Advances in Bayesian methods for complex data: Discussion

giovanni_parmigiani@dfci.harvard.edu

ISBA World Meeting, Montréal 2022

disclosures

- Consultations Related to the Topic:
  Foundation Medicine, Delfi Diagnostics
- Speaker's Bureau: None
- Grant/Research support from: NIH-NCI, NSF
  Licensing of BayesMendel software for genetic counseling.
  Licensing of Ask2me database.
- Stockholder in: Phaeno Biotechnology
- Honoraria from: Academic Only
- co-Founder / Chief Scientific Officer: Phaeno Biotechnology
- Relevant patents: POSTN for debulking in OC
- Employee of: Dana Farber Cancer Institute

Congratulations for three great papers!

Papers do honor the application context

Yet they contain ideas that can be helpful in many other contexts

BORA-GP

Scalable Gaussian Processes on Physically Constrained Domains


PARMIGIANO
REGGIANO

Jin & Tonic !

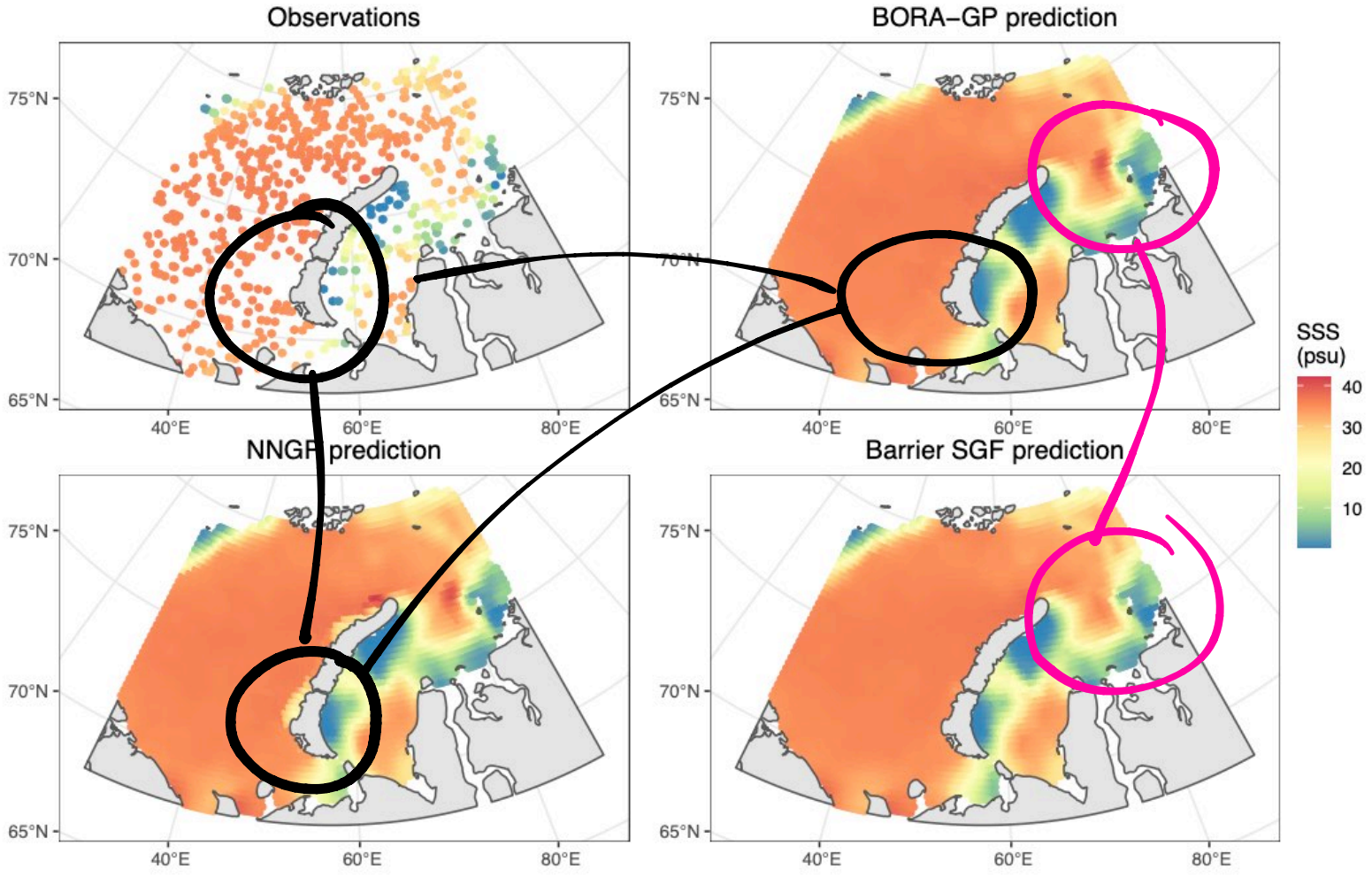


