Brain Connectivity-Informed Adaptive Regularization for Generalized Outcomes

Jaroslaw Harezlak, Ph.D.

Professor and Interim Co-Chair
Department of Epidemiology and Biostatistics
Indiana University School of Public Health
Bloomington, IN, USA

May 22, 2020

Outline

1. Motivating application
2. Brain structure and connectivity
3. Regularization methods
4. riPEER - ridgified Partially Empirical Eigenvectors for Regression
5. Simulation study
6. Brain structure and HIV infection
7. Discussion

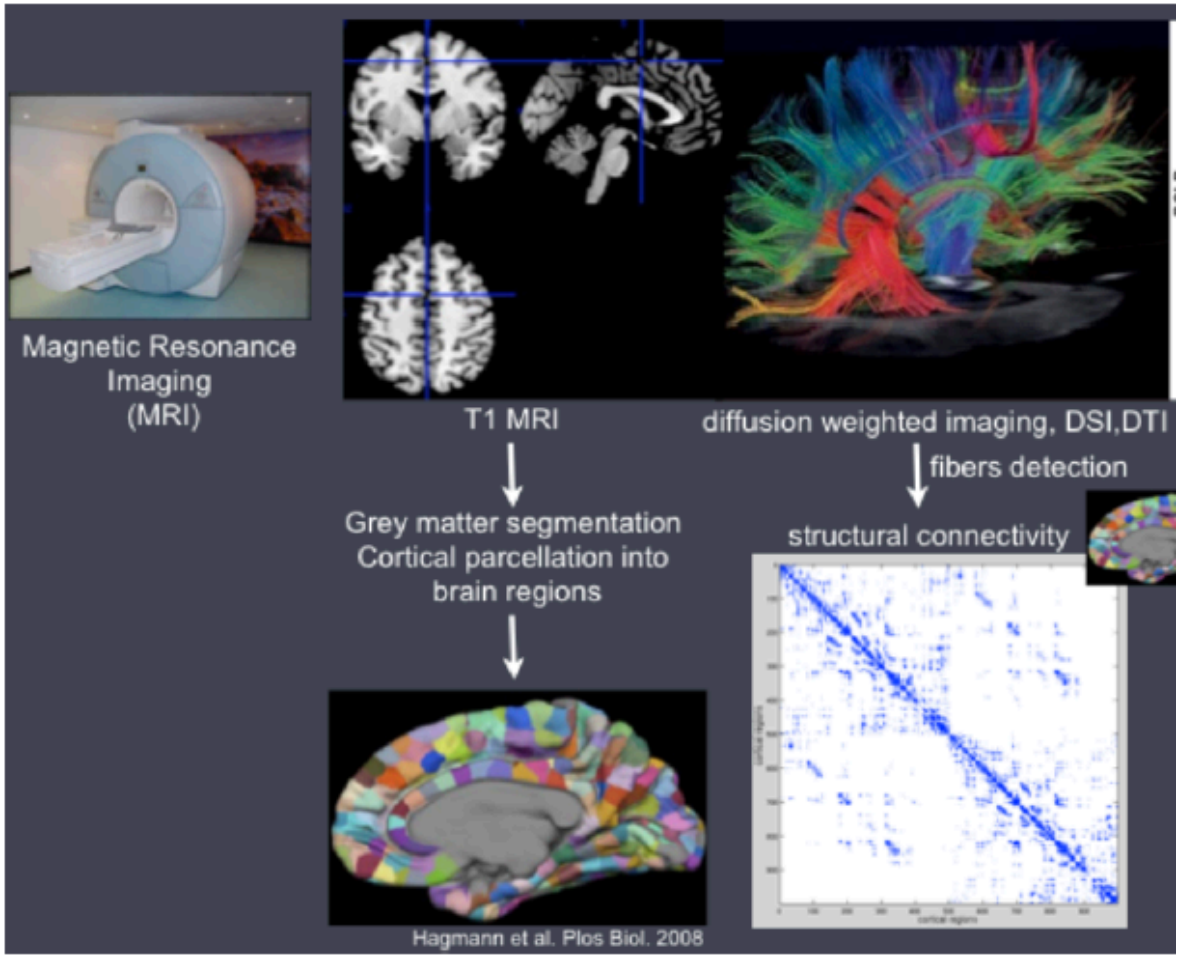
HIV infection study - WUSM

1. N = 299 HIV-infected individuals: 228 males, 71 females
   - Duration of infection range: 0 - 33y (mean: 10.2, sd: 8.7)
   - Age range: 18 - 84 y.o. (mean: 42.3, sd: 16)
2. Imaging modalities
   - T1 - anatomy
   - DTI - structural connectivity

