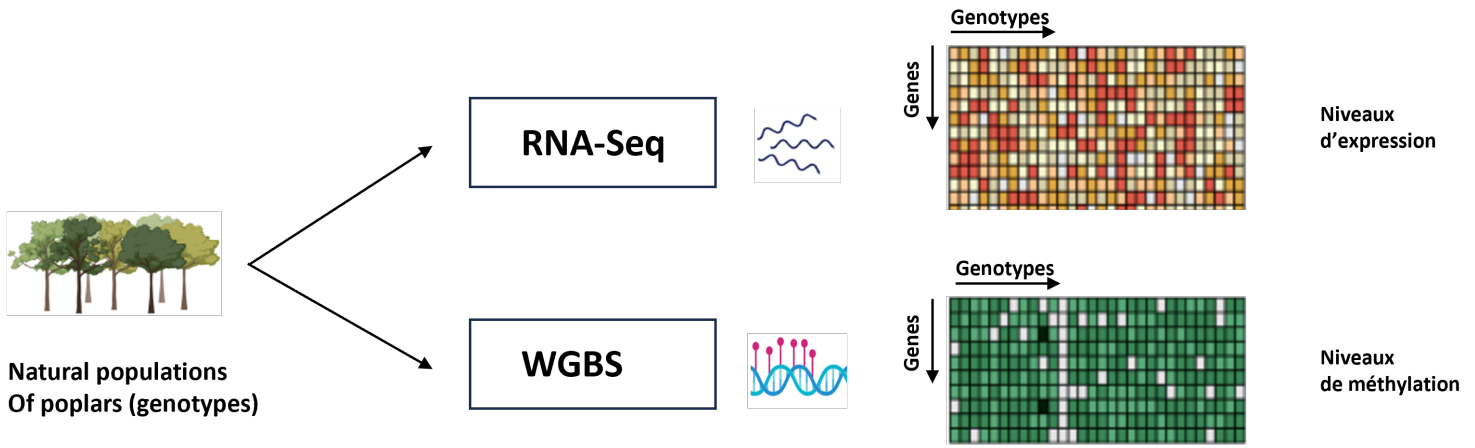
EPITREE project

Evolutionary and functional impact of epigenetic variation in forest trees.

Data

Various omics data produced (transcriptomic, DNA methylation, SNPs) and phenotypes data.

Example

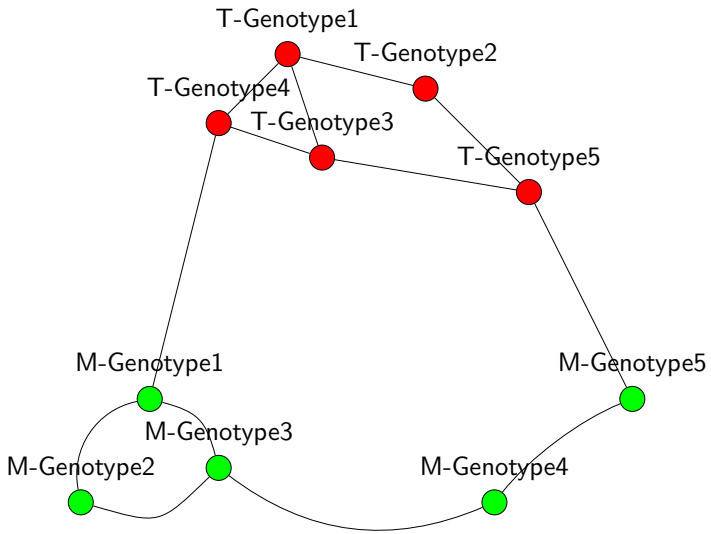


Interactions in multi-omics data

Genotype-Genotype networks

- Co-regulation patterns and identification of strongly associated genotypes across omics.

Example: One-level network on Transcriptomic and Methylation genotypes



Constraints on building the networks

- Sparsity
- Clustering structure

Two-level networks

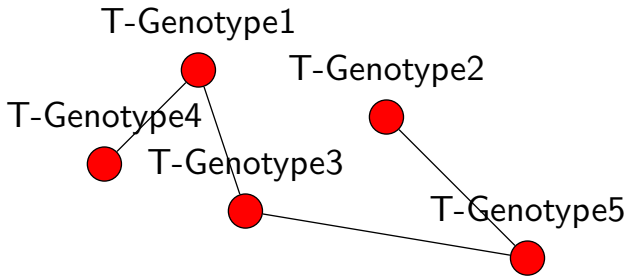
Cheng et al. (2017)

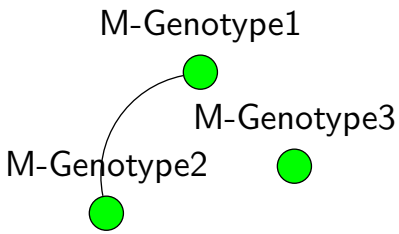
- One level describes networks for variables.

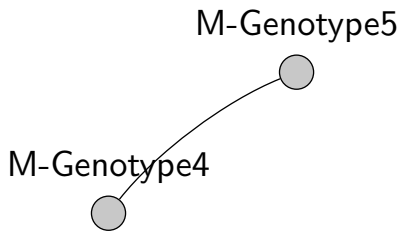
Two-level networks

Cheng et al. (2017)

- One level describes networks for variables.



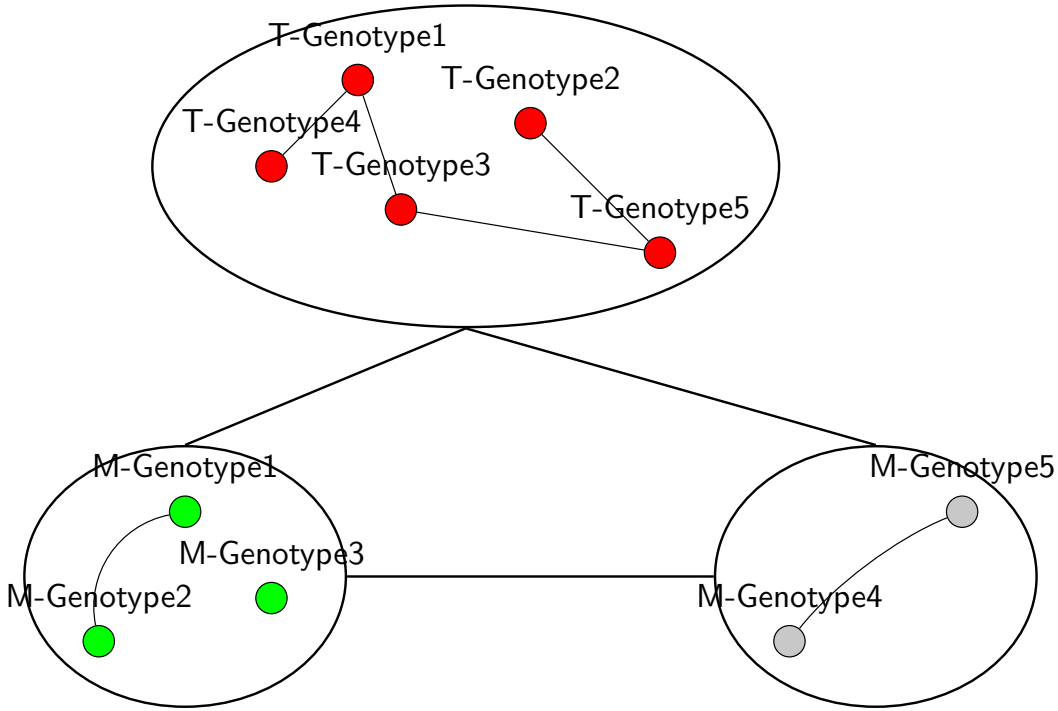




Two-level networks

Cheng et al. (2017)

- One level describes networks for variables.
- The other level describes networks for known groups.



Two-level networks

Cheng et al. (2017)

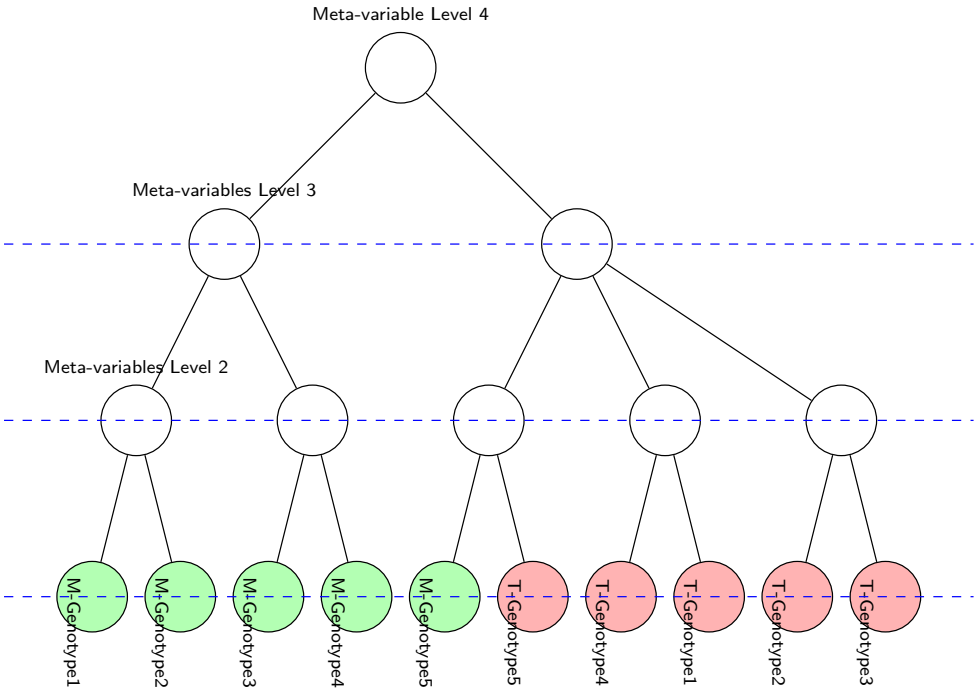
- One level describes networks for variables.
- The other level describes networks for known groups.

Drawbacks

- Known groups
- Only two levels of granularity

Multilevel/Multiscale networks

Sanou et al. (2023)



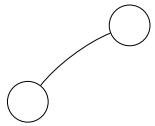


Figure: Level 3: compressed view

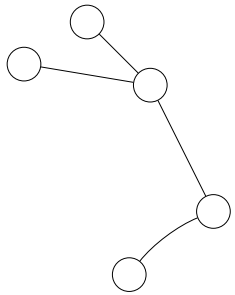


Figure: Level 2: compressed view

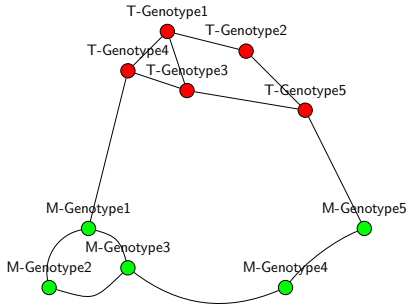
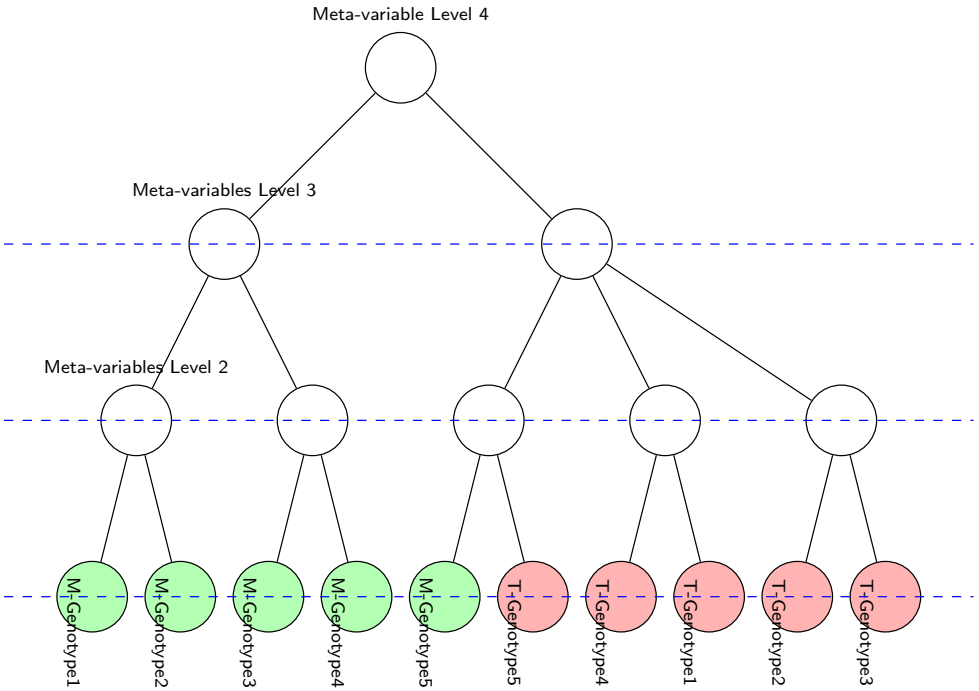


Figure: Level 1

Multilevel/Multiscale networks

Sanou et al. (2023)



Challenge: simultaneous clustering and network inference

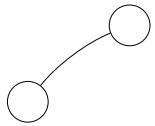


Figure: Level 3: compressed view

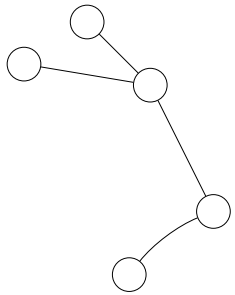


Figure: Level 2: compressed view

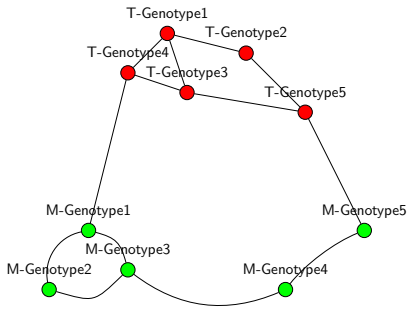


Figure: Level 1

Mathematical framework

i Graphical Models

ii Convex clustering

Undirected Graphical models

- Popular computational analysis tool for biological data.