Figure 3: Predicted results from BORA-GP, NNGP, and Barrier SGF with training data depicted in top-left panel.

Thoughts

I agree that barriers may be very important in this type of application.

BORA-GP seems to be best suited for "neutral" barriers which do not interfere with the spatial correlation of the rest of the process

Perhaps in some cases barriers may inform about the dependence (e.g. island with multiple glaciers)

Questions

What suggestions do you have for diagnosing whether barriers interfere with the spatial correlation of the rest of the process?

Did you encounter cases in which relevant barriers exist but do not happen to land on the edge between two observed points? (e.g. boundary barriers; small islands with big glaciers)

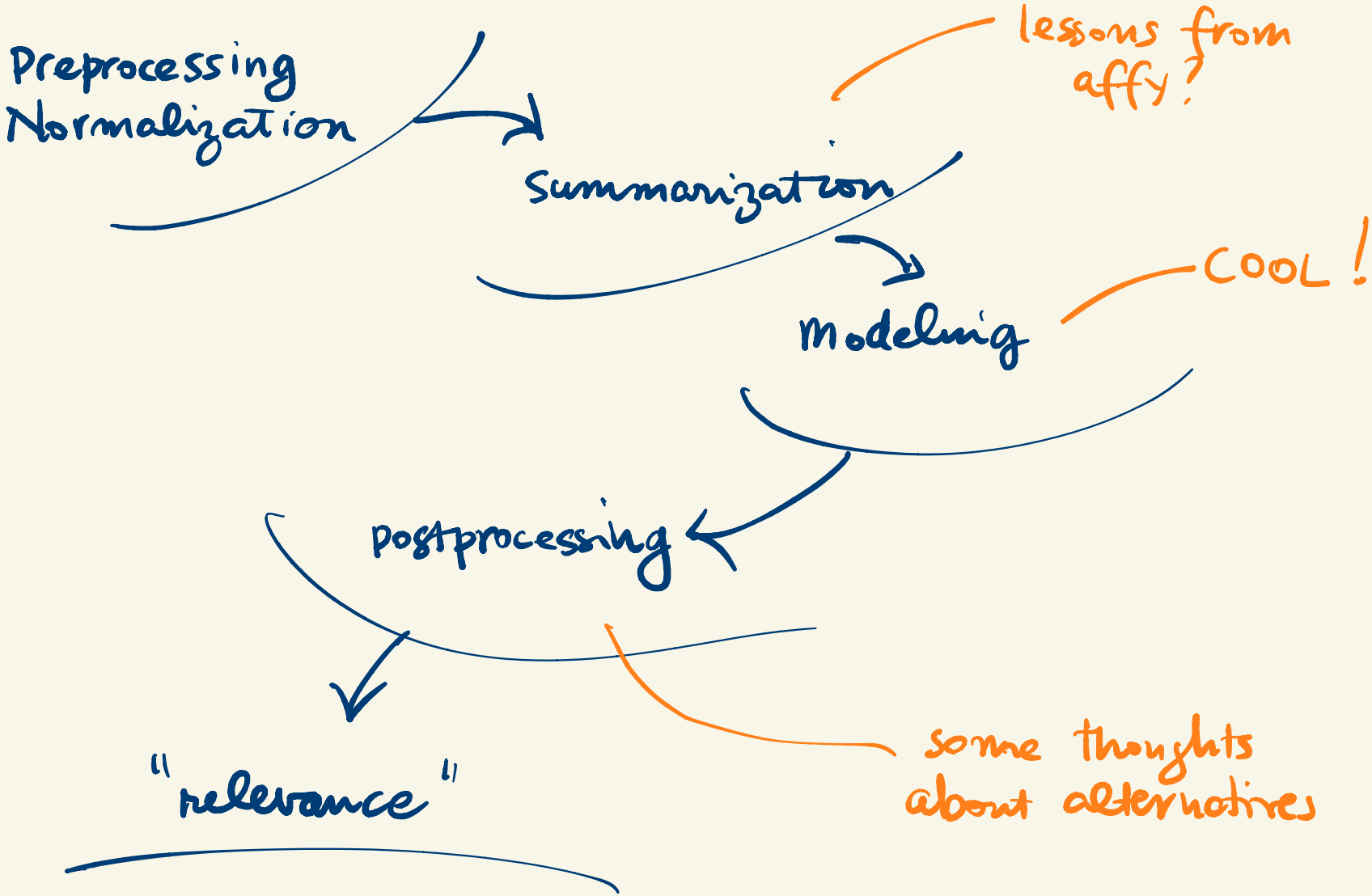
Bayesian screening via mixtures of shrinkage priors with applications to light-sheet fluorescence microscopy in brain imaging

