The data-processing multiverse: achieving reconciliation for Christmas

Brice Ozenne[1,2], Martin Nørgaard[1,3], Cyril Pernet[1], Melanie Ganz[1,4]

[1] Neurobiology Research Unit, University Hospital of Copenhagen, Rigshospitalet.

[2] Section of Biostatistics, Department of Public Health, University of Copenhagen.

[3] Center for Reproducible Neuroscience, Stanford University.

[4] Department of Computer Science, University of Copenhagen.

December 2nd , 2022, NRU Christmas Symposium

The data-processing multiverse

Neuroimaging is used to study brain structure and function

- indirect way of measuring brain signals
- contaminated by multiple sources of noise

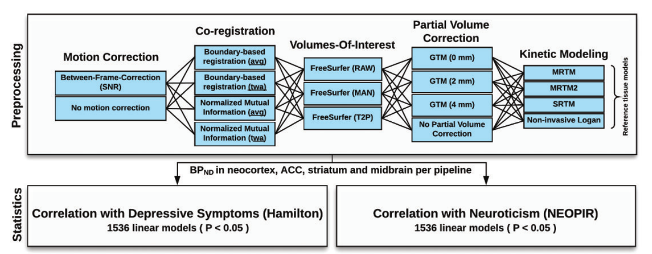
Data preprocessing is critical to decontaminate the signal

The data-processing multiverse

Neuroimaging is used to study brain structure and function

- indirect way of measuring brain signals
- contaminated by multiple sources of noise

Data preprocessing is critical to decontaminate the signal



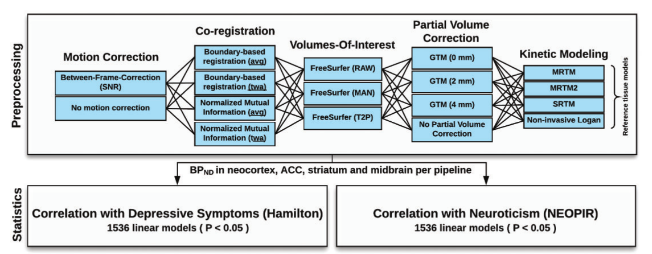
Nørgaard et al 2020. DOI: 10.1177/0271678X19880450

The data-processing multiverse

Neuroimaging is used to study brain structure and function

- indirect way of measuring brain signals
- contaminated by multiple sources of noise

Data preprocessing is critical to decontaminate the signal



Nørgaard et al 2020. DOI: 10.1177/0271678X19880450

⚠ many possibilities!

⚠ impacts the conclusion of the study

How it feels


https://www.outdoorrevival.com/instant-articles/getting-lost-forest-survival-guide.html?firefox=1

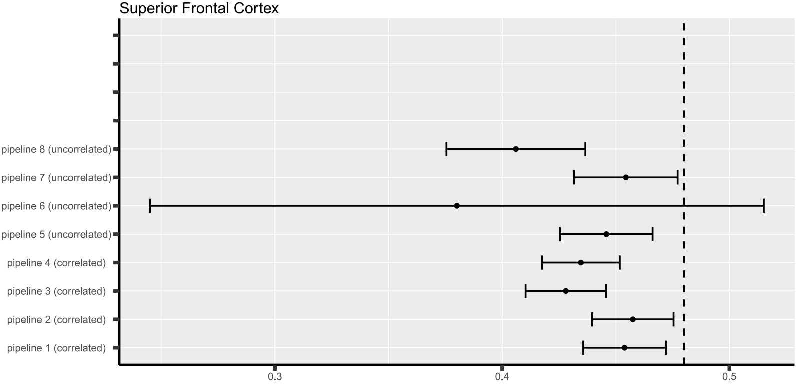
How it feels

Need for a statistical framework:

- aggregate evidence from analyses based on different pipelines

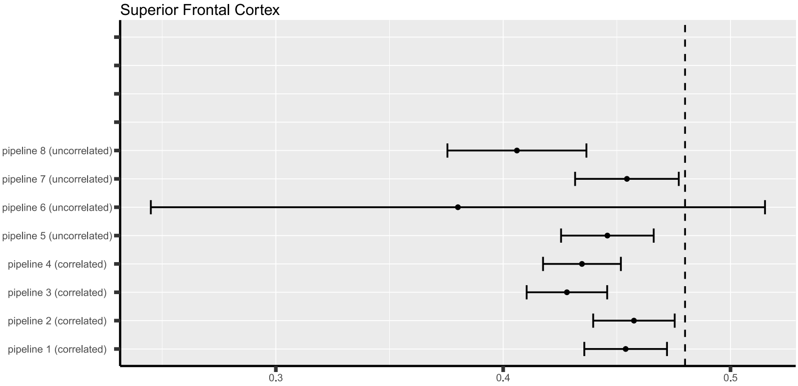
$\rightarrow$ conclusions robust to the choice of pipeline!

