Integrating hierarchical information in the analysis of microbiome data

C. Ambroise[1], A. Bichat[1,2,3], *M. Mariadassou*[2]

[1]LaMME, UEVE, Université Paris-Saclay, Evry, France
[2]MAIAGE, INRAE, Université Paris-Saclay, Jouy-en-Josas, France
[3]Enterome, Paris, France

DATA IA Workshop
November 5th, 2020

Outline

Microbiota

Ecological community of microorganisms that reside in an environmental niche.

Published associations with:

- Inflammatory bowel diseases
- Liver disease
- Vaccine efficiency
- Anxiety
- Muscular strength
- etc

Microbiota

Ecological community of microorganisms that reside in an environmental niche.

Published associations with:

- Inflammatory bowel diseases
- Liver disease
- Vaccine efficiency
- Anxiety
- Muscular strength
- etc

Strong interest in findings (groups of) microbes associated to a given condition.

Metabarcoding data (Ravel et al., 2011)

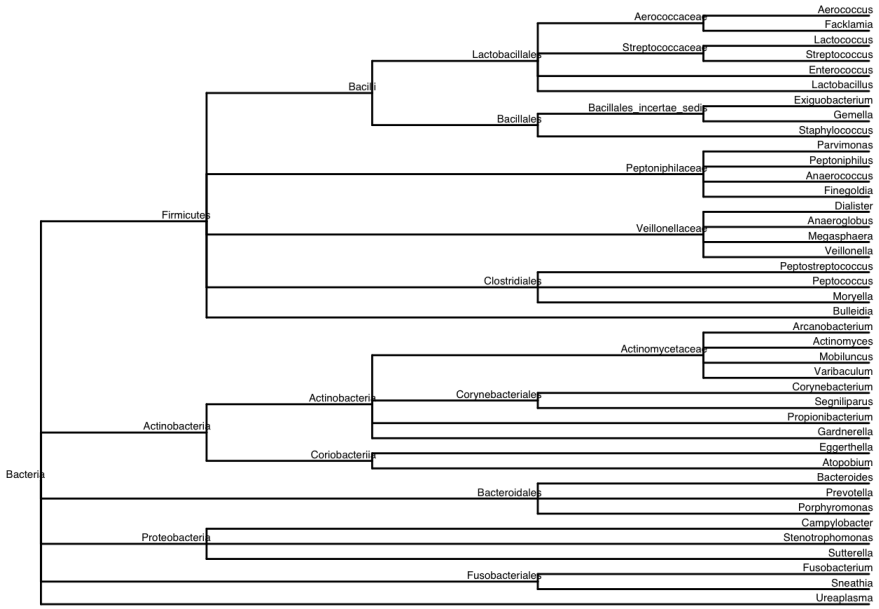
A species × sample count table

	Taxa	A1	A2	A3	B1	B2	B3
1	Lactobacillus	2318	1388	1361	2256	88	1770
2	Prevotella	0	1	1	0	525	7
3	Megasphaera	0	1	0	0	402	0
4	Sneathia	0	0	0	0	302	0
5	Atopobium	0	1	0	0	84	0
6	Streptococcus	0	0	3	0	0	0
7	Dialister	0	1	0	0	152	4
8	Anaerococcus	0	1	3	2	0	9
9	Peptoniphilus	0	1	0	0	7	2
10	Eggerthella	0	0	0	0	2	0

Metabarcoding data (Ravel et al., 2011)

Taxonomic / phylogenetic tree

	Phylum	Class	Order	Family	Genus
1	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Actinobaculum
2	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Actinomyces
3	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Arcanobacterium
4	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Mobiluncus
5	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Varibaculum
6	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium
7	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Gardnerella

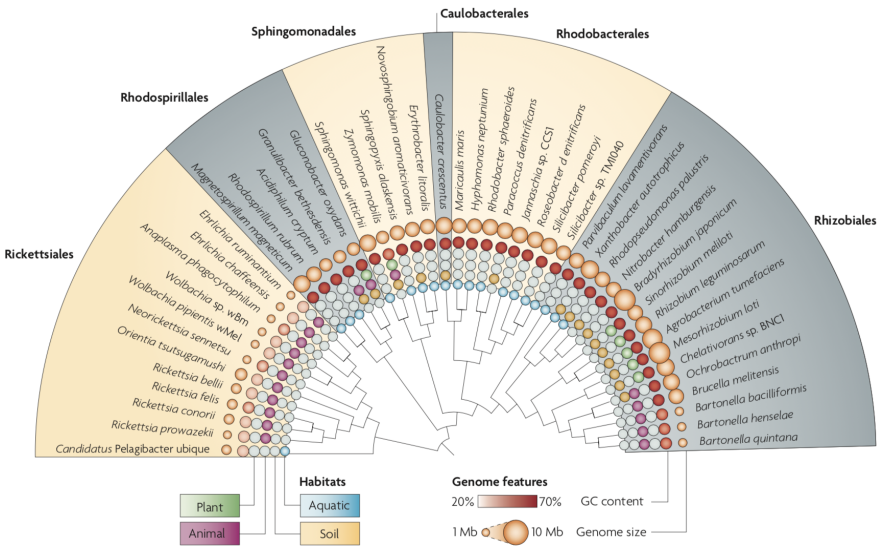


Differential abundance analysis

- For each taxa i (in $\{1, \ldots, n\}$), test
  - H_{0i}: Abundances are equal in groups A and B
  - H_{1i}: Abundances are not equal in groups A and B
- Hundred of univariate tests and p-values
- Need for a multiple testing correction procedure

Differential abundance analysis

- For each taxa i (in $\{1, \ldots, n\}$), test
  - H_{0i}: Abundances are equal in groups A and B
  - H_{1i}: Abundances are not equal in groups A and B
- Hundred of univariate tests and p-values
- Need for a multiple testing correction procedure



- Taxa / group associations may show a **phylogenetic signal**
- Similar taxa $\Rightarrow$ similar levels of association
- Can we leverage the tree when correcting the tests?