When is $>$ 2 better than 2?

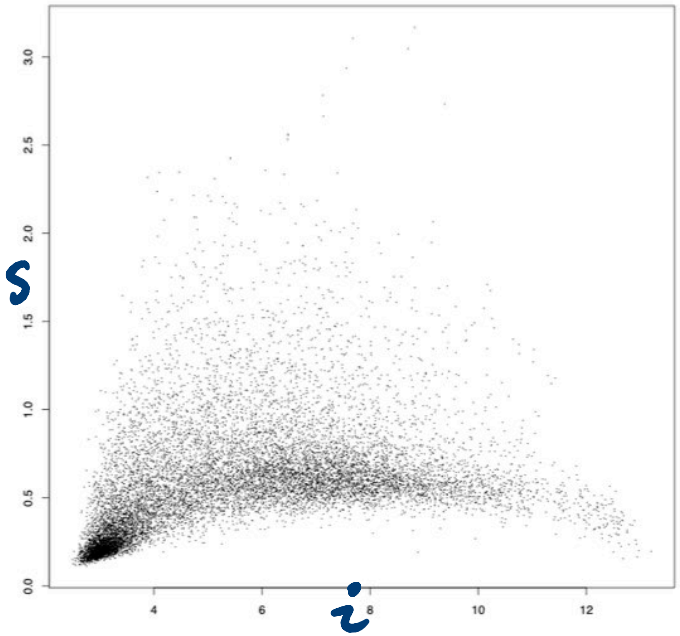
When you have more than two options to pick from

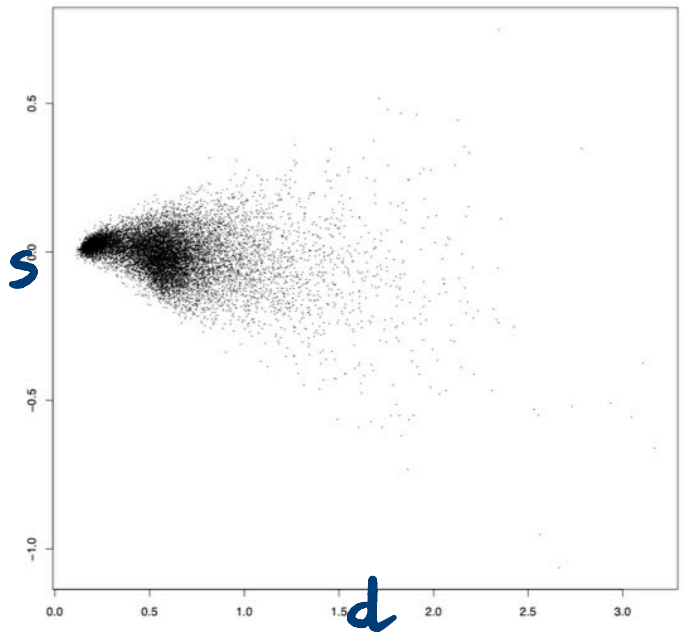
When you have more than two biological regimes to discover