A forest plot!



A common estimate?

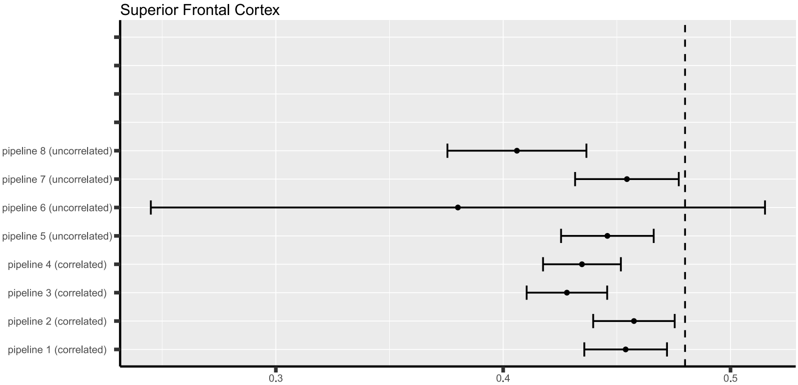
A forest plot!



A common estimate?

$\Psi_{\textbf{average}}$ average

A forest plot!

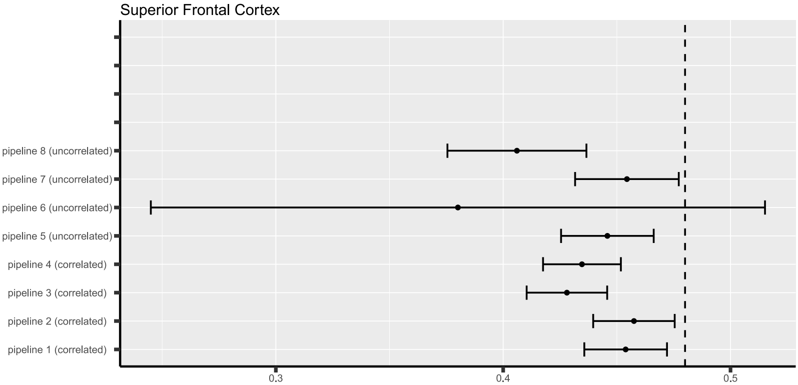


A common estimate?

$\Psi_{\textbf{average}}$ average

$\Psi_{\textbf{pool-se}}$ ...inversely proportional to the uncertainty

A forest plot!



A common estimate?

$\Psi_{\textbf{average}}$ average

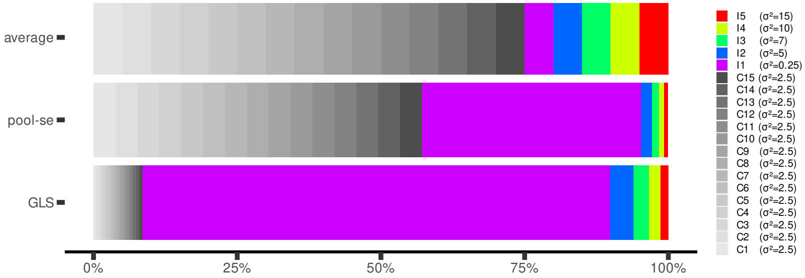
$\Psi_{\textbf{pool-se}}$ ...inversely proportional to the uncertainty

$\Psi_{\textbf{GLS}}$ ...of independent combinations of estimates

Example (scenario 3)

Pipelines:

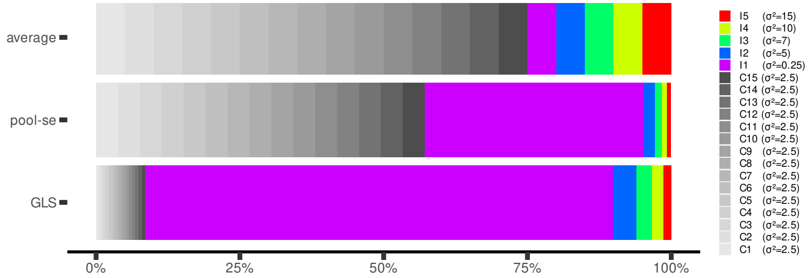
- 15 very correlated with moderate uncertainty ($\rho = 0.95$, $\sigma^2 = 2.5$)
- 5 independent with low to high uncertainty ($\sigma^2 \in [0.25, 15]$)



Example (scenario 3)

Pipelines:

- 15 very correlated with moderate uncertainty ($\rho = 0.95$, $\sigma^2 = 2.5$)
- 5 independent with low to high uncertainty ($\sigma^2 \in [0.25, 15]$)

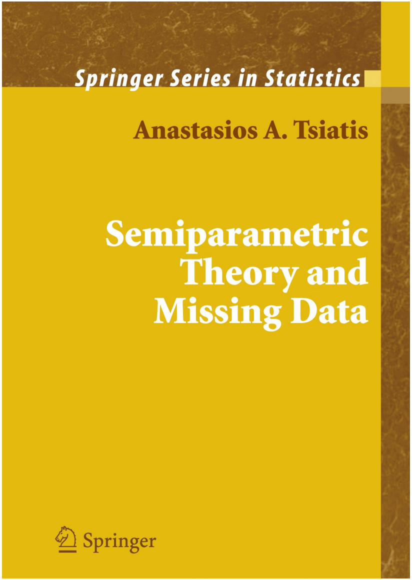


😕 . . . but how do we estimate the correlation?

- we only have one estimate per pipeline

The christmas tale book

The christmas tale book . . . for statisticians



3

The Geometry of Influence Functions

As we will describe shortly, most reasonable estimators for the parameter β, in either parametric or semiparametric models, are asymptotically linear and can be uniquely characterized by the influence function of the estimator. The class of influence functions for such estimators belongs to the Hilbert space of all mean-zero q-dimensional random functions with finite variance that was defined in Chapter 2. As such, this construction will allow us to view estimators or, more specifically, the influence function of estimators, from a geometric point of view. This will give us intuitive insight into the construction of such estimators and a geometric way of assessing the relative efficiencies of the various estimators.

As always, consider the statistical model where $Z_1, \ldots, Z_n$ are iid random vectors and the density of a single Z is assumed to belong to the class $\{p_Z(z;\theta), \theta \,\epsilon\, \Omega\}$ with respect to some dominating measure ν_Z. The parameter θ can be written as $(\beta^T, \eta^T)^T$, where $\beta^{q\times 1}$ is the parameter of interest and η, the nuisance parameter, may be finite- or infinite-dimensional. The truth will be denoted by $\theta_0 = (\beta_0^T, \eta_0^T)^T$. For the remainder of this chapter, we will only consider parametric models where $\theta = (\beta^T, \eta^T)^T$ and the vector θ is p-dimensional, the parameter of interest β is q-dimensional, and the nuisance parameter η is r-dimensional, with $p = q + r$.

An estimator $\hat\beta_n$ of β is a q-dimensional measurable random function of $Z_1, \ldots, Z_n$. Most reasonable estimators for β are *asymptotically linear*; that is, there exists a random vector (i.e., a q-dimensional measurable random function) $\varphi^{q\times 1}(Z)$, such that $E\{\varphi(Z)\} = 0^{q\times 1}$,

$$n^{1/2}(\hat\beta_n - \beta_0) = n^{-1/2} \sum_{i=1}^{n} \varphi(Z_i) + o_p(1), \tag{3.1}$$

where $o_p(1)$ is a term that converges in probability to zero as n goes to infinity and $E(\varphi\varphi^T)$ is finite and nonsingular.

Remark 1. The function $\varphi(Z)$ is defined with respect to the true distribution $p(z, \theta_0)$ that generates the data. Consequently, we sometimes may write

A christmas gift 🥳

$\varphi(Z_i)$: influence function relative to observation i

- pseudo-observation of the individual effect

We now have n "estimates" per pipeline!