Outline

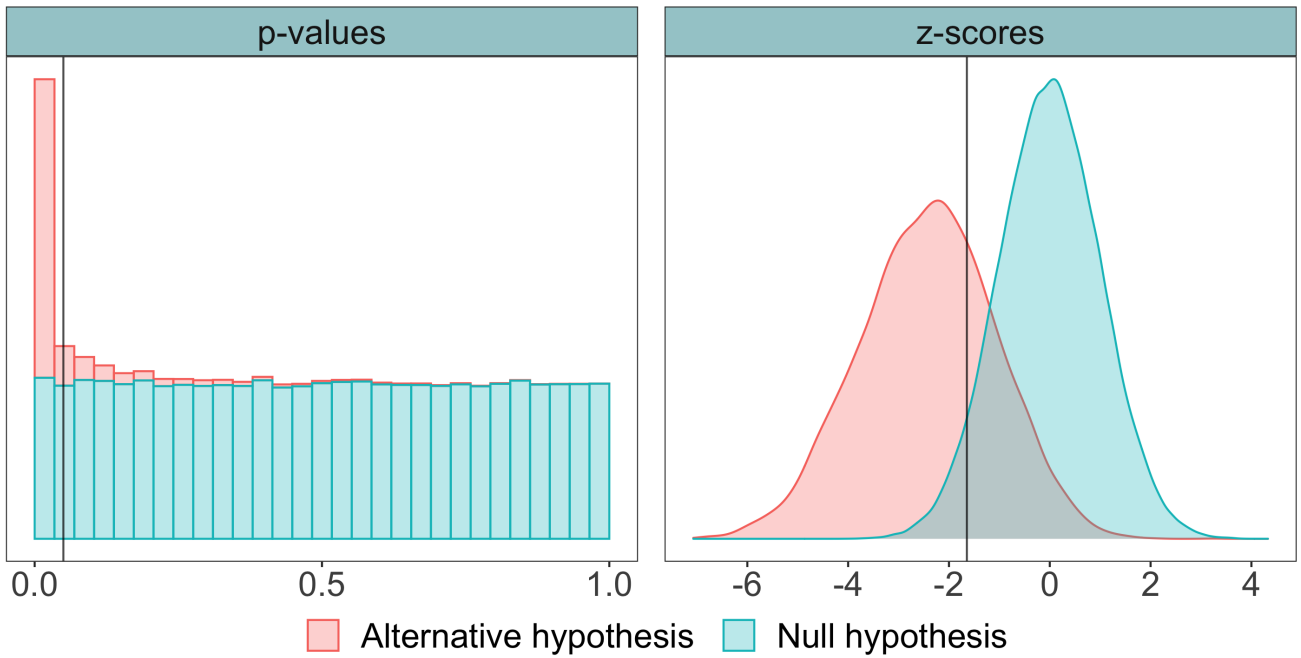
Mathematical Model

Standard assumptions on *p*-values

- Under $H_{0i}, p_i \sim \mathcal{U}(0,1)$
- Under $H_{1i}, p_i \preccurlyeq \mathcal{U}(0,1)$

Standard assumptions on *z*-scores

- Under H_{0i}, $z_i = \Phi^{-1}(p_i) \sim \mathcal{N}(0,1)$
- Under H_{1i}, $z_i = \Phi^{-1}(p_i) \sim \mathcal{N}(m_i, 1)$ with $m_i < 0$



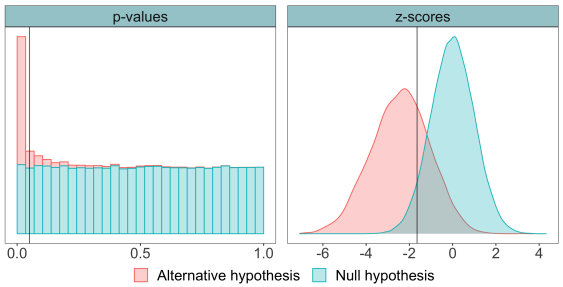
Mathematical Model

Standard assumptions on p-values

- Under $H_{0i}, p_i \sim \mathcal{U}(0,1)$
- Under $H_{1i}, p_i \preccurlyeq \mathcal{U}(0,1)$

Standard assumptions on z-scores

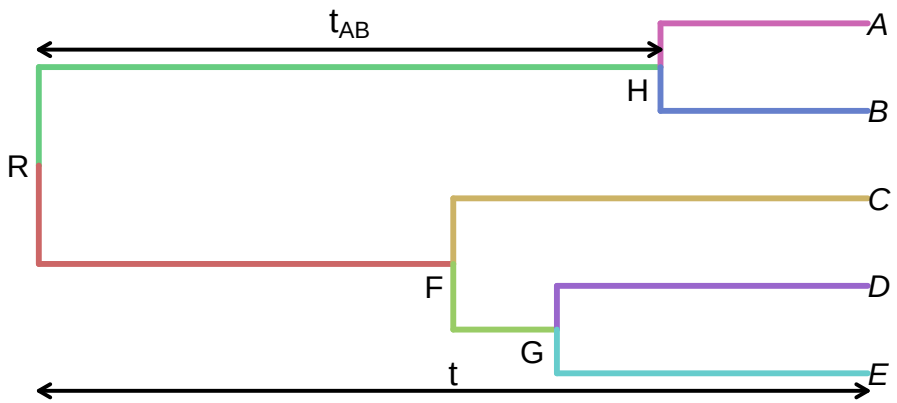
- Under H_{0i}, $z_i = \Phi^{-1}(p_i) \sim \mathcal{N}(0,1)$
- Under H_{1i}, $z_i = \Phi^{-1}(p_i) \sim \mathcal{N}(m_i, 1)$ with $m_i < 0$