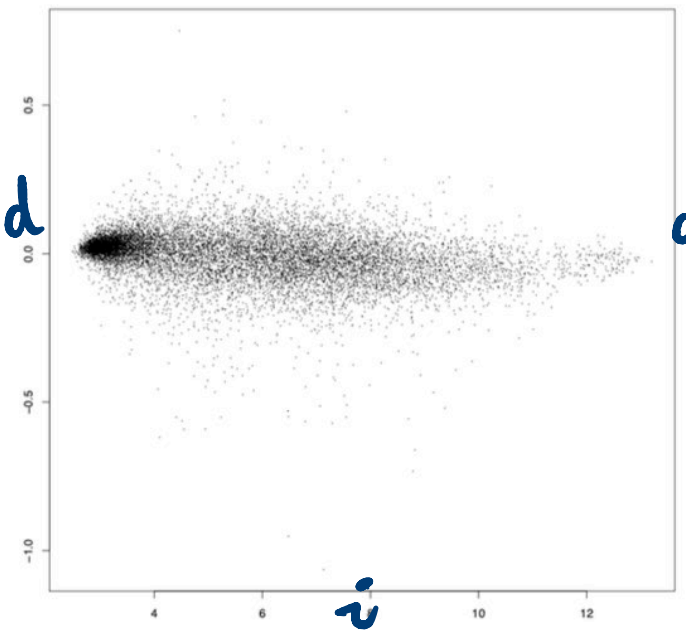
When you can communicate rough categories of "risk" better than real-valued probabilities

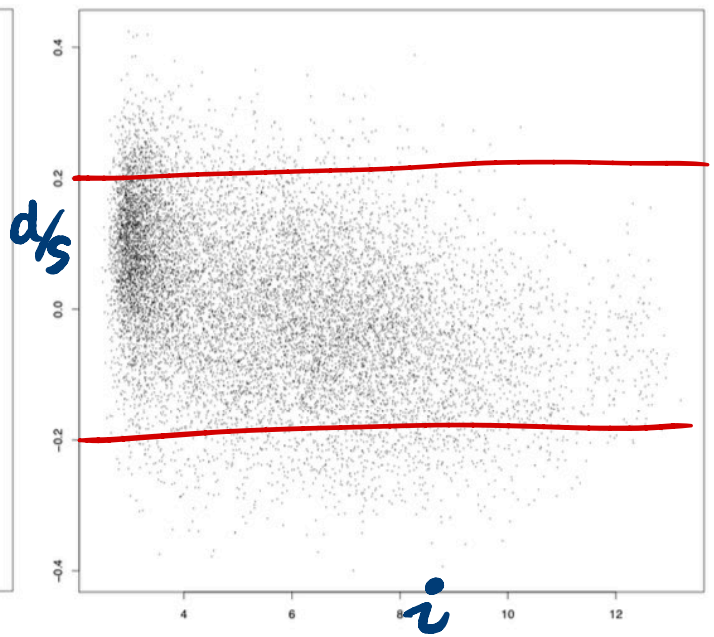

Preprocessing
Normalization
lessons from affy?
Summarization
Modeling
COOL!
postprocessing
"relevance"
some thoughts about alternatives

A lesson from old fluorescence-based expression?









$z \propto d/s$

d = difference

s = standard deviation

i = average intensity

Caveat!
here points are genes, not cells

from:
CuratedOvarianData

False discovery rates: a new deal

Matthew Stephens

Biostatistics, Volume 18, Issue 2, April 2017, Pages 275–294,
https://doi.org/10.1093/biostatistics/kxw041
Published: 17 October 2016 **Article history**

Johns Hopkins University, Dept. of Biostatistics Working Papers

4-16-2004

Screening for Differentially Expressed Genes: Are Multilevel Models Helpful?