Anatomy and connectivity

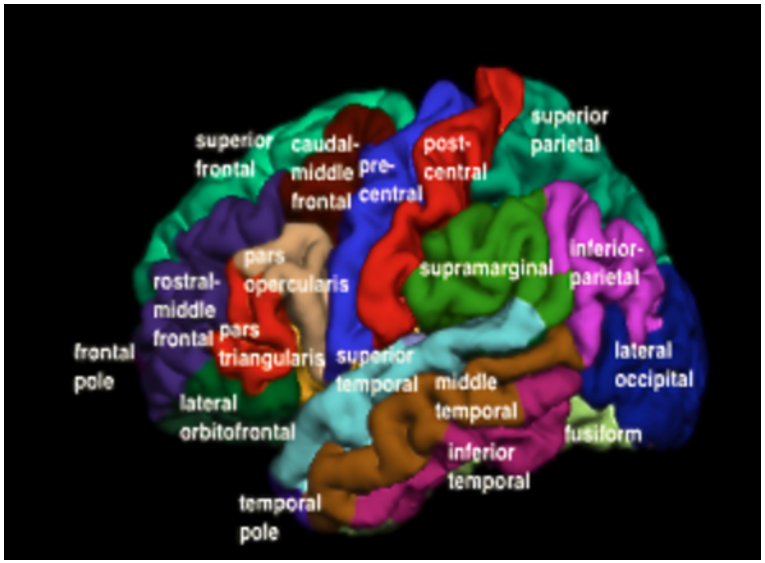
1. Anatomy
   - MPRAGE protocol
   - Processing using FreeSurfer software (version 5.1)
   - Desikan-Killiany atlas - 66 cortical regions
2. Structural connectivity
   - DTI and maximal diffusion coherence model
   - Density of connections between each pair of regions

MRI data

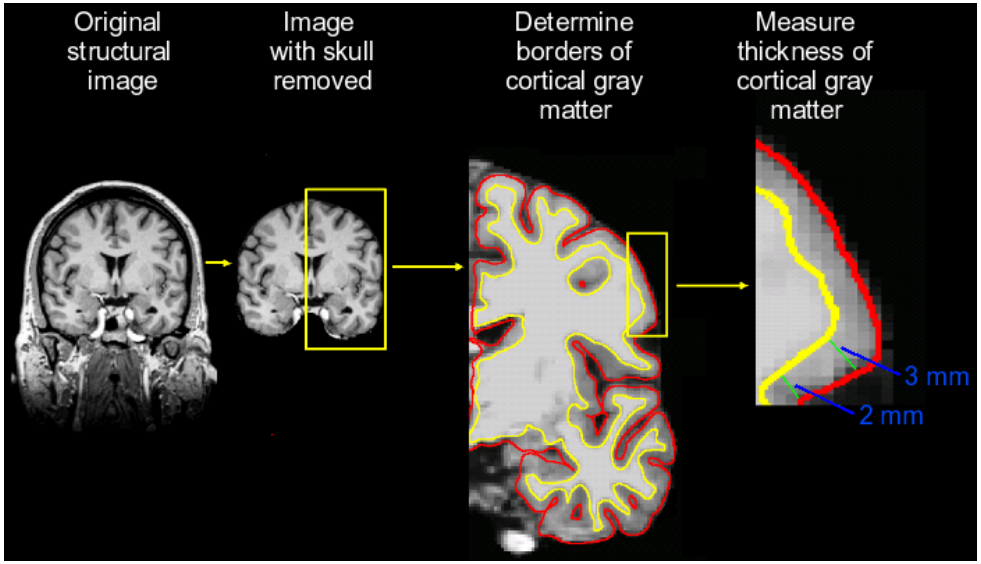

Magnetic Resonance Imaging (MRI)
T1 MRI
diffusion weighted imaging, DSI,DTI
fibers detection
Grey matter segmentation
Cortical parcellation into brain regions
structural connectivity
Hagmann et al. Plos Biol. 2008