- easy to evaluate correlation between pipeline estimates

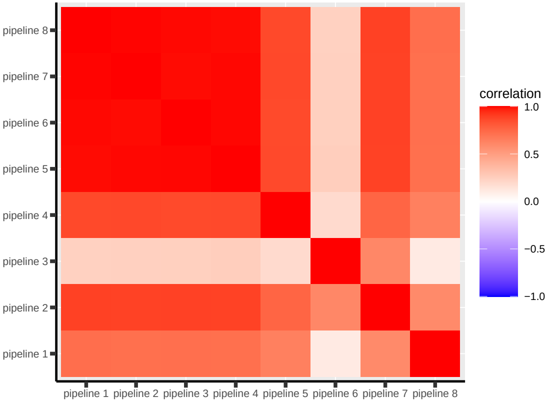
A christmas gift 🥳

$\varphi(Z_i)$: influence function relative to observation i

- pseudo-observation of the individual effect

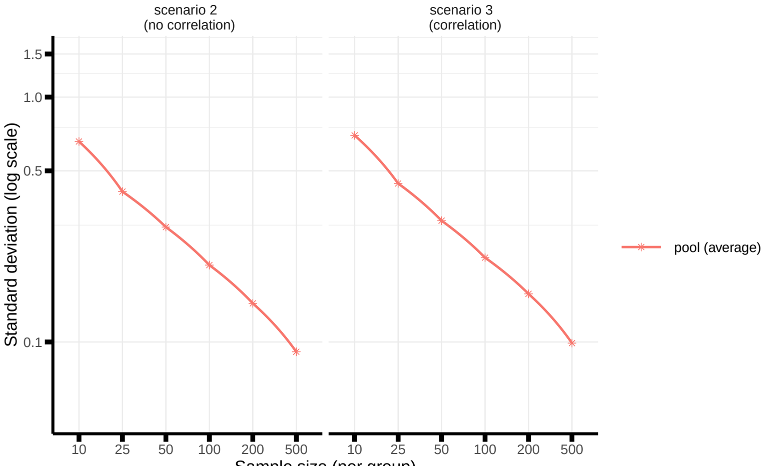
We now have n "estimates" per pipeline!

- easy to evaluate correlation between pipeline estimates



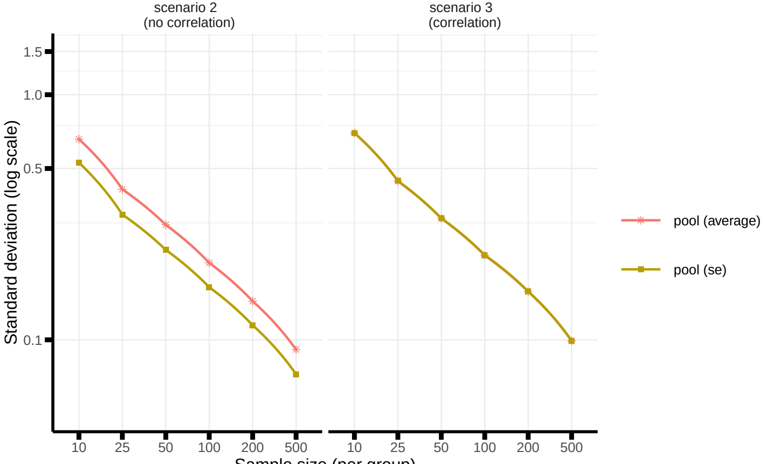
Simulation results

- No bias
- Uncertainty (lower is better)



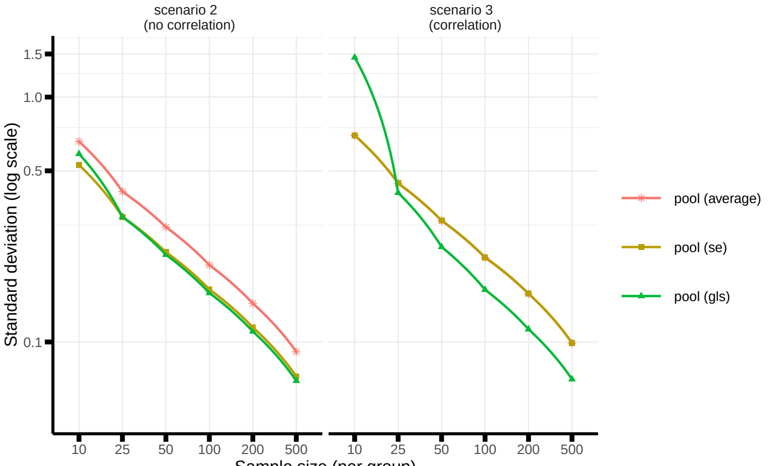
Simulation results

- No bias
- Uncertainty (lower is better)