Dongmei Liu
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Giovanni Parmigiani
The Sydney Kimmel Comprehensive Cancer Center, Johns Hopkins University, gp@jimmy.harvard.edu

Brian Caffo
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Questions

Do you see a way to use the latent class labels for the final clustering?
(Perhaps after "freezing" the chain at a preferred L).

How do you think about setting a threshold on the cluster posterior probabililty (in your case .05)?

MOReTreeS prior

Applications and Case Studies

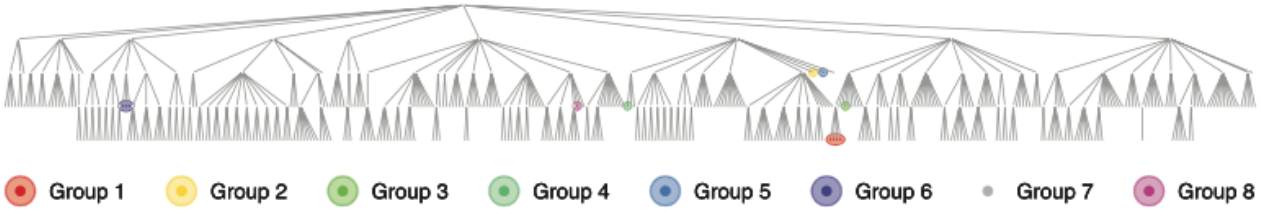
Estimating the Effects of Fine Particulate Matter on 432 Cardiovascular Diseases Using Multi-Outcome Regression With Tree-Structured Shrinkage

Emma G. Thomas, Lorenzo Trippa, Giovanni Parmigiani & Francesca Dominici

Pages 1689-1699 | Received 17 Jan 2019, Accepted 22 Jan 2020, Accepted author version posted online: 24 Feb 2020, Published online: 26 Feb 2020

Download citation https://doi.org/10.1080/01621459.2020.1722134 Check for updates

Table 1. Estimated increase in odds of hospitalization for diseases of the circulatory system associated with a 10 $\mu g\, m^{-3}$ increase in short-term exposure to $PM_{2.5}$.