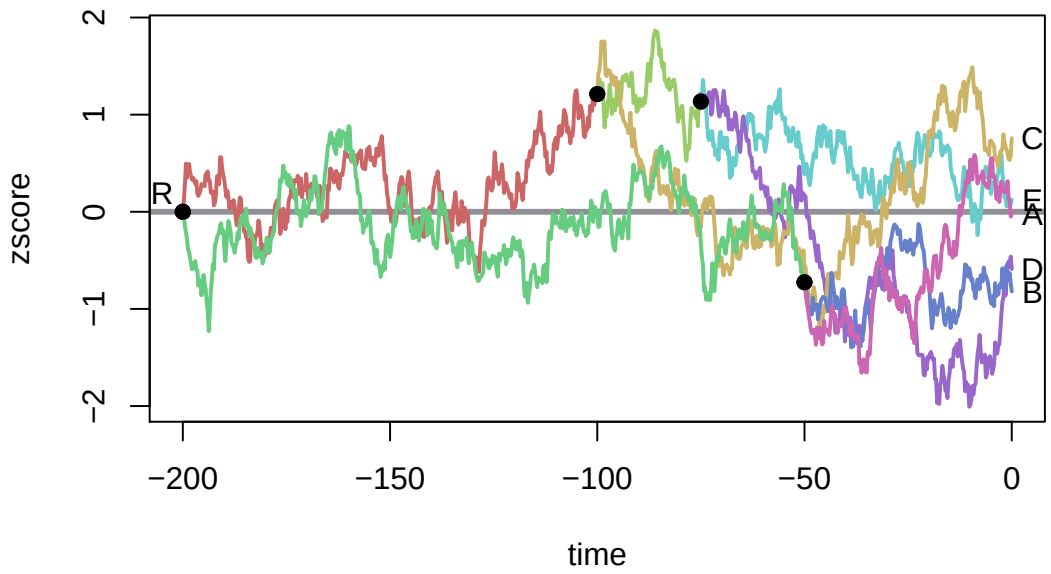


Hierarchical assumptions on z-scores

- The (z_i) are correlated according to a tree.

Stochastic Process on a Tree (Felsenstein, 1985)





The tree is known.

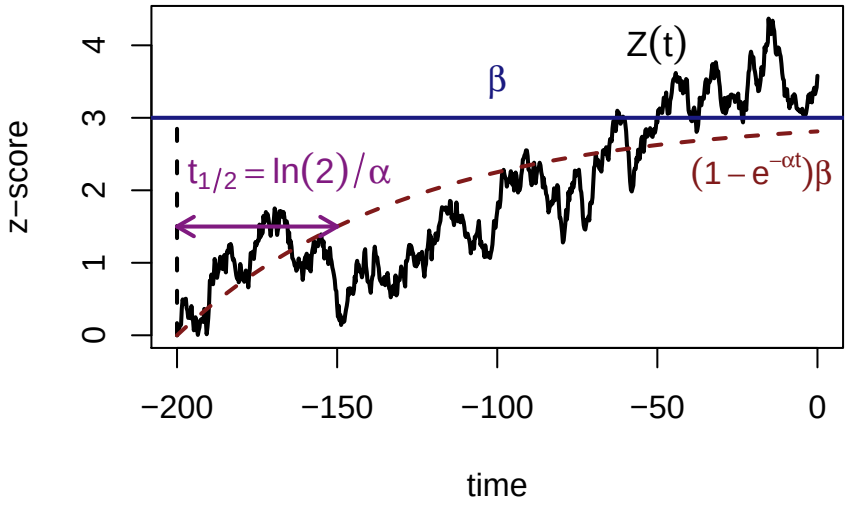
Only *tip* values are observed

Ornstein Uhlenbeck:

$$\mathbb{V}\mathrm{ar}\left[A \mid R\right] = \frac{\sigma^2}{2\alpha}(1-e^{-2\alpha t})$$

$$\mathbb{C}\mathrm{or}(A, B|R) = e^{-2\alpha(t-t_{AB})}$$

OU Modeling (Hansen, 1997)



$$dZ(t) = \alpha[\beta(t) - Z(t)]dt + \sigma dB(t)$$

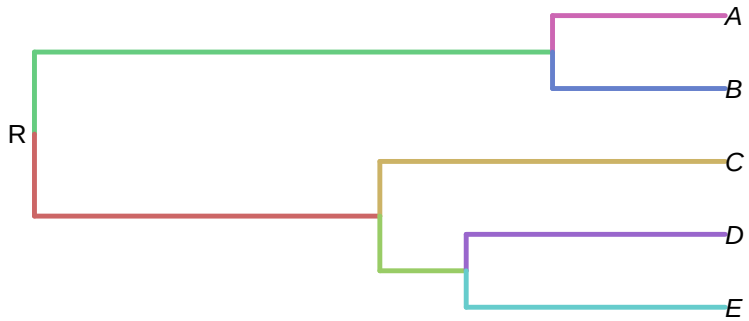
Deterministic part:

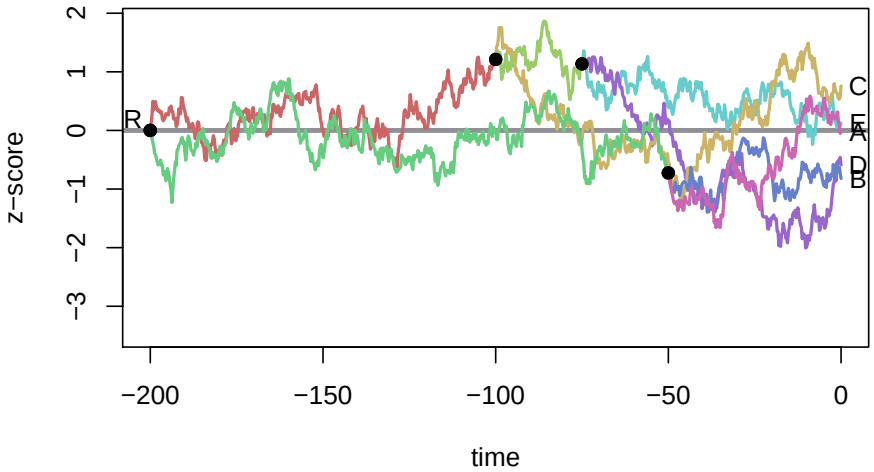
- $\beta(t)$: Effect size
- $\ln(2)/\alpha$: phylogenetic half live.