Undirected Graphical models

- Popular computational analysis tool for biological data.

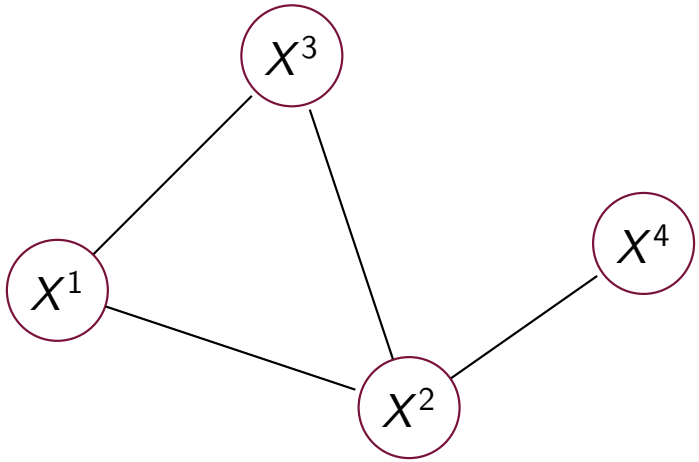


Figure: Undirected graph.

Undirected Graphical models

- Popular computational analysis tool for biological data.
- Describe statistical dependencies between variables.

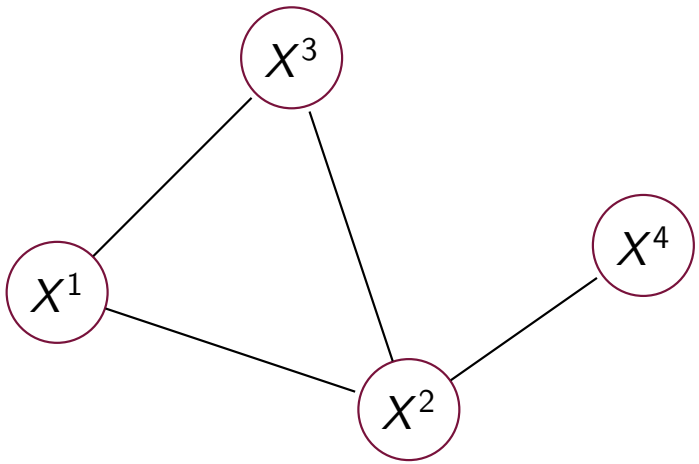


Figure: Undirected graph.

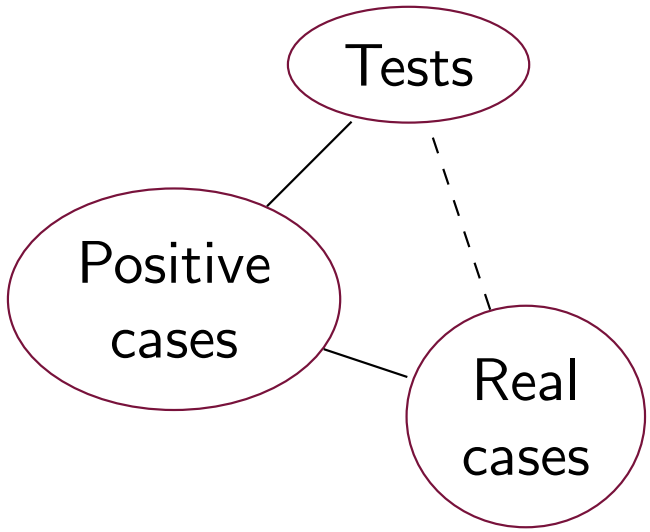
- Connected: conditional dependence

  Markov property: X^4 is independent from (X^1, X^3) conditionally to X^2

Undirected Graphical models

- Popular computational analysis tool for biological data.
- Describe statistical dependencies between variables.

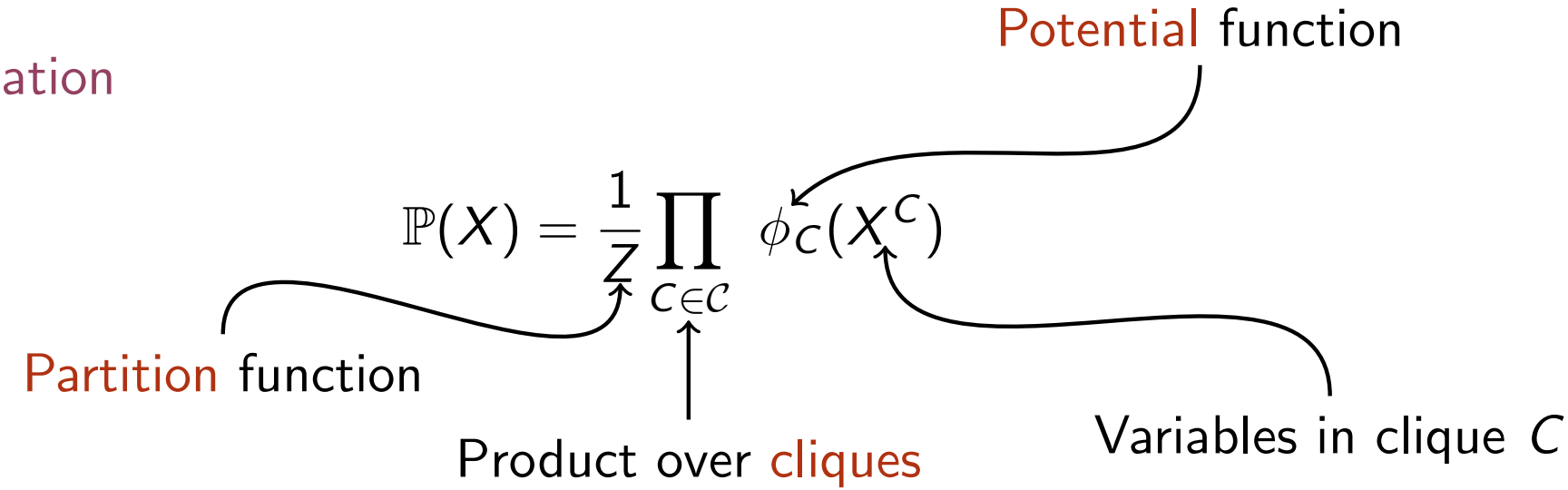
Example: spurrious correlation



Undirected Graphical Models

Pair $(G, \mathbb{P})$ where $\mathbb{P}$ is a distribution that factorizes with respect to the graph G.
Let $X = (X^1, \ldots, X^p)$ be a random vector taking values in χ.

Factorization



Some distributions for ϕ_C

- Ising model: $\chi = \{0, 1\}$
- Potts model: $\chi = \{1, \ldots, m\}$
- Gaussian model: $\chi = \mathbb{R}$

Gaussian Graphical Models (GGM)

Gaussian distribution

Let $\boldsymbol{X} \sim \mathcal{N}(0_p, \boldsymbol{\Sigma})$ with precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$.

$$\mathbb{P}_{\boldsymbol{\Omega}}(\boldsymbol{X}) = \frac{(\det(\boldsymbol{\Omega}))^{1/2}}{(2\pi)^{p/2}} \exp\left\{-\frac{1}{2}\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X}\right\}$$

Gaussian Graphical Models (GGM)

Gaussian distribution

Let $\boldsymbol{X} \sim \mathcal{N}(0_p, \boldsymbol{\Sigma})$ with precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$.

$$\mathbb{P}_{\boldsymbol{\Omega}}(\boldsymbol{X}) = \frac{1}{Z_{\boldsymbol{\Omega}}} \exp\left\{ -\frac{1}{2} \sum_{s,t=1}^{p} \Omega_{st} X^s X^t \right\}$$

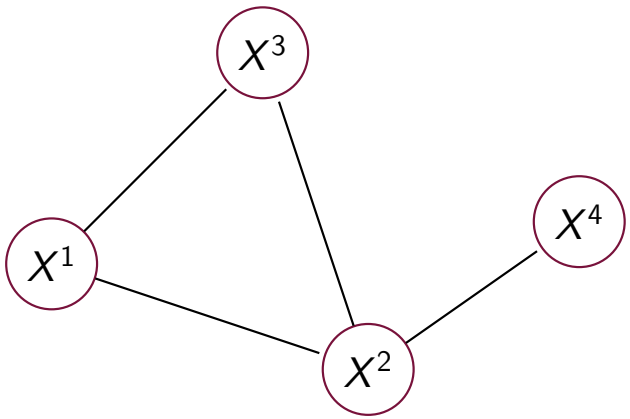
Gaussian Graphical Models (GGM)

Gaussian distribution

Let $\boldsymbol{X} \sim \mathcal{N}(0_p, \boldsymbol{\Sigma})$ with precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$.

$$\mathbb{P}_{\boldsymbol{\Omega}}(\boldsymbol{X}) = \frac{1}{Z_{\boldsymbol{\Omega}}} \exp\left\{ -\frac{1}{2} \sum_{s,t=1}^{p} \Omega_{st} X^s X^t \right\}$$

Based on the factorization (Hammersley-Clifford)



$$\Longleftrightarrow \qquad \boldsymbol{\Omega} = \begin{pmatrix} * & * & * & 0 \\ * & * & * & * \\ * & * & * & 0 \\ 0 & * & 0 & * \end{pmatrix}$$

Maximum likelihood estimator

- Log-likelihood

$$\log \mathcal{L}(\boldsymbol{X}, \boldsymbol{\Omega}) = \sum_{i=1}^{n} \log \mathbb{P}(X_i; \boldsymbol{\Omega}) \propto \log \det(\boldsymbol{\Omega}) - \operatorname{tr}(\boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\Omega})$$

- Solution

$$\widehat{\boldsymbol{\Omega}}_{MLE} = \underset{\boldsymbol{\Omega}}{\operatorname{argmax}} \log \mathcal{L}(\boldsymbol{X}, \boldsymbol{\Omega}) = \mathbf{S}^{-1}$$

Drawbacks

- Singularity of the covariance matrix in high dimension
- Complete graphs

Maximum pseudo-likelihood estimator

- Conditional distribution

$$X^i | X^{\setminus i} = X^{\setminus i} \sim \mathcal{N}\left(-\Omega_{ii}^{-1}\Omega_{i\setminus i}X^{\setminus i}, \quad \Omega_{ii}^{-1}\right).$$

- Solving p linear regressions

$$\boldsymbol{X}^i = \boldsymbol{X}^{\setminus i}\boldsymbol{\beta}^i + \boldsymbol{\epsilon}, \qquad \text{where} \quad \boldsymbol{\beta}^i = -\boldsymbol{\Omega}_{i\setminus i}/\Omega_{ii}, \forall i = 1, \ldots, p$$

- A pseudo-likelihood criterion (Ambroise et al., 2009)

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \log \tilde{\mathcal{L}}(\boldsymbol{X}, \boldsymbol{\beta}) = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \sum_{i=1}^{n}\sum_{j=1}^{p} \log P\left(X_j^i | X_j^{\setminus i}; \boldsymbol{\beta}^i\right)$$

Estimation with sparsity constraint

Neighborhood selection (Meinshausen and Bühlmann, 2006)

- Solve p independant Lasso regression problems:

$$\hat{\boldsymbol{\beta}}^i(\lambda) = \underset{\boldsymbol{\beta}^i}{\operatorname{argmin}} \frac{1}{n} ||\boldsymbol{X}^i - \mathbf{X}^{\setminus i}\boldsymbol{\beta}^i||_2^2 + \lambda ||\boldsymbol{\beta}^i||_1, \forall i = 1, \ldots, p$$

Drawbacks

- Limitations of Lasso in the presence of strongly correlated variables

Alternative approaches in single task regression problems

- In two-step: clustering followed by regression e.g. Cluster representative Lasso (Bühlmann et al., 2013)
- In one-step: penalty functions for grouping e.g. Fused-Lasso (Tibshirani et al., 2005)

Challenges in multitask learning

Framework

- p responses variables: $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^p \in \mathbb{R}^n$
- p sets of explanatory variables $\boldsymbol{X}^{\setminus 1}, \ldots, \boldsymbol{X}^{\setminus p} \in \mathbb{R}^{n,p-1}$
- p regression vectors $\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^p \in \mathbb{R}^{p-1}$ giving the neighborhood structure the variables
- The pseudo log-likelihood writes:

$$\frac{1}{2} \sum_{i=1}^{p} \left\| \mathbf{x}^i - \mathbf{x}^{\setminus i} \boldsymbol{\beta}^i \right\|_2^2 + \lambda_1 \sum_{i=1}^{p} \left\| \boldsymbol{\beta}^i \right\|_1$$

Challenge: penalty for grouping over non-fixed set of covariates

$$\frac{1}{2} \sum_{i=1}^{p} \left\| \mathbf{x}^i - \mathbf{x}^{\setminus i} \boldsymbol{\beta}^i \right\|_2^2 + \lambda_1 \sum_{i=1}^{p} \left\| \boldsymbol{\beta}^i \right\|_1 + pen(\boldsymbol{\beta}^i, \ldots, \boldsymbol{\beta}^p)$$

Convex clustering

Let $\boldsymbol{x}_i \in \mathbb{R}^p$ be datapoints and $\boldsymbol{\alpha}_i \in \mathbb{R}^p$ corresponding centroids with $i = 1, \ldots, n$.
Convex clustering (Pelckmans et al., 2005; Hocking et al., 2011; Lindsten et al., 2011) solves:

$$\underset{\boldsymbol{\alpha} \in \mathbb{R}^{n \times p}}{\operatorname{argmin}} \underbrace{\frac{1}{2} \sum_{i=1}^{n} ||x_i - \boldsymbol{\alpha}_i||_2^2}_{\textit{Clustering loss}} + \underbrace{\lambda \sum_{i<j} ||\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_j||_q}_{\textit{Shrinkage term}}$$

- Global optimum reached and independent of initialization.
- λ determines the number of clusters.
- Choice of the q-norm influences solutions' path.