MRI-derived data: cortical thickness

1. Parcellation of the cortex into 66 regions
2. Average cortical thickness

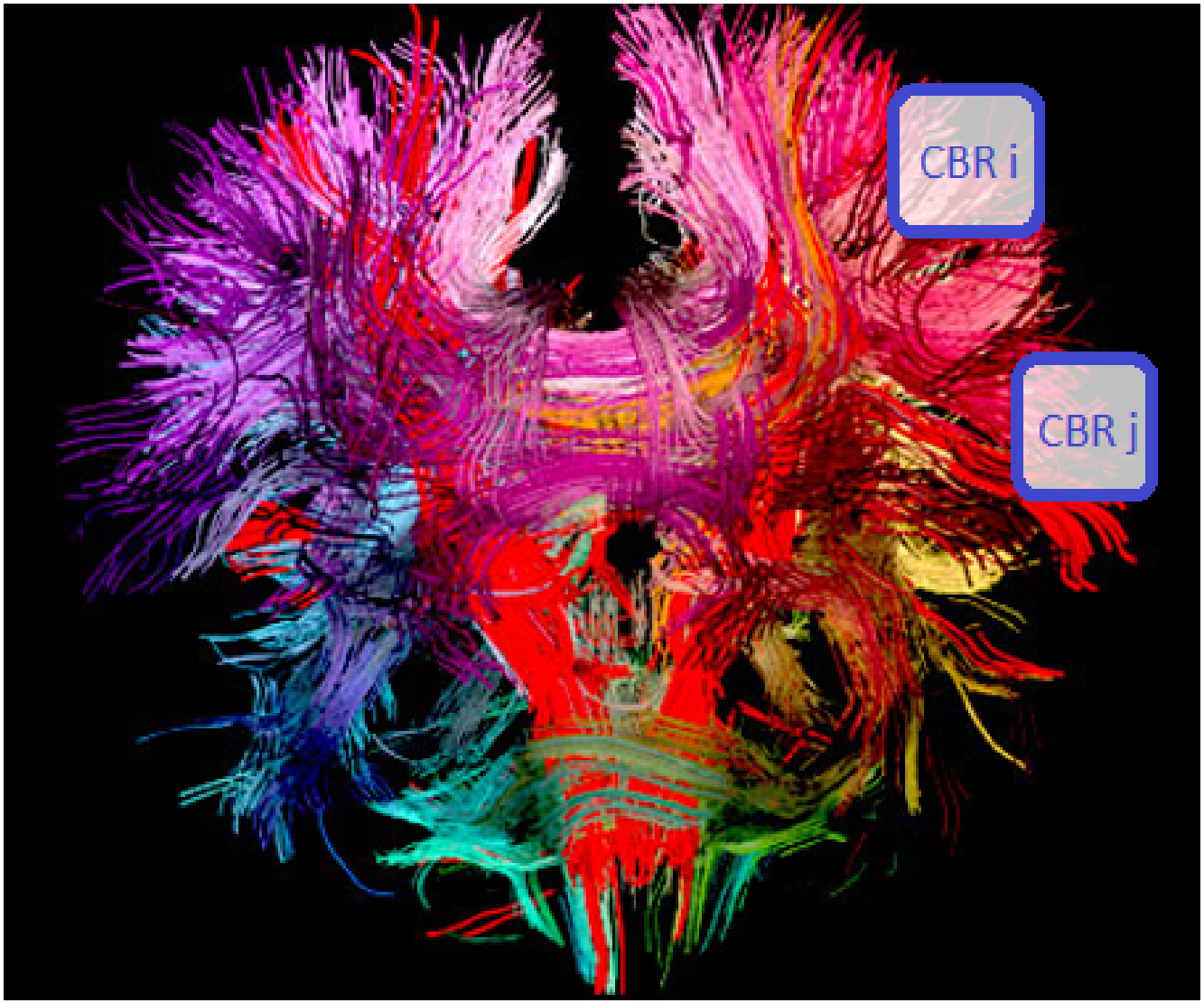