Stochastic part:

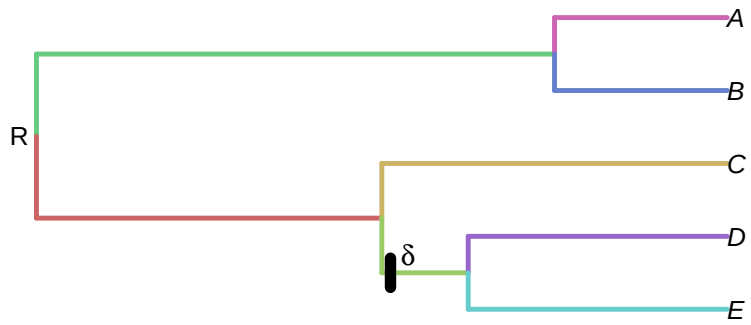
- $Z(t)$: z-scores $\simeq$ estimated effect size.
- $\sigma dB(t)$ Brownian fluctuations.

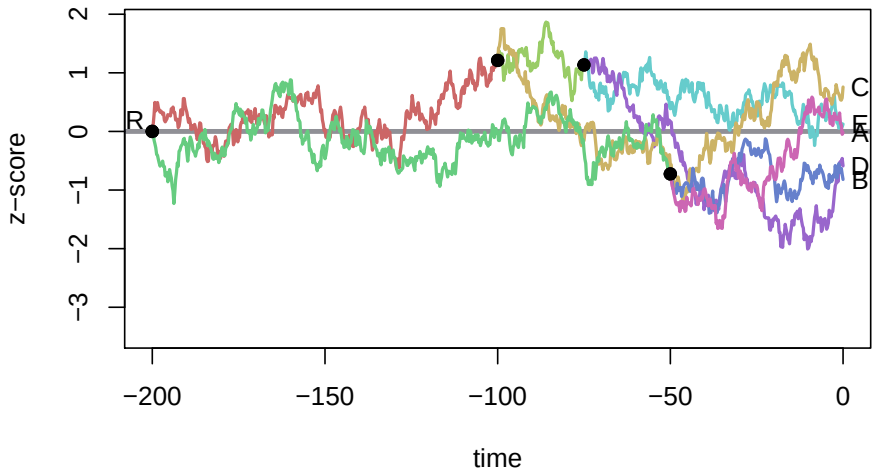
Shifts

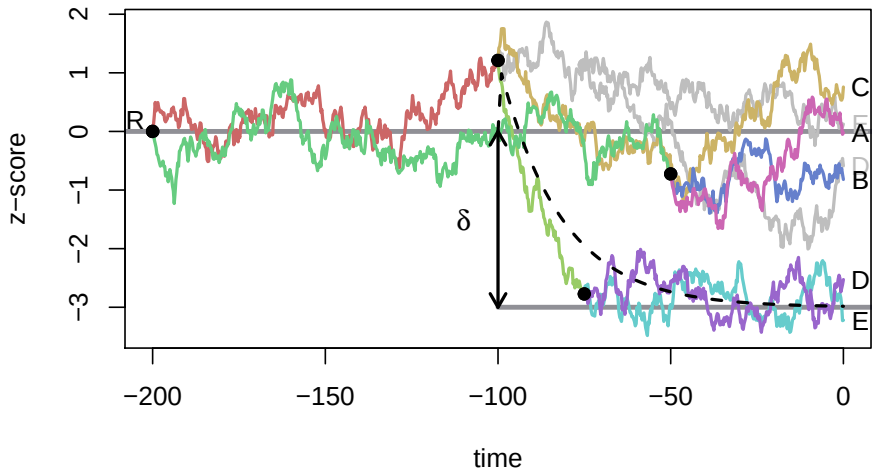




Shifts







Negative shift $\Rightarrow E[Z] \leq 0 \Rightarrow$ Small p-values $\Rightarrow$ Differential abundance

Mathematical Model (Cont'd)

Statistical Model

Assume that z-scores evolve as an OU on the tree with a sign constraint on the mean.

- $Z = (z_1, \ldots, z_n) \sim \mathcal{N}(M, \Sigma_\alpha)$ where
  - $M = (m_1, \ldots, m_n) \in \mathbb{R}^n_-$
  - Σ_α is the variance matrix of an OU process on a tree.

Mathematical Model (Cont'd)

Statistical Model

Assume that z-scores evolve as an OU on the tree with a sign constraint on the mean.

- $Z = (z_1, \ldots, z_n) \sim \mathcal{N}(M, \Sigma_\alpha)$ where
  - $M = (m_1, \ldots, m_n) \in \mathbb{R}^n_-$
  - Σ_α is the variance matrix of an OU process on a tree.

Goal

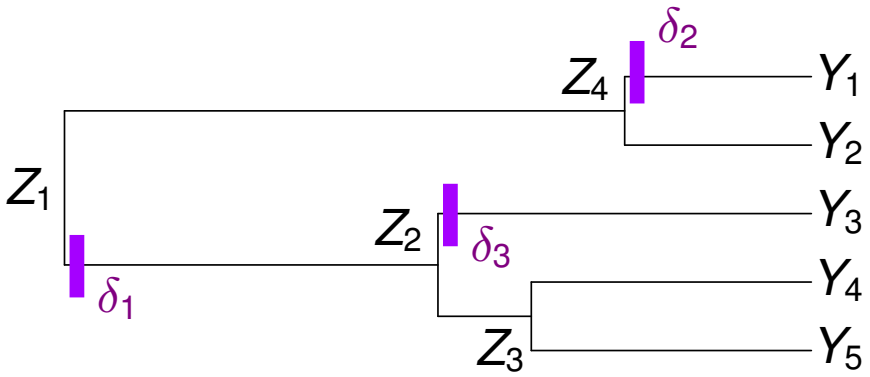
Find the negative entries of M

Linear regression model

Tree-structure enforced by decomposition $M = TW(\alpha)\Delta$.