Clustering path

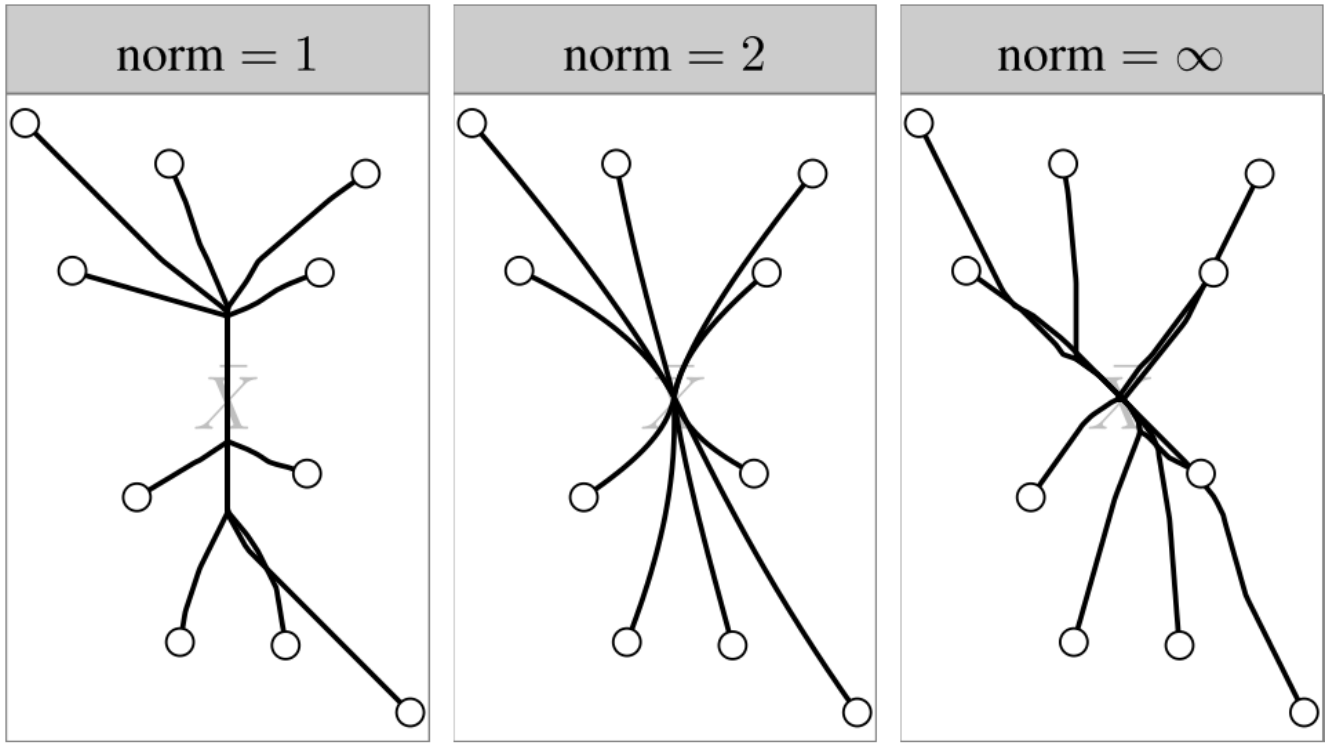


Figure: Clustering path for different norms (Hocking et al., 2011)

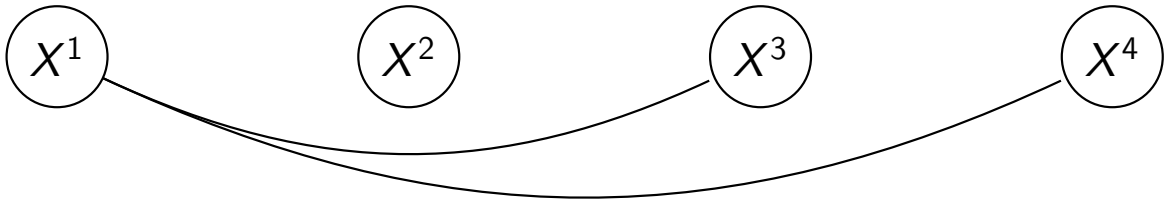
Multiscale Gaussian Graphical Model

i Model

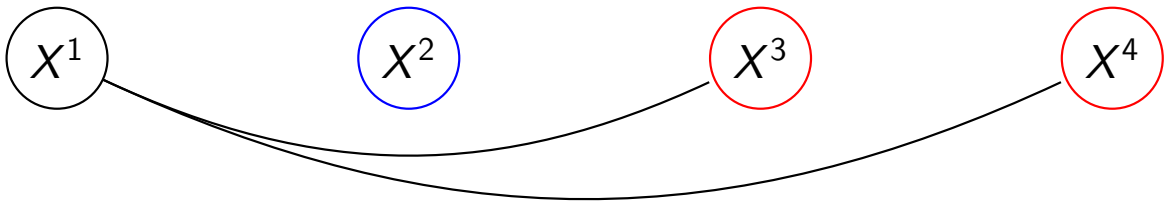
ii Inference

Intuition

- Identify the neighborhood of each node
  $\rightarrow$ via Lasso selection: $\sum_{i=1}^{p} \left\| \beta^i \right\|_1$



- Group variables with similar neighborhood structure (Local constancy)
  $\rightarrow$ via group-fused Lasso: $\sum_{i<j} \left\| \beta^i - \tau_{ij} \beta^j \right\|_2$



Formulation

The Multiscale Graphical LASSO (MGLASSO, Sanou et al. (2023)) problem minimizes the following pseudo-likelihood criterion:

$$\mathcal{J}_{\lambda_1,\lambda_2}(\boldsymbol{\beta};\mathbf{X}) = \underbrace{\frac{1}{2}\sum_{i=1}^{p}\left\|\mathbf{X}^i - \mathbf{X}^{\setminus i}\boldsymbol{\beta}^i\right\|_2^2 + \lambda_1\sum_{i=1}^{p}\left\|\boldsymbol{\beta}^i\right\|_1}_{\textit{Neighborhood selection}} + \underbrace{\lambda_2\sum_{i<j}\left\|\boldsymbol{\beta}^i - \boldsymbol{\tau}_{ij}\boldsymbol{\beta}^j\right\|_2}_{\textit{convex clustering}},$$

with $\left\|\boldsymbol{\beta}^i - \boldsymbol{\tau}_{ij}\boldsymbol{\beta}^j\right\|_2 = \sqrt{\sum_{k\in\{1,\ldots,p\}\setminus\{i,j\}}(\beta_k^i - \beta_k^j)^2 + (\beta_j^i - \beta_i^j)^2}$.

Formulation

The Multiscale Graphical LASSO (MGLASSO, Sanou et al. (2023)) problem minimizes the following pseudo-likelihood criterion:

$$\mathcal{J}_{\lambda_1,\lambda_2}(\boldsymbol{\beta};\mathbf{X}) = \underbrace{\frac{1}{2}\sum_{i=1}^{p}\left\|\mathbf{X}^i - \mathbf{X}^{\setminus i}\boldsymbol{\beta}^i\right\|_2^2}_{Smooth} + \underbrace{\lambda_1\sum_{i=1}^{p}\left\|\boldsymbol{\beta}^i\right\|_1 + \lambda_2\sum_{i<j}\left\|\boldsymbol{\beta}^i - \boldsymbol{\tau}_{ij}\boldsymbol{\beta}^j\right\|_2}_{Non\ smooth}$$

Non-smooth convex optimization

- Subgradient method (Shor, 2012)
- Alternating Direction Method of Multipliers (Boyd et al., 2011)
- Continuation method based on smoothing criterion (Hadj-Selem et al., 2018)

Optimization

- CONESTA: COntinuation with NEsterov smoothing in a Shrinkage-Thresholding Algorithm (Hadj-Selem et al., 2018)
- Dedicated to regression problems with structured sparsity.

CONESTA problem form

$$f(\theta) = \underbrace{g(\theta)}_{smooth} + \underbrace{\lambda_1 h(\theta) + \lambda_2 s(\theta)}_{non-smooth},$$

with

- $g(\theta)$: convex and smooth criterion
- $h(\theta)$: convex and non-smooth penalty whose proximal is known
- $s(\theta)$: convex and non-smooth penalty with an explicit max-structure:

$$s(\theta) = \max_{\mathbf{u} \in U} \left\{ \mathbf{u}^\top \mathbf{A}\theta - \phi(\mathbf{u}) \right\}$$

Reformulation

The objective of MGLASSO reformulates as follows:

$$f(\widetilde{\boldsymbol{\beta}}) = \frac{1}{2}||\mathbf{Y} - \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}}||_2^2 + \lambda_1||\widetilde{\boldsymbol{\beta}}||_1 + \lambda_2 \sum_{i<j} ||\boldsymbol{D}_{ij}\widetilde{\boldsymbol{\beta}}||_2$$

$$= \frac{1}{2}||\mathbf{Y} - \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}}||_2^2 + \lambda_1||\widetilde{\boldsymbol{\beta}}||_1 + \lambda_2 \max_{\boldsymbol{\alpha}} \left\{ \boldsymbol{\alpha}^\top \mathbf{D}\widetilde{\boldsymbol{\beta}} \right\},$$

Steps

- Nesterov smoothing:
  - $f_\mu(\widetilde{\boldsymbol{\beta}}) = \frac{1}{2}||\mathbf{Y} - \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}}||_2^2 + \lambda_1||\widetilde{\boldsymbol{\beta}}||_1 + \lambda_2 \max_{\boldsymbol{\alpha}} \left\{ \boldsymbol{\alpha}^\top \mathbf{D}\widetilde{\boldsymbol{\beta}} - \frac{\mu}{2}\|\boldsymbol{\alpha}\|_2^2 \right\}$
  - Good approximation of $f(\widetilde{\boldsymbol{\beta}})$ when μ is small.
- Key idea of CONESTA: Start with a relative large value of μ and decrease it with respect to the distance to the minimum.
- The duality gap provides an upper bound of the distance: $\text{GAP}(\widetilde{\boldsymbol{\beta}}^{(k)}) \geq f(\widetilde{\boldsymbol{\beta}}^{(k)}) - f(\widetilde{\boldsymbol{\beta}}^\star) \geq 0$.

In practice

MGLASSO learning

The selection problem operates at two levels:

- λ_1 chosen via model selection using the StARS approach
- λ_2 varies across a grid of values to obtain graphs with different levels of granularity for a fixed value of λ_1.

Performances in support recovery (stochastic block model)

- Adding a fusion penalty improves AUC compared to GLASSO

Performances in clustering (hierarchically structured simulation model)

- Good adjusted Rand indices for $n/p = 0.5$ and higher levels in the hierarchical clustering tree
- Can't conclude for lower levels

Illustration on multi-omics data

Natural populations of trees

- Model tree: poplar

Figure: Poplar and map of natural populations.