(a) Parcellation of the brain

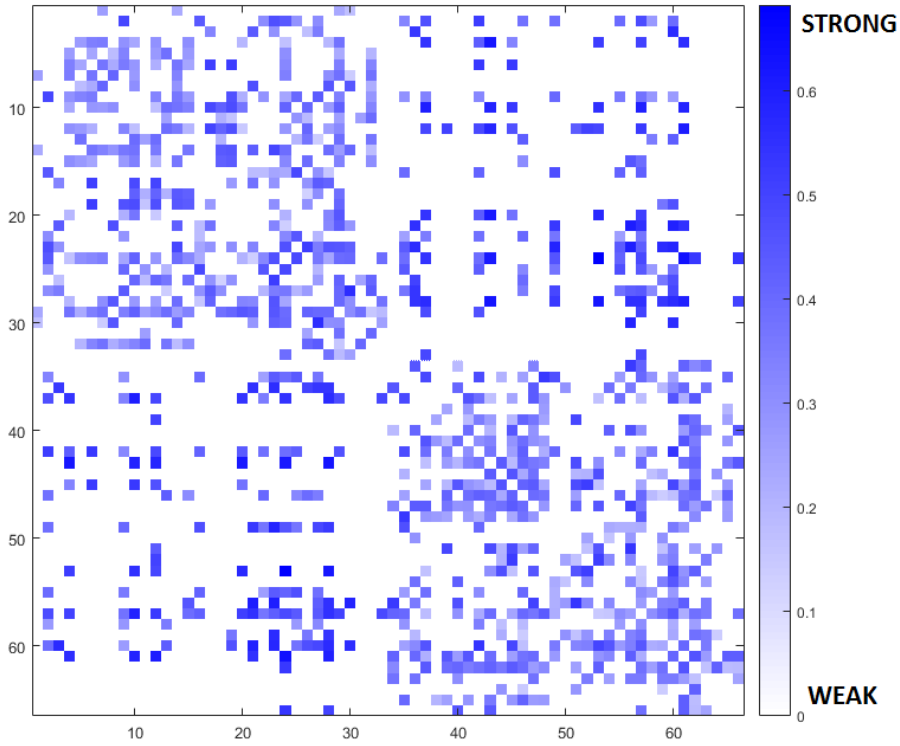


(b) Cortical thickness

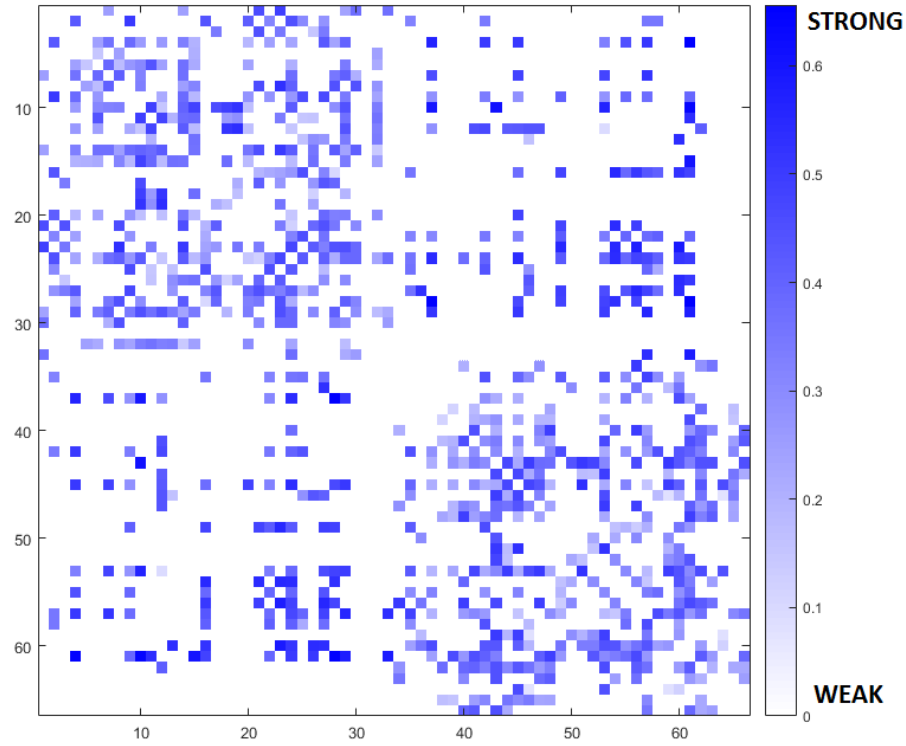
Connections in the brain