Linear regression model

Tree-structure enforced by decomposition $M = TW(\alpha)\Delta$.



$$\Delta = \begin{pmatrix} 0 \\ \delta_1 \\ 0 \\ 0 \\ \delta_2 \\ 0 \\ \delta_3 \\ 0 \\ 0 \end{pmatrix} \qquad TW(\alpha)\Delta = \begin{pmatrix} w_5\delta_2 \\ 0 \\ w_2\delta_1 + w_7\delta_3 \\ w_2\delta_1 \\ w_2\delta_1 \end{pmatrix}$$

$$T = \begin{array}{c} \\ Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{array} \begin{array}{c} \begin{array}{ccccccccc} Z_1 & Z_2 & Z_3 & Z_4 & Y_1 & Y_2 & Y_3 & Y_4 & Y_5 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

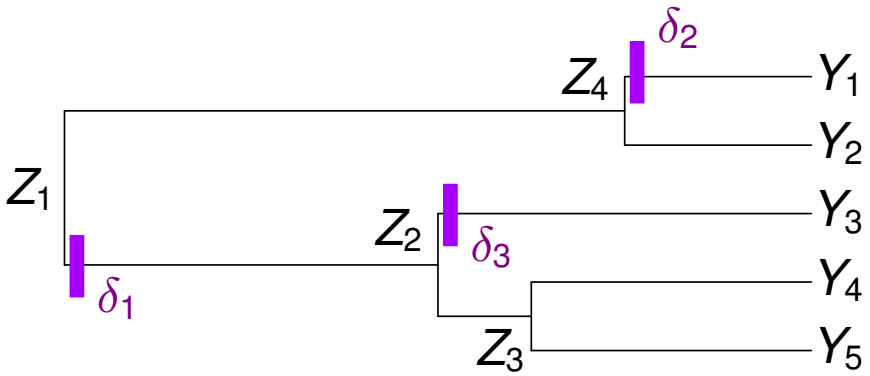
$$OU: \quad Z = TW(\alpha)\Delta + E$$

$$E \sim \mathcal{N}(0, V(\alpha))$$

$$W(\alpha) = \text{Diag}(1 - e^{-\alpha(h - t_{\text{pa}(i)})}, 1 \leq i \leq m + n)$$

Linear regression model

Tree-structure enforced by decomposition $M = TW(\alpha)\Delta$.



$$\Delta = \begin{pmatrix} 0 \\ \delta_1 \\ 0 \\ 0 \\ \delta_2 \\ 0 \\ \delta_3 \\ 0 \\ 0 \end{pmatrix} \qquad TW(\alpha)\Delta = \begin{pmatrix} w_5\delta_2 \\ 0 \\ w_2\delta_1 + w_7\delta_3 \\ w_2\delta_1 \\ w_2\delta_1 \end{pmatrix}$$

$$T = \begin{array}{c} \\ Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{array} \begin{array}{c} \begin{array}{ccccccccc} Z_1 & Z_2 & Z_3 & Z_4 & Y_1 & Y_2 & Y_3 & Y_4 & Y_5 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

$$OU: \quad Z = TW(\alpha)\Delta + E$$

$$E \sim \mathcal{N}(0, V(\alpha))$$

$$W(\alpha) = \text{Diag}(1 - e^{-\alpha(h - t_{\text{pa}(i)})}, 1 \leq i \leq m+n)$$

Goal: Find $\{i : m_i < 0\}$ through $\{j : \Delta_j \neq 0\}$

Outline

Estimating M

The MLE of Δ (and in turn M) is solution to

$$\underset{\Delta \text{ s.t. } TW(\alpha)\Delta \leq 0}{\operatorname{argmax}} \|Z - TW(\alpha)\Delta\|^2_{2,\Sigma_\alpha^{-1}}$$

Equivalent to[1]:

$$\underset{\Delta \text{ s.t. } C\Delta \leq 0}{\operatorname{argmax}} \|Y - X\Delta\|^2_2$$

[1] with C, Y and X some simple transforms of Z and $TW(\alpha)$, Σ_α

[2] Using a variant of the LASSO shooting algorithm

Estimating M

The MLE of Δ (and in turn M) is solution to

$$\underset{\Delta \text{ s.t. } TW(\alpha)\Delta \leq 0}{\operatorname{argmax}} \|Z - TW(\alpha)\Delta\|^2_{2,\Sigma_\alpha^{-1}}$$

Equivalent to[1]:

$$\underset{\Delta \text{ s.t. } C\Delta \leq 0}{\operatorname{argmax}} \|Y - X\Delta\|_2^2$$

Add a ℓ_1-penalty to sparsify the solution and solve[2]