Natural populations of trees

- Model tree: poplar
- 10 natural populations sampled in Western Europe

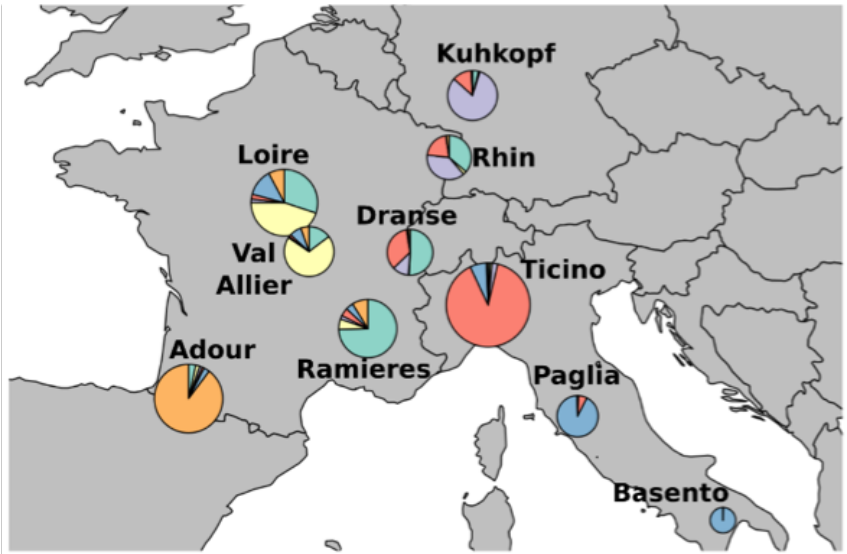


Figure: Poplar and map of natural populations.

Natural populations of trees

- Model tree: poplar
- 10 natural populations sampled in Western Europe
- 2 genotypes per population $\rightarrow$ 20 genotypes.

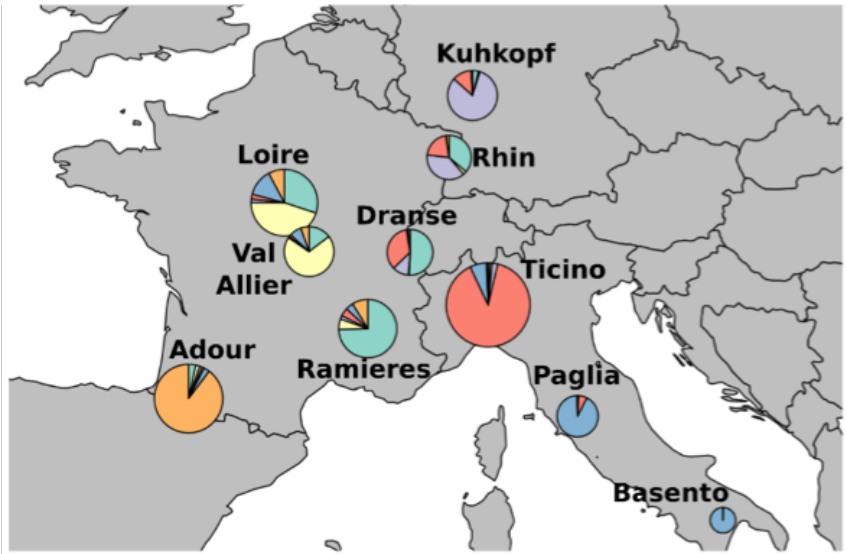


Figure: Poplar and map of natural populations.

Multi-omic data

Omics measured for 20 genotypes (**variables**) for genotype-genotype networks

Transcriptomic (Gene expression): RNA-Seq data (Chateigner et al., 2020)

- 34229 genes (**samples**).
- Provided data: continuous and Gaussian.

Epigenomic (DNA methylation): WGBS data (Sow et al., 2023)

Multi-omic data

Omics measured for 20 genotypes (**variables**) for genotype-genotype networks

Transcriptomic (Gene expression): RNA-Seq data (Chateigner et al., 2020)

- 34229 genes (**samples**).
- Provided data: continuous and Gaussian.

Epigenomic (DNA methylation): WGBS data (Sow et al., 2023)

Multi-omic data

Omics measured for 20 genotypes (**variables**) for genotype-genotype networks

Transcriptomic (Gene expression): RNA-Seq data (Chateigner et al., 2020)

- 34229 genes (**samples**).
- Provided data: continuous and Gaussian.

Epigenomic (DNA methylation): WGBS data (Sow et al., 2023)

Multi-omic data

Omics measured for 20 genotypes (**variables**) for genotype-genotype networks

Transcriptomic (Gene expression): RNA-Seq data (Chateigner et al., 2020)

- 34229 genes (**samples**).
- Provided data: continuous and Gaussian.

Epigenomic (DNA methylation): WGBS data (Sow et al., 2023)

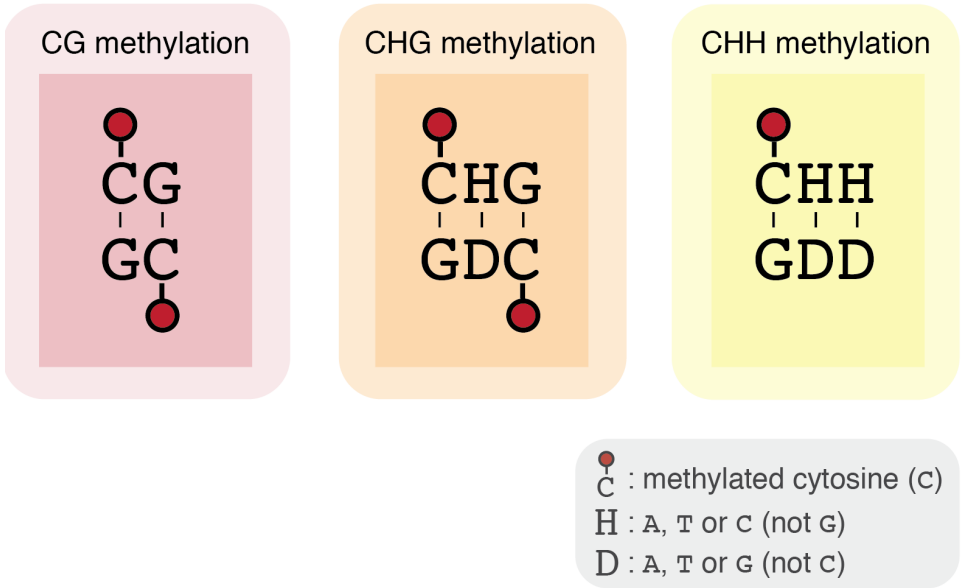
Multi-omic data

Omics measured for 20 genotypes (**variables**) for genotype-genotype networks

Transcriptomic (Gene expression): RNA-Seq data (Chateigner et al., 2020)

- 34229 genes (**samples**).
- Provided data: continuous and Gaussian.

Epigenomic (DNA methylation): WGBS data (Sow et al., 2023)



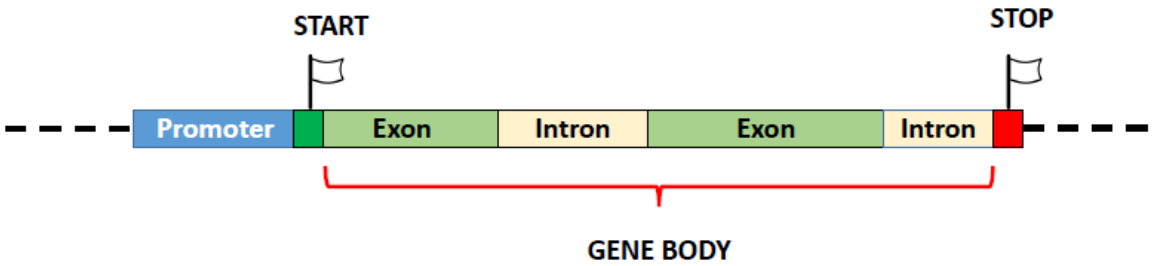
Multi-omic data

Omics measured for 20 genotypes (**variables**) for genotype-genotype networks

Transcriptomic (Gene expression): RNA-Seq data (Chateigner et al., 2020)

- 34229 genes (**samples**).
- Provided data: continuous and Gaussian.

Epigenomic (DNA methylation): WGBS data (Sow et al., 2023)



Multi-omic data

Omics measured for 20 genotypes (**variables**) for genotype-genotype networks

Transcriptomic (Gene expression): RNA-Seq data (Chateigner et al., 2020)

- 34229 genes (**samples**).
- Provided data: continuous and Gaussian.

Epigenomic (DNA methylation): WGBS data (Sow et al., 2023)

- Three methylation contexts measured on gene-body and gene-promoter.
- $\rightarrow$ 6 data-sets each containing about 40000 genes (**samples**).

Multi-omic data

Omics measured for 20 genotypes (**variables**) for genotype-genotype networks

Transcriptomic (Gene expression): RNA-Seq data (Chateigner et al., 2020)

- 34229 genes (**samples**).
- Provided data: continuous and Gaussian.

Epigenomic (DNA methylation): WGBS data (Sow et al., 2023)

- Three methylation contexts measured on gene-body and gene-promoter.
- $\rightarrow$ 6 data-sets each containing about 40000 genes (**samples**).
- Provided data: normalized counts per gene.
- To gaussian: $log_2(x+1)$ with $x \in \mathbb{R}$.

Multi-omic data

Omics measured for 20 genotypes (**variables**) for genotype-genotype networks

Transcriptomic (Gene expression): RNA-Seq data (Chateigner et al., 2020)

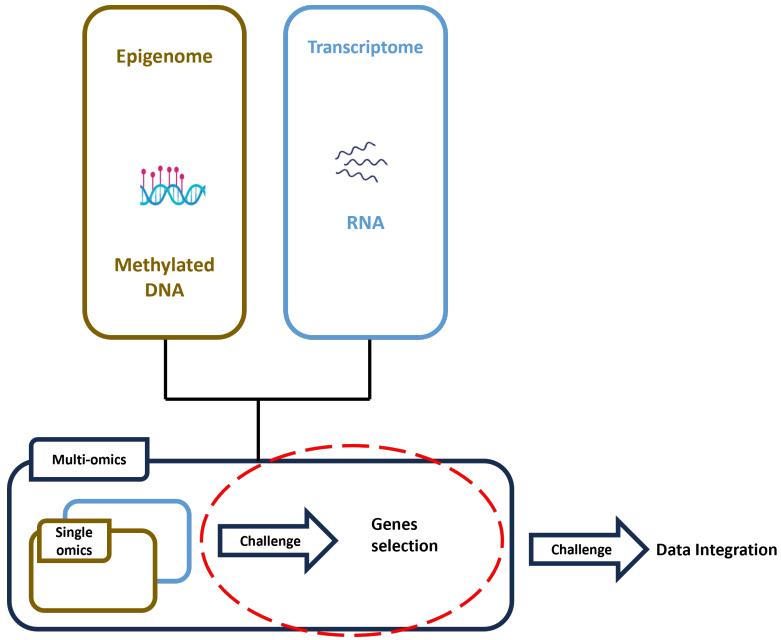
- 34229 genes (**samples**).
- Provided data: continuous and Gaussian.

Epigenomic (DNA methylation): WGBS data (Sow et al., 2023)

- Three methylation contexts measured on gene-body and gene-promoter.
- $\rightarrow$ 6 data-sets each containing about 40000 genes (**samples**).
- Provided data: normalized counts per gene.
- To gaussian: $log_2(x+1)$ with $x \in \mathbb{R}$.

For each omic, average measures per natural populations.

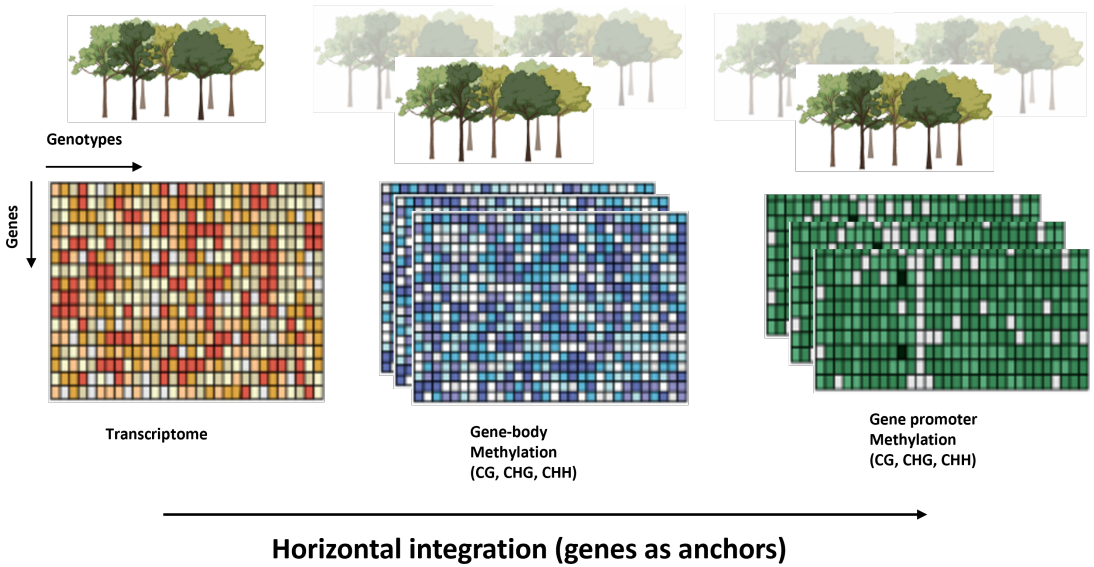
Complex omics data integration



Genes (**samples**) selection

- For scalability issue of MGLASSO.
- For each omic data, apply sparse PCA and select 15 top-genes for the 3 PCs using `mixomics`.
- Merge genes lists for all the omics and remove duplicates: 151 genes.