Simulation results

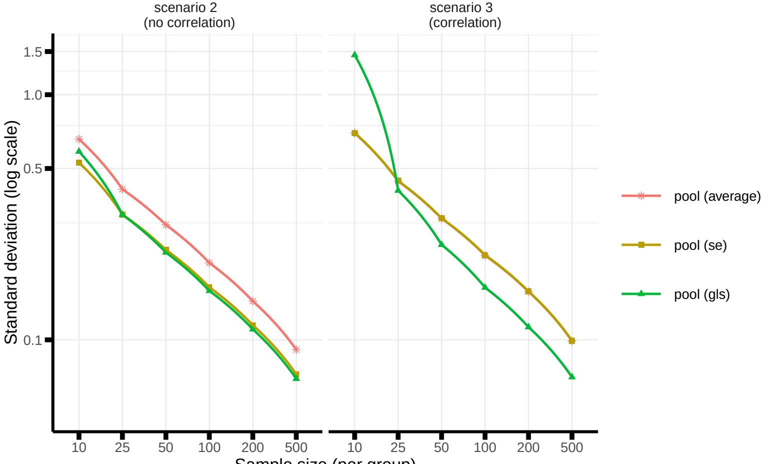
- No bias
- Uncertainty (lower is better)



Simulation results

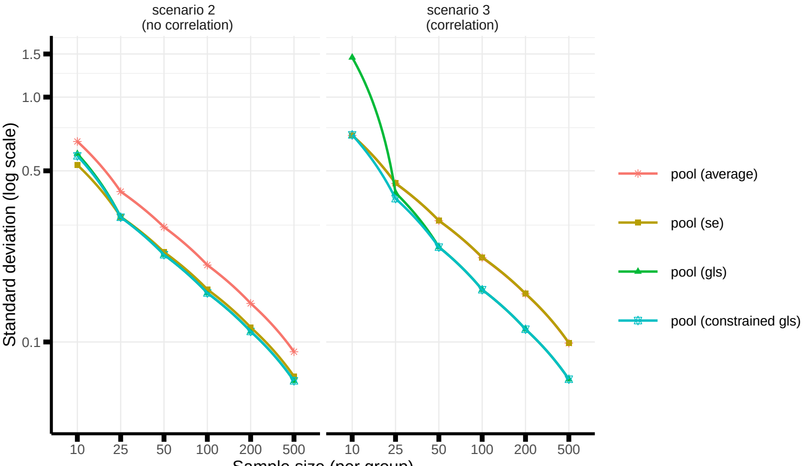
- No bias
- Uncertainty (lower is better)

  → GLS: poor performance with NRU typical sample size 😱

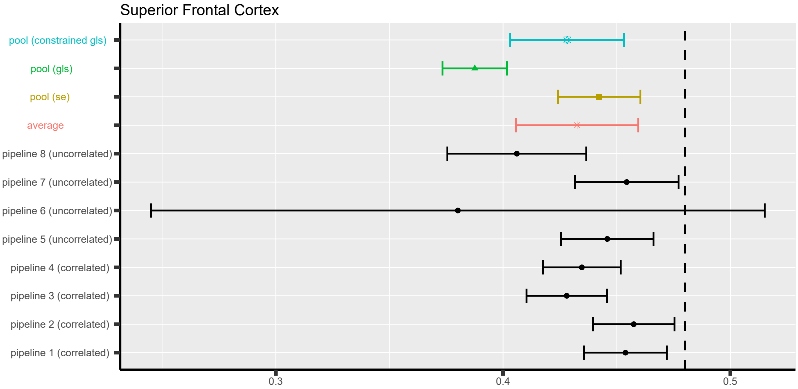


Simulation results

- No bias
- Uncertainty (lower is better)



Real data results



Wrap-up

A statistical framework for "sensitivity analysis" for neuroimaging

- **visualize heterogeneity** across pipelines
- estimate a **global effect** across pipelines
- quantify **proportion of pipelines with evidence** for an effect
- **test hypotheses** across pipelines

On-going project

- working paper & software (R package `LMMstar`)

Future

- handling "biased" pipelines

Reference I