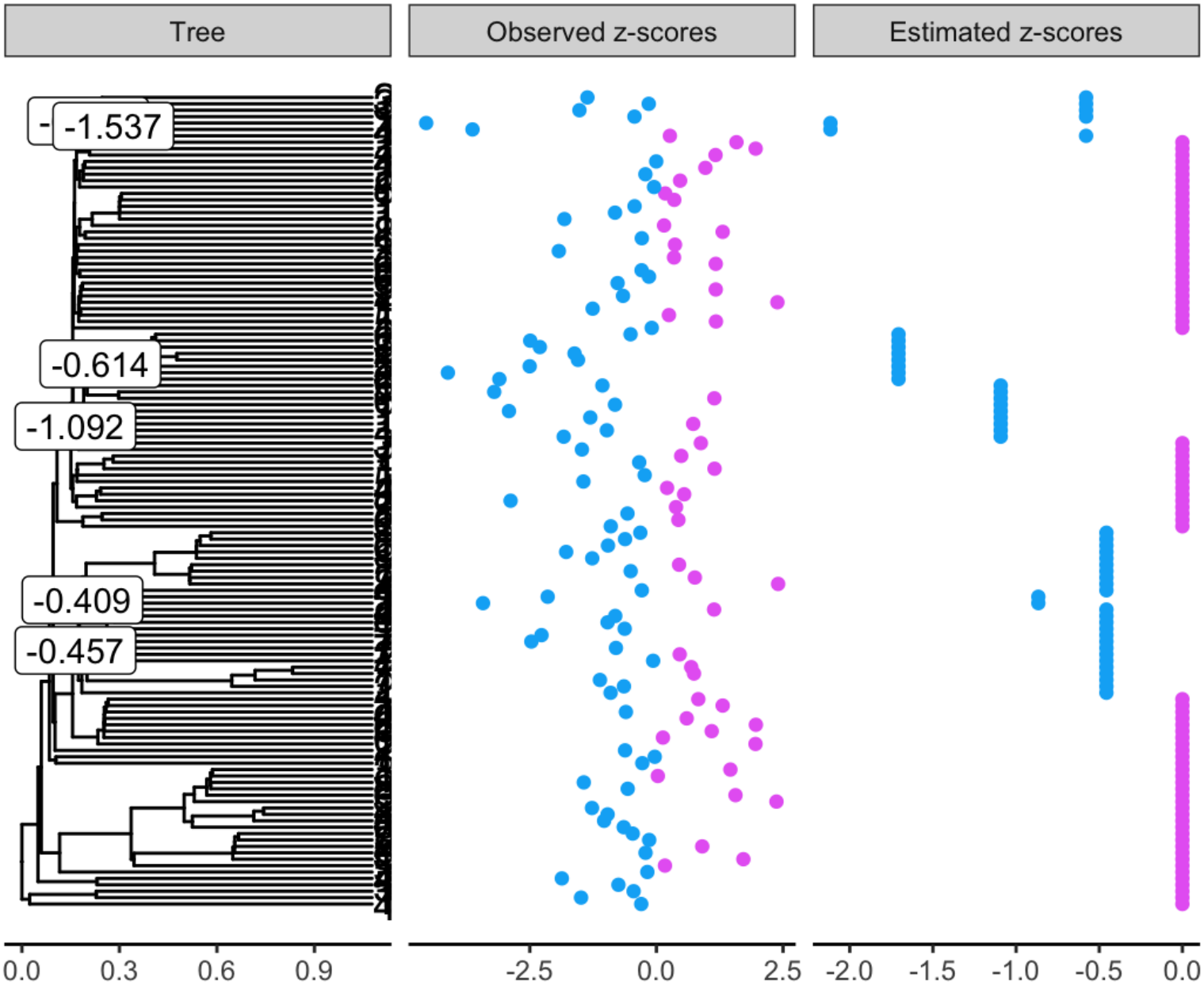
$$\hat{\Delta} = \underset{\Delta \text{ s.t. } C\Delta \leq 0}{\operatorname{argmax}} \|Y - X\Delta\|_2^2 + \lambda\|\Delta\|_1$$

using penalized likelihood for selection of α and λ

[1] with C, Y and X some simple transforms of Z and $TW(\alpha)$, Σ_α

[2] Using a variant of the LASSO shooting algorithm

Illustration



Confidence intervals for m_i

Estimates of $\hat{\Delta}$ (and $\hat{M}$) are **biased** and lack **confidence intervals**

Confidence intervals for m_i

Estimates of $\hat{\Delta}$ (and $\hat{M}$) are **biased** and lack **confidence intervals**

Debiasing (Zhang and Zhang, 2014)

- One-step correction of the LASSO estimator

$$\hat{\Delta}_j^{debias} = \hat{\Delta}_j + \frac{S_j^\top (Y - X\hat{\Delta})}{S_j^\top X_j}$$

- Where S is a relaxed Graham-Schmidt orthogonalization of X.

Confidence intervals for m_i

Estimates of $\hat{\Delta}$ (and $\hat{M}$) are **biased** and lack **confidence intervals**

Debiasing (Zhang and Zhang, 2014)

- One-step correction of the LASSO estimator

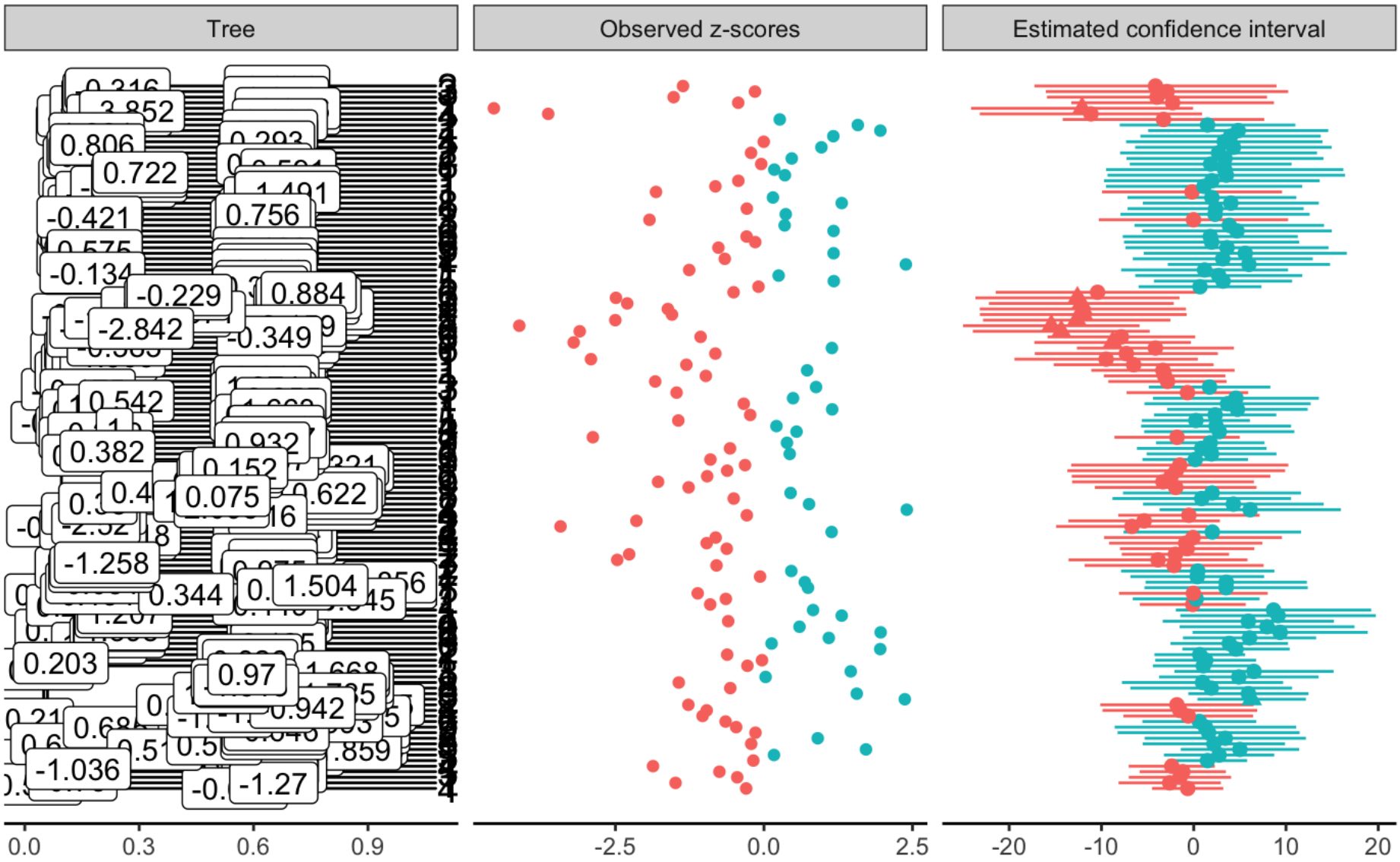
$$\hat{\Delta}_j^{debias} = \hat{\Delta}_j + \frac{S_j^\top (Y - X\hat{\Delta})}{S_j^\top X_j}$$