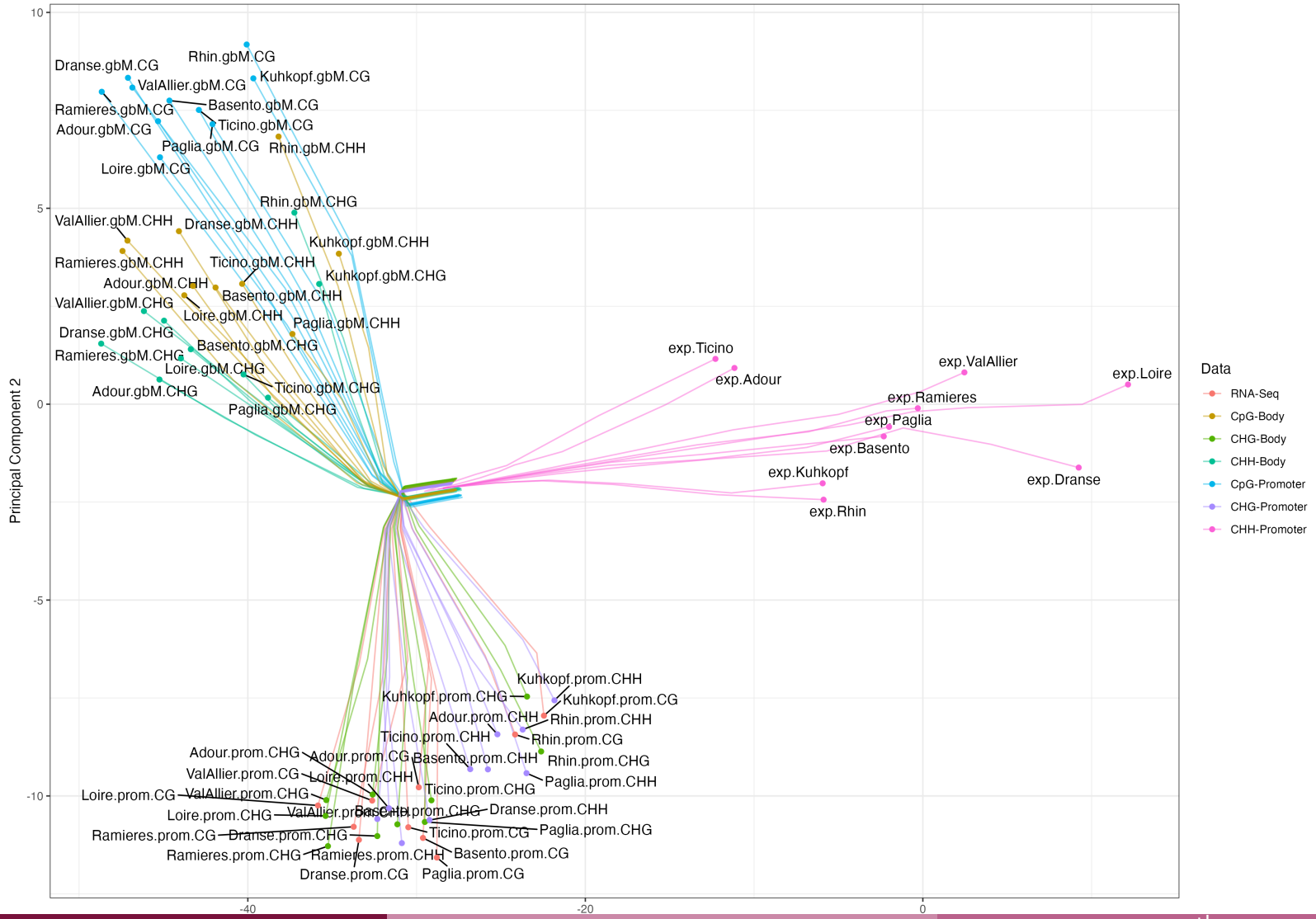
Integration across genes



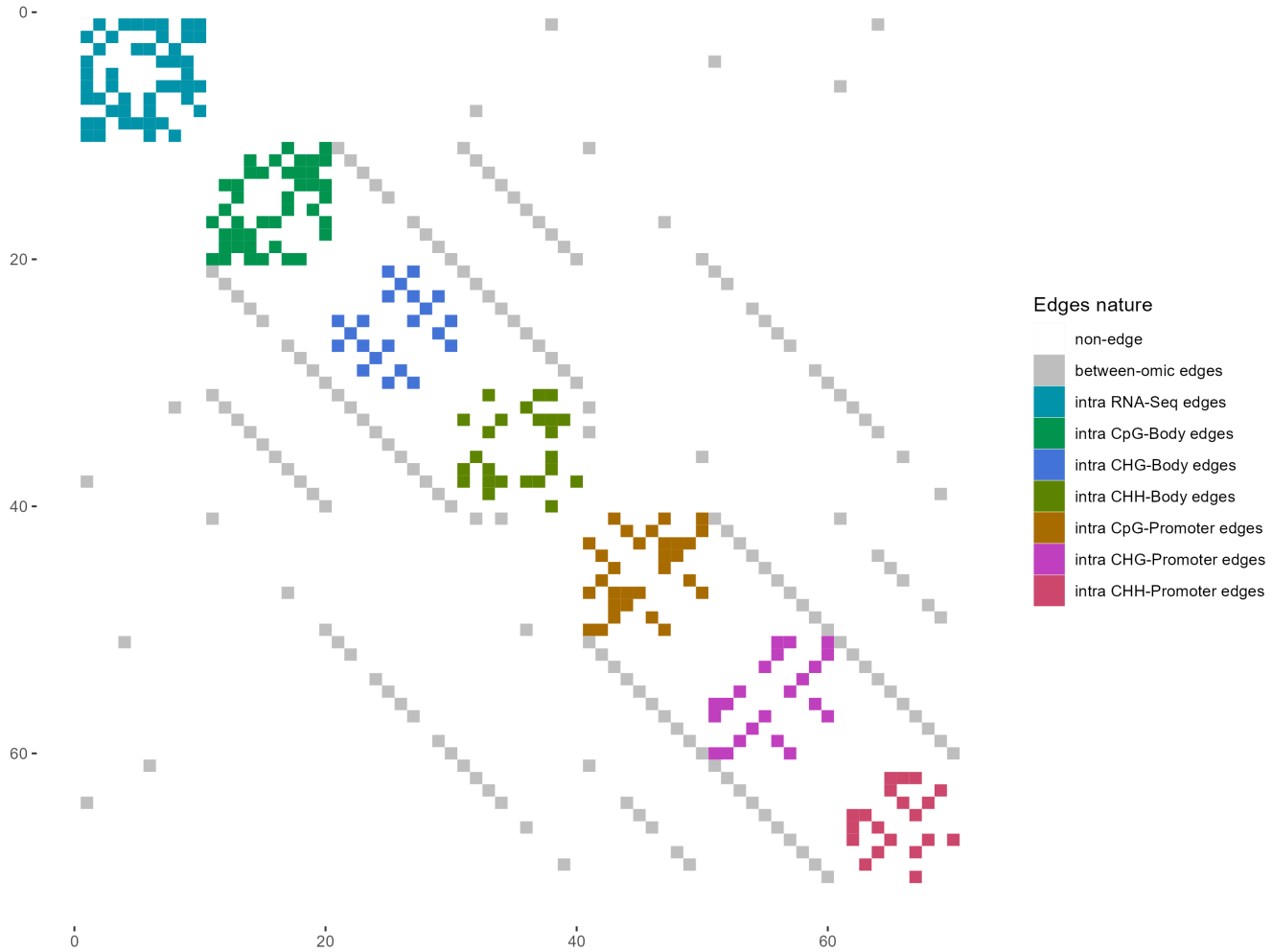
- **X** : 151 (genes) observations of 70 Gaussian methylation and transcriptomic profiles (variables)

Clustering path

MGLASSO identifies three coherent groups of omics profiles.

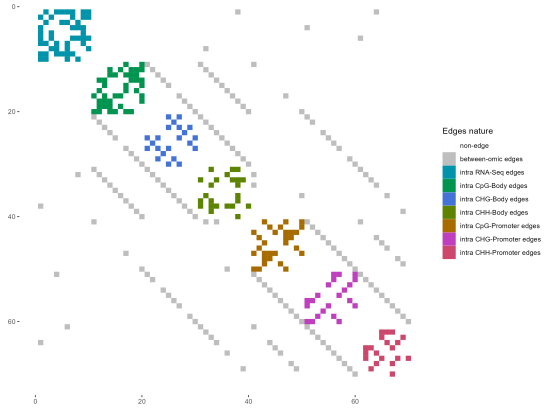


Multiscale Networks: non-compressed view (1)

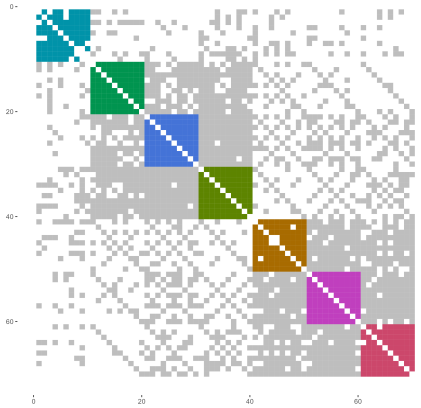


(a) $\lambda_2 = 0$

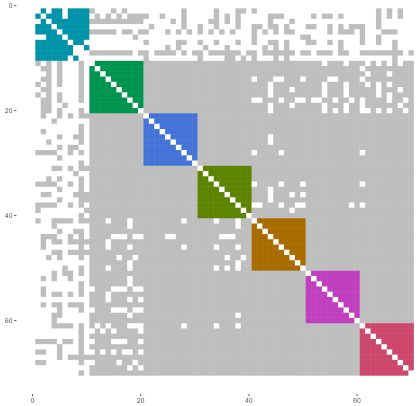
Multiscale Networks: non-compressed view (2)



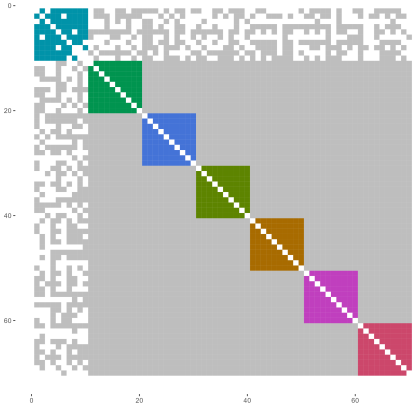
(a) $\lambda_2 = 0$



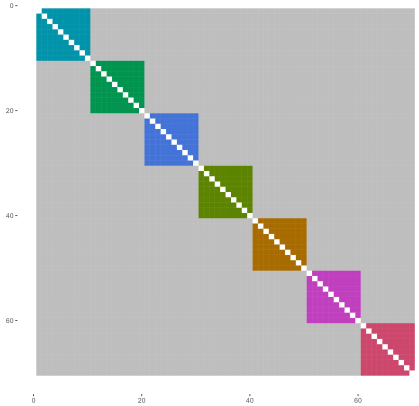
(b) $\lambda_2 = 1.63$



(c) $\lambda_2 = 3.26$



(d) $\lambda_2 = 4.89$



(e) $\lambda_2 = 30.94$

Conclusion and Perspectives

Conclusion

Mathematics

Applications: Life sciences

- Epigenetics can differentiate natural populations of poplars.

Through

- Differential analysis of methylation data.
- Gene set enrichment analysis of selected genes.

Conclusion

Mathematics

A probabilistic model for:

- Inferring multiscale networks for continuous data.
- Estimating simultaneously clustering partition.

Based on:

- Gaussian graphical modeling via pseudo log-likelihood maximization.
- Convex clustering theory.

Applications: Life sciences

- Epigenetics can differentiate natural populations of poplars.

Through

- Differential analysis of methylation data.
- Gene set enrichment analysis of selected genes.

Conclusion

Mathematics

A probabilistic model for:

- Inferring multiscale networks for continuous data.
- Estimating simultaneously clustering partition.

Based on:

- Gaussian graphical modeling via pseudo log-likelihood maximization.
- Convex clustering theory.

Applications: Life sciences

- Epigenetics can differentiate natural populations of poplars.

Through

- Differential analysis of methylation data.
- Gene set enrichment analysis of selected genes.

Extensions

Convex clustering

- Conditions for which one can recover a tree structure.
- Bounds on regularization parameters.

Network inference

- Penalized maximum likelihood estimator.
- Fitting mixed graphical models for heterogeneous data.

Optimization

- Scale to high dimensional data.

Contributions

Articles

- E. Sanou, C. Ambroise, G. Robin *"Inference of Multiscale Gaussian Graphical Model."* Computo (2023)
- M. Sow et al. *"Epigenetic Variation in Tree Evolution: a case study in black poplar (Populus nigra)."* bioRxiv 2023.07.16.549253. Submitted to New Phytologist (2023).