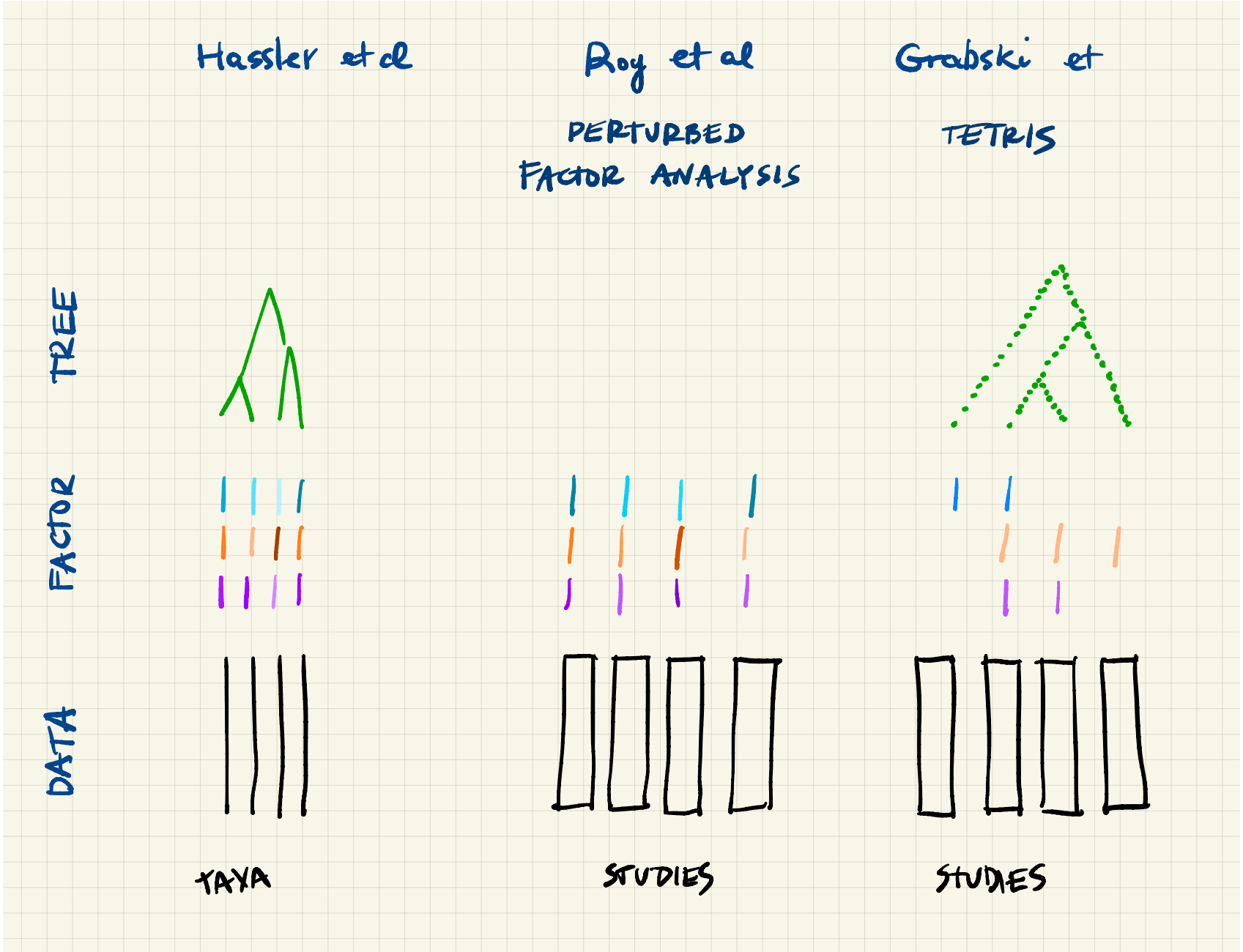


Group	ICD9 codes	#Cases	Odds ratio (95% credible/confidence interval) ssMOReTreeS	Odds ratio (95% credible/confidence interval) Maximum likelihood
1	Dissection of aorta (441.00, 441.01, 441.02, 441.03)	41,358	0.940 (0.917,0.963)	0.932 (0.907,0.956)
2	Subarachnoid hemorrhage (430)	61,488	0.945 (0.926,0.965)	0.940 (0.920,0.960)
3	Abdominal aneurysm, ruptured (441.3)	39,902	0.949 (0.925,0.972)	0.941 (0.917,0.967)
4	Unspecified intracranial hemorrhage (432.9)	29,030	0.959 (0.930,0.988)	0.950 (0.919,0.981)
5	Intracerebral hemorrhage (431)	317,447	0.960 (0.951,0.969)	0.959 (0.950,0.968)
6	Essential hypertension (401.0, 401.1, 401.9)	310,148	0.977 (0.968,0.986)	0.977 (0.967,0.986)
7	All 420 remaining codes	19,056,207	1.011 (1.010,1.012)	1.011 (1.010,1.012)
8	Congestive heart failure, unspecified (428.0)	4,083,787	1.019 (1.017,1.022)	1.019 (1.017,1.022)

NOTE: Results are shown for groups of ICD9 codes discovered using ssMOReTreeS. Supplement Section I shows a full list of ICD9 codes associated with each group. Estimates and credible intervals from ssMOreTreeS are shown along with point estimates and confidence intervals from fitting maximum likelihood conditional logistic regression models to the same outcome groups. The plot above the table shows the tree of outcomes included in our analysis (ICD9 codes 350–459).

Principled, practical, flexible, fast: a new approach to phylogenetic factor analysis

Related methods from multi-study factor analysis



Thoughts

The Hassler *et al.* tree structure could be used in other contexts to leverage domain knowledge about similarity of subjects (studies) and relax exchangeability.

I agree with Hassler *et al.* that we need to think in a principled way about postprocessing chain results. This is an important frontier of Bayesian foundations.

Questions

Does your method already cover the case where you have multiple phenotypic profiles generated in organisms from the same taxon?

Have you made any preliminary inroads in the "unknown tree" or "approximately known tree" cases?