- Where S is a relaxed Graham-Schmidt orthogonalization of X.

Confidence intervals (Javanmard et al., 2019)

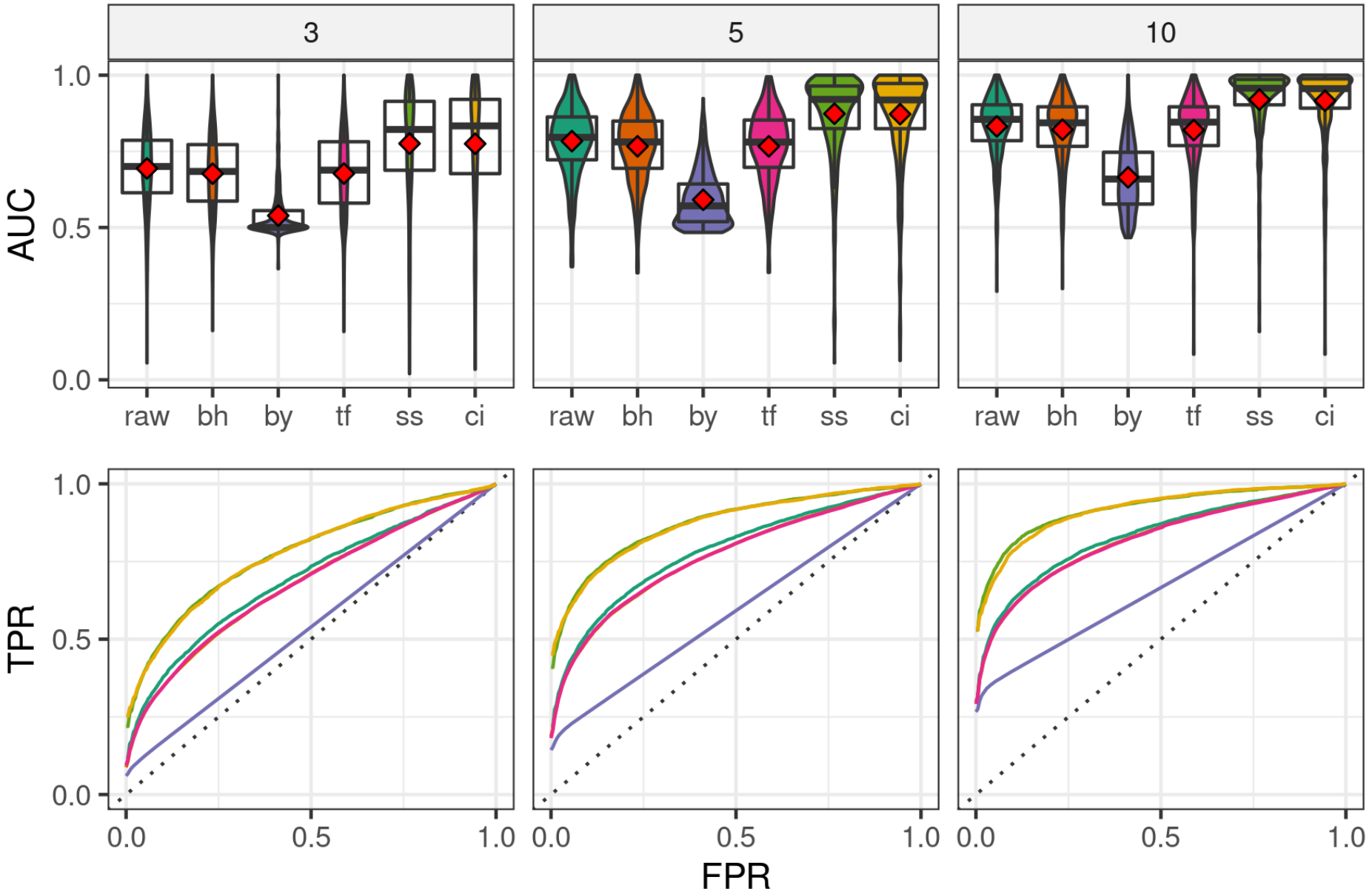
- BH-like procedure based on asymptotic normality of the $\hat{\Delta}_j^{debias}$