R packages

- E. Sanou *"mglasso: Multiscale Graphical Lasso."*: CRAN, 2022 `https://CRAN.R-project.org/package=mglasso`.

Thank you for your attention.

References I

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160.

Ambroise, C., Chiquet, J., and Matias, C. (2009). Inferring sparse gaussian graphical models with latent structure. *Electronic Journal of Statistics*, 3(0):205–238.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122.

Bühlmann, P., Rütimann, P., Van De Geer, S., and Zhang, C.-H. (2013). Correlated variables in regression: clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143(11):1835–1858.

Chateigner, A., Lesage-Descauses, M.-C., Rogier, O., Jorge, V., Leplé, J.-C., Brunaud, V., Roux, C. P.-L., Soubigou-Taconnat, L., Martin-Magniette, M.-L., Sanchez, L., et al. (2020). Gene expression predictions and networks in natural populations supports the omnigenic theory. *BMC genomics*, 21(1):1–16.

Cheng, L., Shan, L., and Kim, I. (2017). Multilevel Gaussian graphical model for multilevel networks. *Journal of Statistical Planning and Inference*, 190:1–14.

Chi, E. C. and Steinerberger, S. (2019). Recovering trees with convex clustering. *SIAM Journal on Mathematics of Data Science*, 1(3):383–407.

Chiquet, J., Gutierrez, P., and Rigaill, G. (2017). Fast tree inference with weighted fusion penalties. *Journal of Computational and Graphical Statistics*, 26(1):205–216.

Hadj-Selem, F., Lofstedt, T., Dohmatob, E., Frouin, V., Dubois, M., Guillemot, V., and Duchesnay, E. (2018). Continuation of Nesterov's Smoothing for Regression with Structured Sparsity in High-Dimensional Neuroimaging. *IEEE Transactions on Medical Imaging*, 2018.

Hocking, T., Vert, J.-P., Bach, F., and Joulin, A. (2011). Clusterpath: an algorithm for clustering using convex fusion penalties. In *ICML*.

Lin, M., Sun, D., Toh, K.-C., and Wang, C. (2020). Estimation of sparse gaussian graphical models with hidden clustering structure. *arXiv preprint arXiv:2004.08115*.

References II

Lindsten, F., Ohlsson, H., and Ljung, L. (2011). Clustering using sum-of-norms regularization: With application to particle filter output computation. In *2011 IEEE Statistical Signal Processing Workshop (SSP)*, pages 201–204.

McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., Aksenov, A. A., Behsaz, B., Brennan, C., Chen, Y., et al. (2018). American gut: an open platform for citizen science microbiome research. *Msystems*, 3(3):e00031–18.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462.

Pelckmans, K., De Brabanter, J., Suykens, J. A., and De Moor, B. (2005). Convex clustering shrinkage. In *PASCAL workshop on statistics and optimization of clustering workshop*.

Sanou, E., Ambroise, C., and Robin, G. (2023). Inference of Multiscale Gaussian Graphical Models. *Computo*.

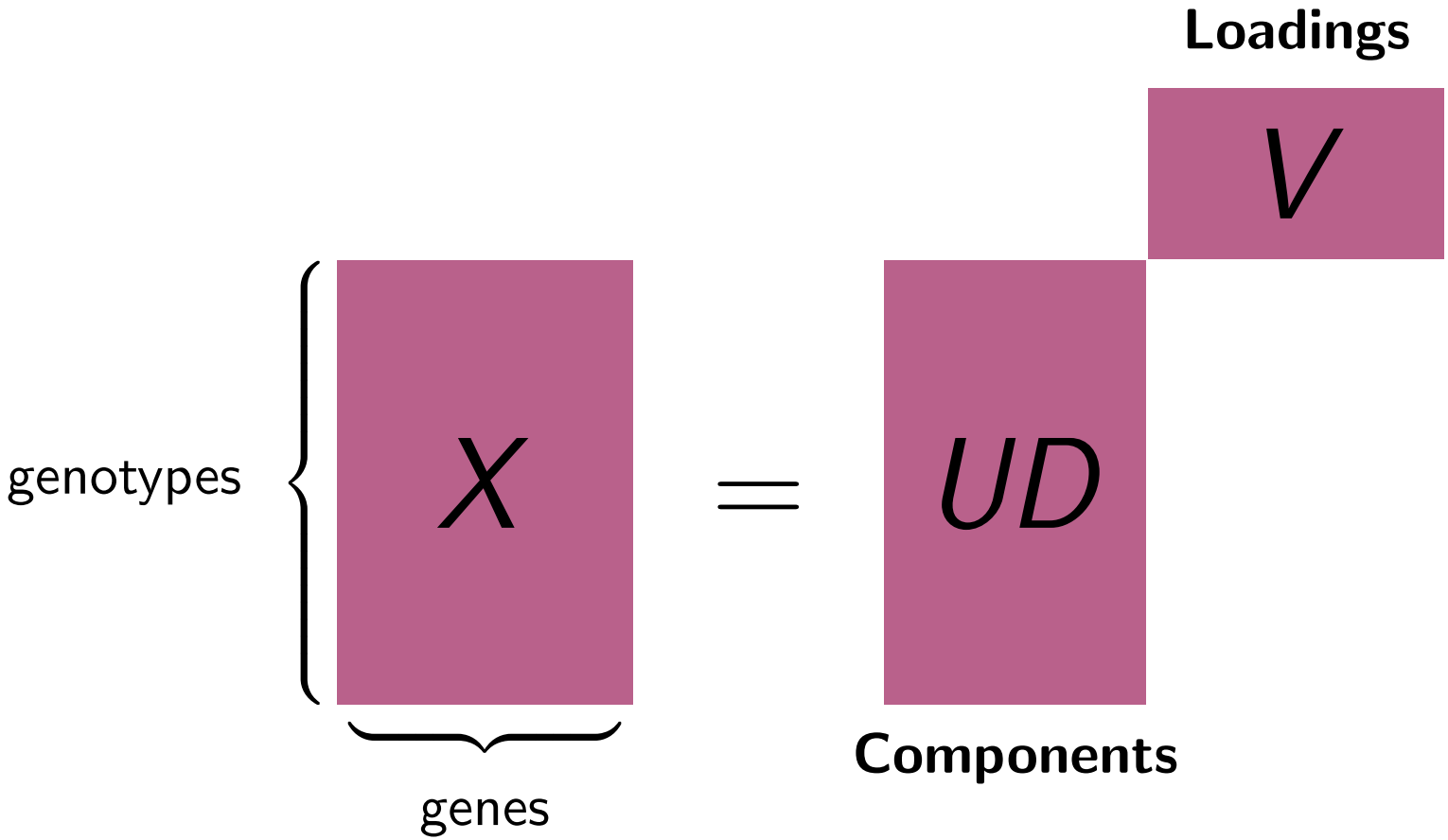
Shor, N. Z. (2012). *Minimization methods for non-differentiable functions*, volume 3. Springer Science & Business Media.

Sow, M. D., Rogier, O., Lesur, I., Daviaud, C., Mardoc, E., Sanou, E., Duvaux, L., Civan, P., Delaunay, A., Lesage-Descauses, M.-C., Benoit, V., Le-Jan, I., Buret, C., Besse, C., Durufle, H., Fichot, R., Le-Provost, G., Guichoux, E., Boury, C., Garnier, A., Senhaji-Rachik, A., Jorge, V., Ambroise, C., Tost, J., Plomion, C., Segura, V., Maury, S., and Salse, J. (2023). Epigenetic variation in tree evolution: a case study in black poplar (populus nigra). *bioRxiv*.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B*, pages 91–108.

Yao, T. and Allen, G. I. (2019). Clustered gaussian graphical model via symmetric convex clustering. In *2019 IEEE Data Science Workshop (DSW)*, pages 76–82.

sparse PCA for gene selection (1)



PCA: Dimension reduction

- Via SVD of the data matrix $\mathbf{X} = \mathbf{UDV}^\top$
- PCs are linear combinations of all the genes

sparse PCA for gene selection (2)

PCA as a low rank matrix approximation problem

- Sum of rank 1 matrices

$$\begin{aligned}\mathbf{X} &= \mathbf{U}(\mathbf{D}_1 + \mathbf{D}_2 + \ldots + \mathbf{D}_p)\mathbf{V}^\top \\ &= d_1\mathbf{u}_1\mathbf{v}_1^\top + \cdots + d_p\mathbf{u}_p\mathbf{v}_p^\top\end{aligned}$$

- PCA seeks the best rank 1 approximation of $\mathbf{X}$:

$$\min \mathbf{a}, \mathbf{v} \left\| \mathbf{X} - \mathbf{a}\mathbf{v}^\top \right\|_F^2$$

for the first pair of singular vectors $\mathbf{a}_1 = d_1\mathbf{u}_1$ and $\mathbf{v}_1$.

- Then find $(\mathbf{a}_2, \mathbf{v}_2)$ as the best rank 1 approximation of $\mathbf{X} - \mathbf{a}_1\mathbf{v}_1^\top$

sparse PCA for gene selection (2)

PCA as a low rank matrix approximation problem

sparse PCA

$$\min \mathbf{a}, \mathbf{v} \left\| \mathbf{X} - \mathbf{a}\mathbf{v}^\top \right\|_F^2 + \lambda \|\mathbf{v}\|_1$$

subject to $\|\mathbf{a}\|_2 = 1$.

Procedure

- Apply sparse PCA and select 15 top-genes for the 3 PCs using `mixomics`
- Merge genes lists for all the omics and remove duplicates: 151 genes

Application on microbial abundance data

Microbial abundance data

Microbial abundance data collected as part of the American Gut project (McDonald et al., 2018).

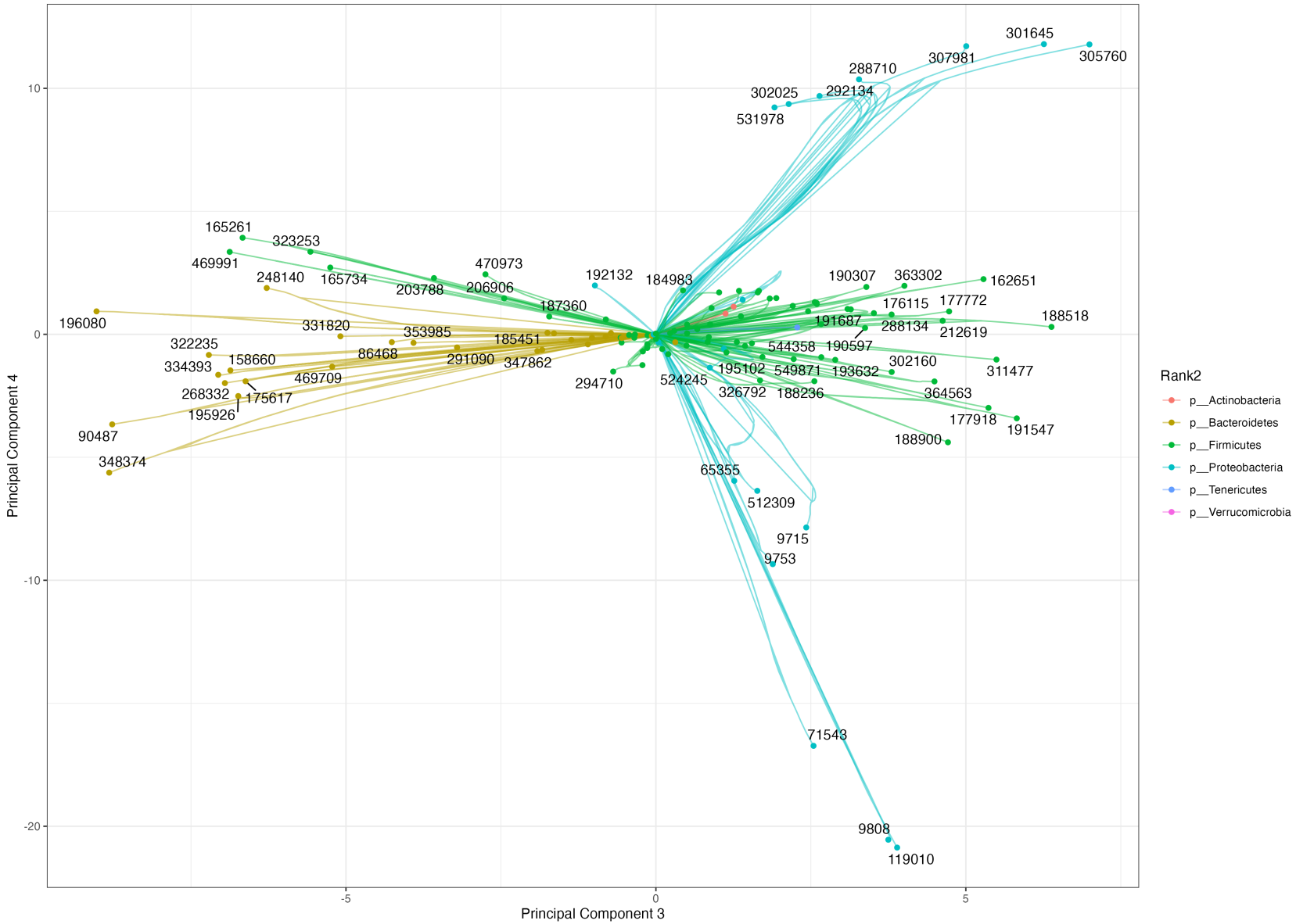
- $n = 289$ samples via an abundance table for $p = 127$ types of microbes (operational taxonomic units).

OTU 1	OTU 2	...	OTU 127
0	7	...	0
0	43	...	4
8	0	...	2
0	0	...	0
...	...	...	...
0	0	...	346

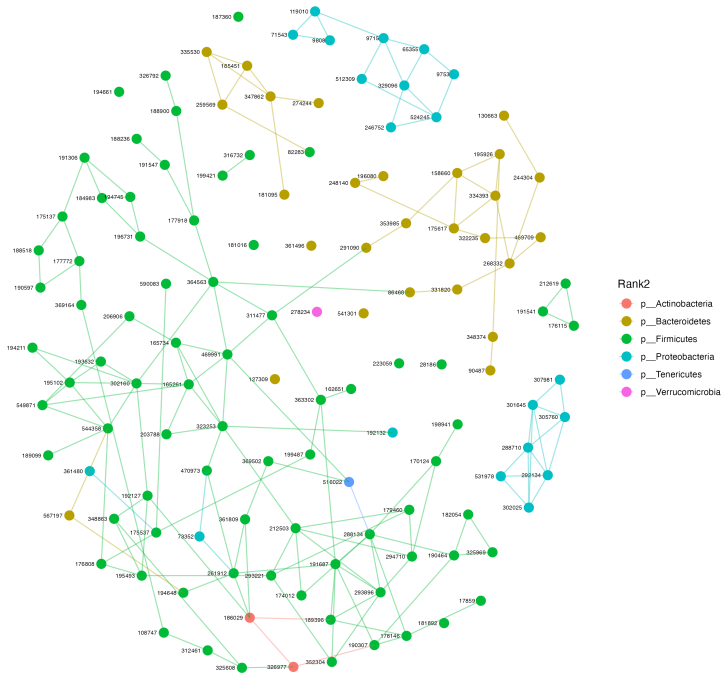
Table: Matrix of OTUs

- Data transformed with the centered log-ratio (clr, Aitchison (1982)).

OTUs clustering path

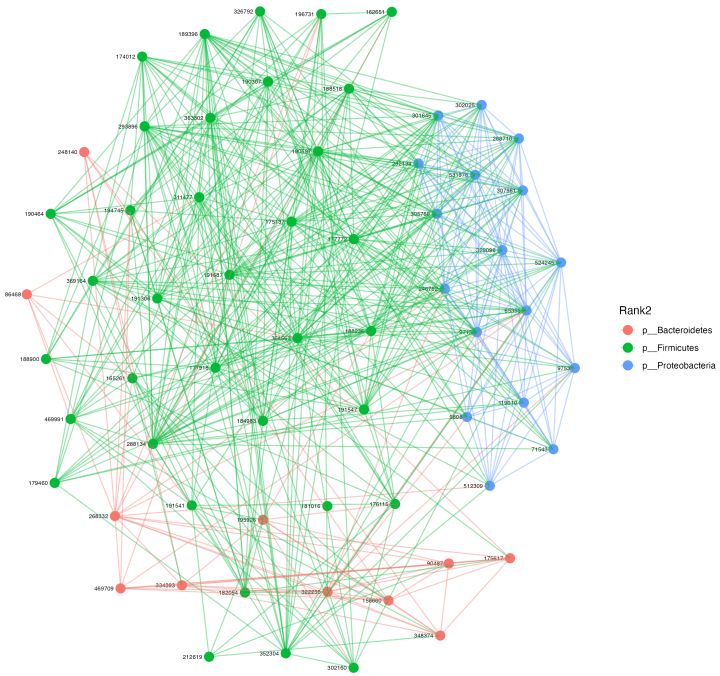


OTUs networks

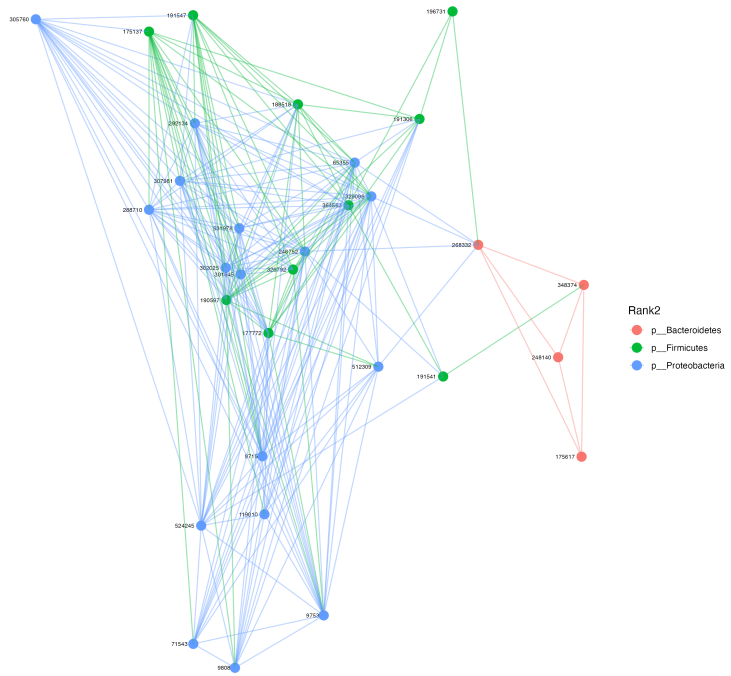


(a) $\lambda_2 = 0$

OTUs networks

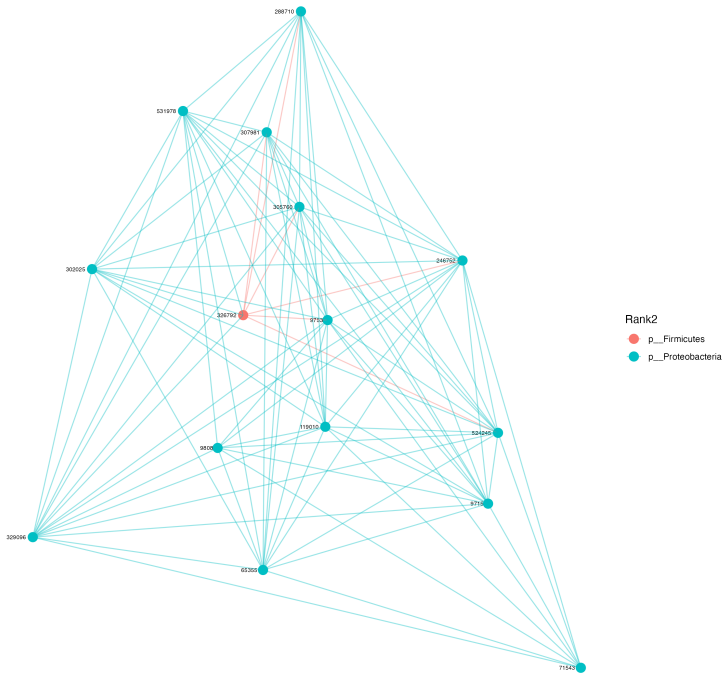


(a) 63 clusters

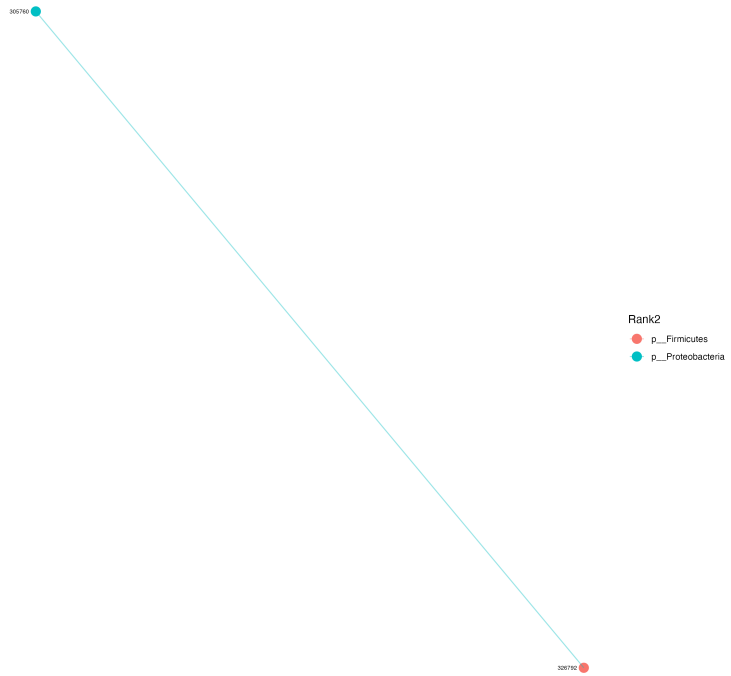


(b) 31 clusters

OTUs networks



(a) 15 clusters



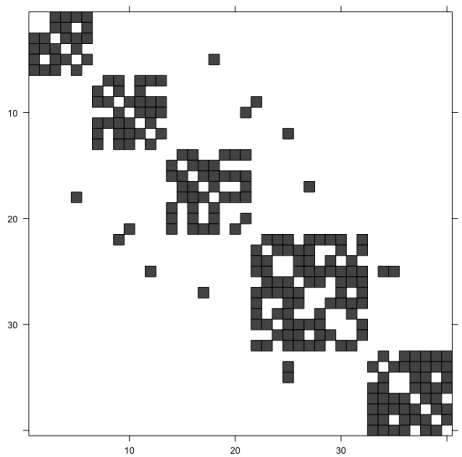
(b) 2 clusters

Simulations

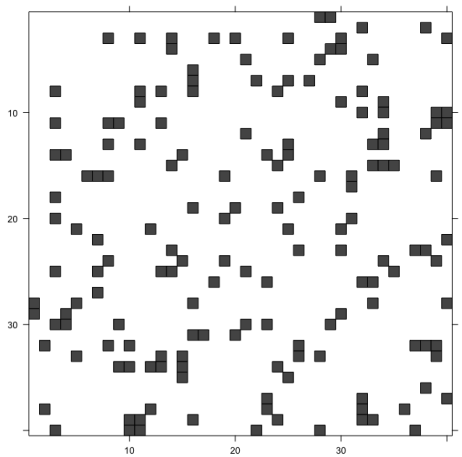
Performances evaluated in term of:

- Support recovery:
  - Methods: GLASSO
  - Simulation model: Stochastic block, Erdos-Renyi and Scale free models
  - Criterion: ROC curve
- Clustering:
  - Methods: HAC, $k-$means, Convex clustering, spectral clustering
  - Simulation model: Stochastic block and hierarchically structured models
  - Criterion: Adjusted Rand Index