Illustration

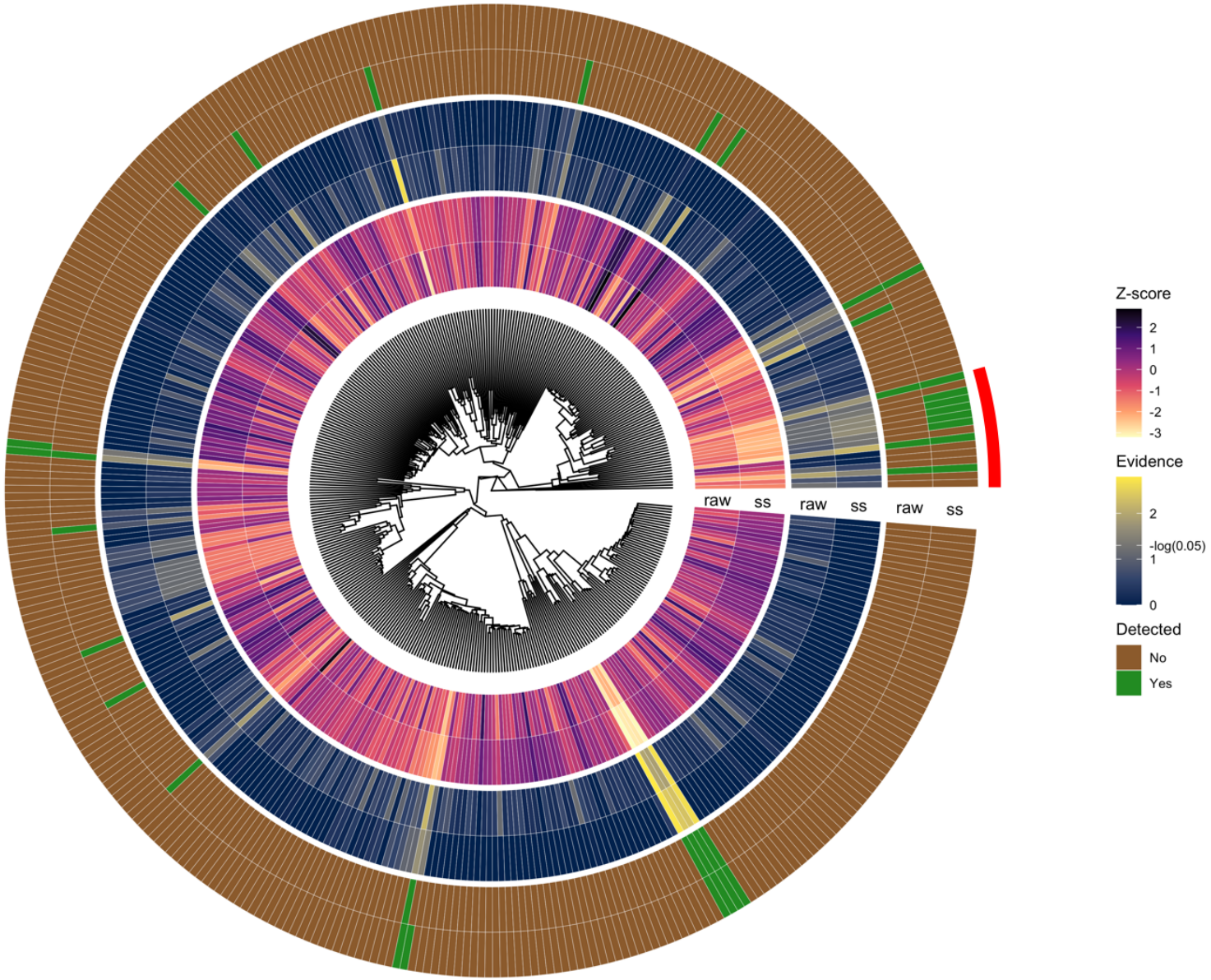


Outline

Simulation results: zazou (ss/ci) is competitive



Fiji adult and children gut microbiome (Brito et al., 2016)



Conclusion

- A framework for differential analyses that integrates taxonomic information
- Based on combining LASSO, trees and stochastic processes
- That performs well on simulations
- Implemented as a github R package (`abichat/zazou`)
- Technical details available in Bichat et al. (`arXiv:2009.13335`)

Bibliography

A. Bichat, C. Ambroise, and M. Mariadassou. Hierarchical correction of *p*-values via a tree running ornstein-uhlenbeck process.

I. L. Brito, S. Yilmaz, K. Huang, L. Xu, S. D. Jupiter, A. P. Jenkins, W. Naisilisili, M. Tamminen, C. S. Smillie, J. R. Wortman, B. W. Birren, R. J. Xavier, P. C. Blainey, A. K. Singh, D. Gevers, and E. J. Alm. Mobile genes in the human microbiome are structured from global to individual scales. *Nature*, 535(7612):435–439, July 2016. doi: 10.1038/nature18927. URL https://doi.org/10.1038/nature18927.

J. Felsenstein. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39(4):783–791, July 1985. doi: 10.2307/2408678. URL http://links.jstor.org/sici?sici=0014-3820(198507)39:4%3C783:CLOPAA%3E2.0.CO;2-L.

T. F. Hansen. Stabilizing selection and the comparative analysis of adaptation. *Evolution*, 51(5):1341–1351, Oct. 1997. URL http://www.jstor.org/stable/2411186.

A. Javanmard, H. Javadi, et al. False discovery rate control via debiased lasso. *Electronic Journal of Statistics*, 13(1):1212–1253, 2019.

J. Ravel, P. Gajer, Z. Abdo, G. M. Schneider, S. S. K. Koenig, S. L. McCulle, S. Karlebach, R. Gorle, J. Russell, C. O. Tacket, and et al. Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences*, 108(Suppl. 1): 4680–4687, March 2011. ISSN 1091-6490. doi: 10.1073/pnas.1002611107. URL http://dx.doi.org/10.1073/pnas.1002611107.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.