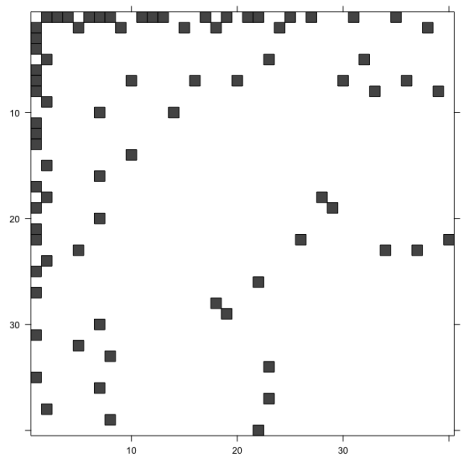
Graph models



(a) Stochastic block model



(b) Erdos-Renyi model



(c) Scale free model

Figure: Different graph models

Stochastic block model: settings

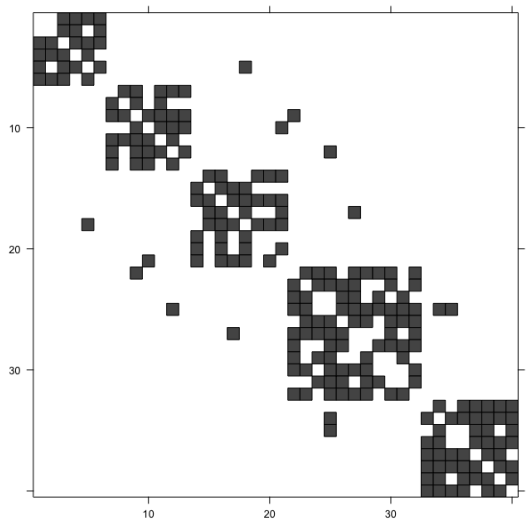
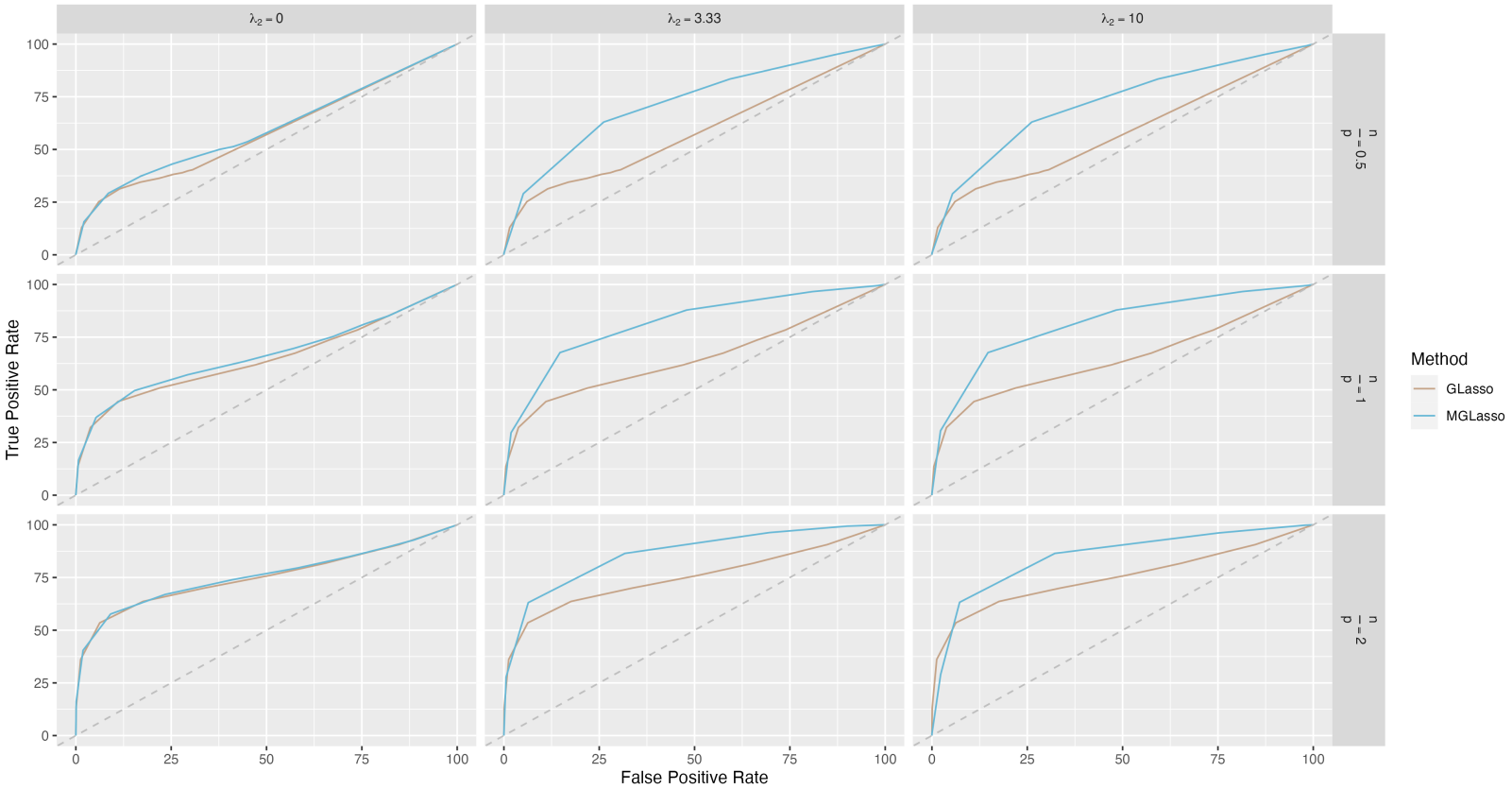


Figure: Stochastic block model

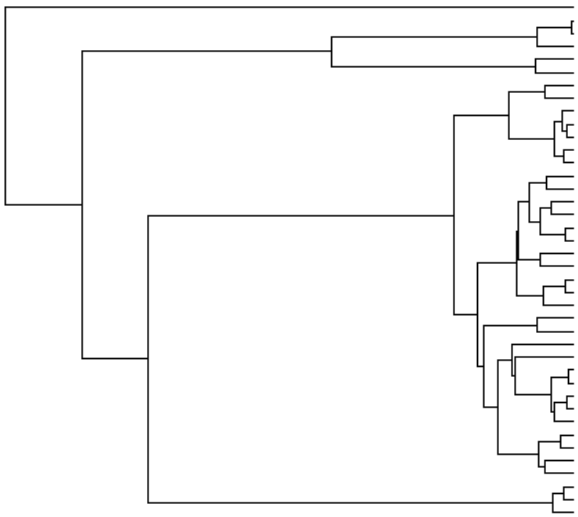
Settings

- $p = 40$ variables, $\frac{n}{p} \in \{0.5, 1, 2\}$
- 50 simulations per setting

Stochastic block model: results



Hierarchically structured model



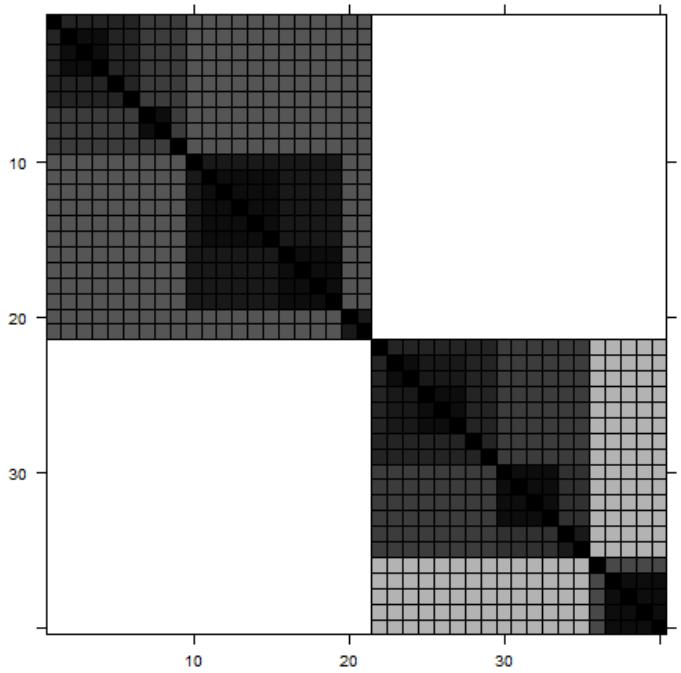
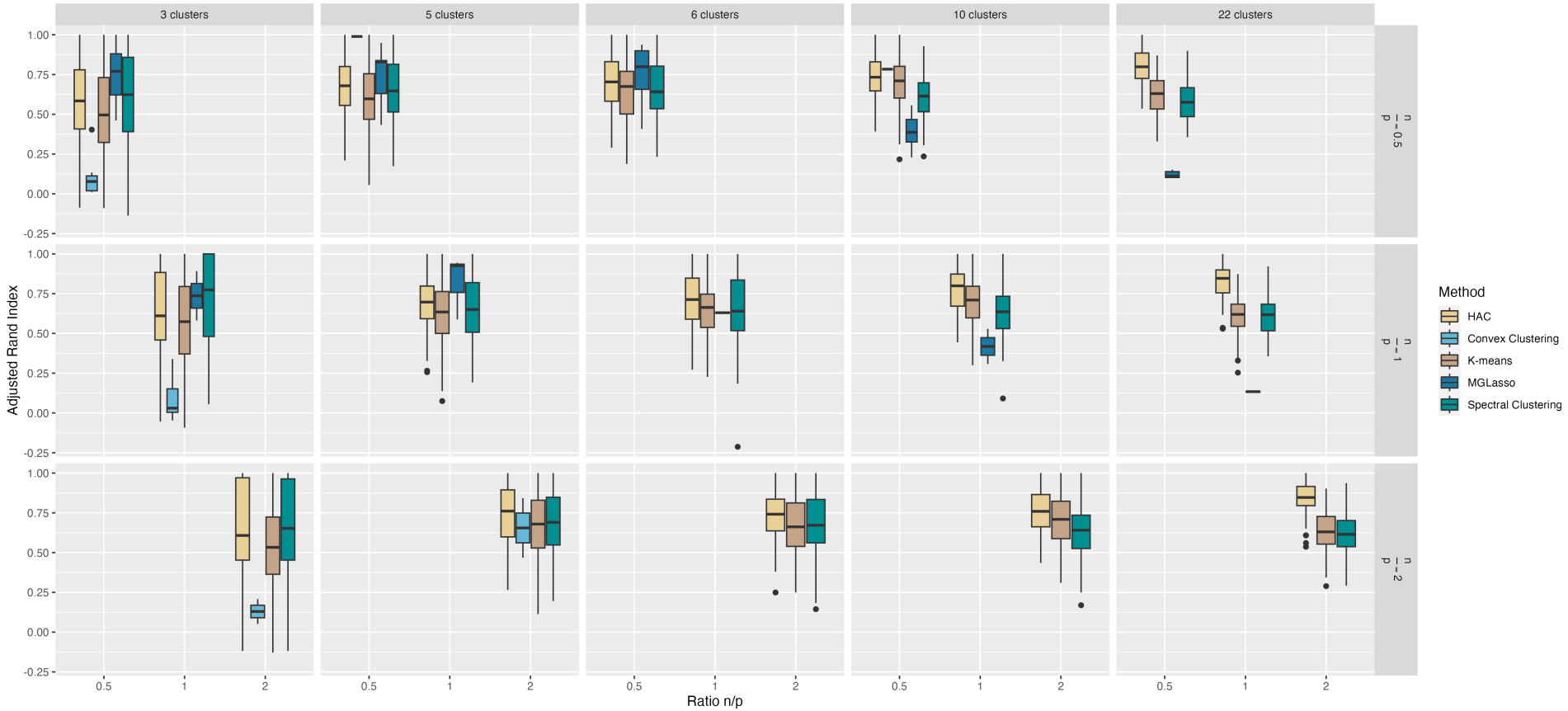


Figure: Tree, Covariance matrix of the phylogeny based hierarchical model

Hierarchically structured model: results



Graphical models + Convex clustering: Relevant Works

- Clustered Gaussian Graphical Model Via Symmetric Convex Clustering (Yao and Allen, 2019):

$$\min_{\Theta,\{\Psi_l\}} -\log\det\Omega + \operatorname{tr}\hat{\Sigma}\Omega + \lambda\sum_{l\in\mathcal{M}} w_l \left\|\Psi_{l,1} - \Psi_{l,2}\right\|_2 \tag{1}$$

subject to $(Q_l\Theta R_l) - \Psi_l = 0, \forall l \in \mathcal{M}$.

- Estimation of Sparse Gaussian Graphical Models with Hidden Clustering Structure (Lin et al., 2020):

$$\max_{\mathbf{\Omega}\in\mathbb{S}^p_{\succ 0}} \left\{ \log\det(\mathbf{\Omega}) - \operatorname{tr}(\mathbf{S\Omega}) - \lambda_1\sum_{i<j}|\mathbf{\Omega}_{ij}| - \lambda_2\sum_{i<j}\sum_{s<t}|\mathbf{\Omega}_{ij} - \mathbf{\Omega}_{st}| \right\} \tag{2}$$

Convex Clustering with Weights

$$\frac{1}{2}\sum_{i=1}^{n}\|x_i-\alpha_i\|_2^2+\lambda\sum_{i<j}w_{ij}\|\alpha_i-\alpha_j\|_q \tag{3}$$

Distance Decreasing Weights:

- Empirical approach Hocking et al. (2011); Chi and Steinerberger (2019): $w_{ij}=\exp\left(-\gamma\|x_i-x_j\|_2^2\right)$ (Gaussian kernel)
- Theoretical approach Chiquet et al. (2017): $w_{ij}=f\left(\|x_i-x_j\|_q\right)$ (ℓ_1 case, and multimensional clusterpath problem).

$$\sum_{i=1}^{p}\left[\frac{1}{2}\sum_{i=1}^{n}(x_{ik}-\alpha_{ik})^2+\lambda\sum_{i<j}w_{ij}\left|\alpha_{ik}-\alpha_{jk}\right|\right] \tag{4}$$

Conditional Gaussian Distribution

Given $\mathbf{X}^{\setminus j} = \boldsymbol{Z}$, the conditional distribution of $\mathbf{X}^j = Y$ is Gaussian.
Partitioning $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZY} & \boldsymbol{\Sigma}_{ZZ} \end{pmatrix}$$

The conditional distribution:

$$Y|\boldsymbol{Z} = z \sim \mathcal{N}\left((z-\mu_Z)^T \boldsymbol{\Sigma}_{ZZ}^{-1}\Sigma_{ZY}, \Sigma_{YY} - \Sigma_{ZY}^T \boldsymbol{\Sigma}_{ZZ}^{-1}\Sigma_{ZY}\right)$$

Using Schur complement and partitioning $\boldsymbol{\Omega}$:

$$Y|\boldsymbol{Z} = z \sim \mathcal{N}\left((z-\mu_Z)^T(-\boldsymbol{\Omega}_{YZ}/\Omega_{YY}), \Omega_{YY}^{-1}\right)$$

Log-likelihood of univariate-conditional normal distribution:

$$\log p(Y|\boldsymbol{Z}) = \frac{1}{2}\log(\Omega_{YY}) - \frac{1}{2}\Omega_{YY}\left(y + \frac{\boldsymbol{\Omega}_{YZ}}{\Omega_{YY}}z\right)^2 + \text{const}$$