Connectivity matrices



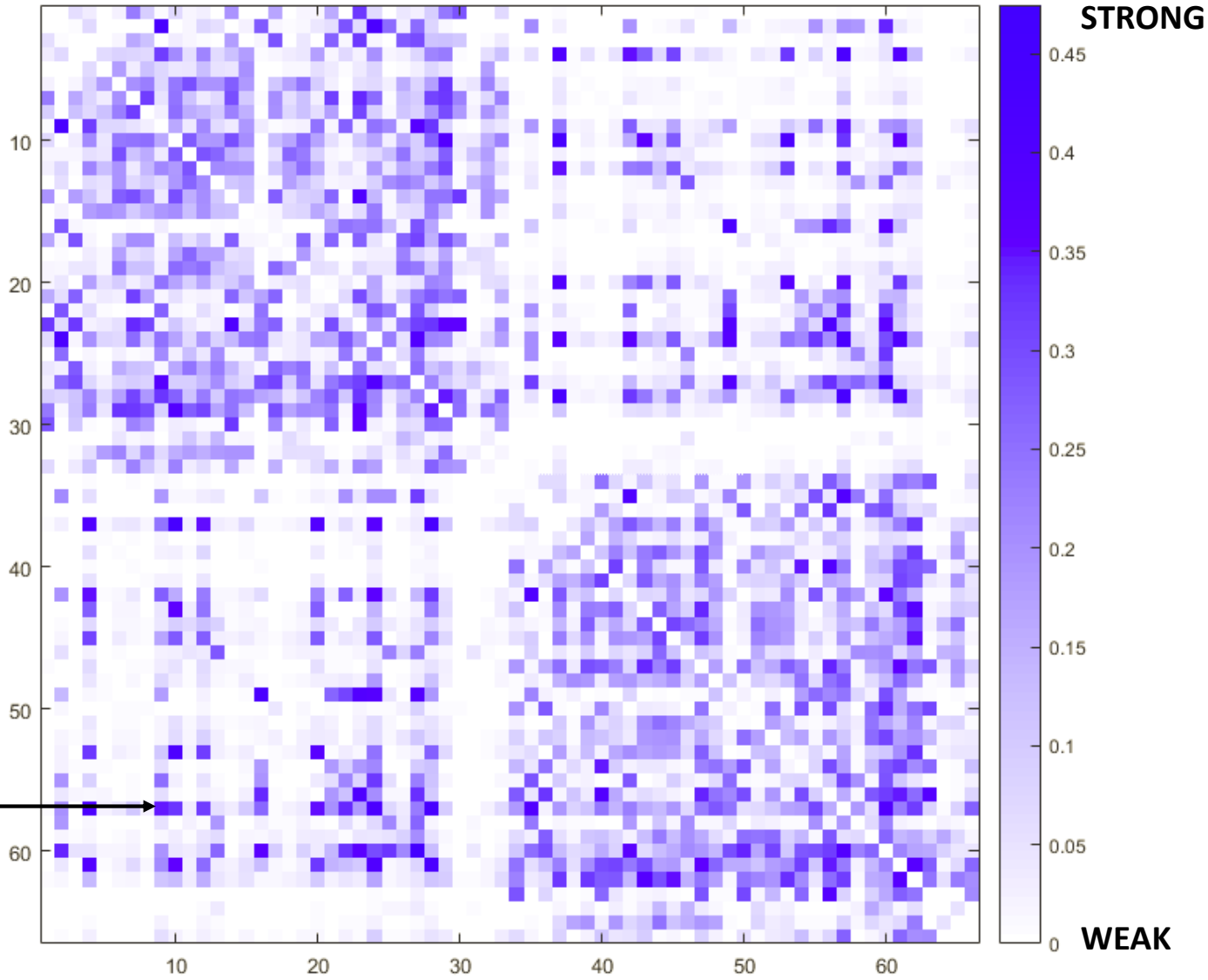
(a) Connectivity matrix: subject 1



(b) Connectivity matrix: subject 2

Population connectivity matrix

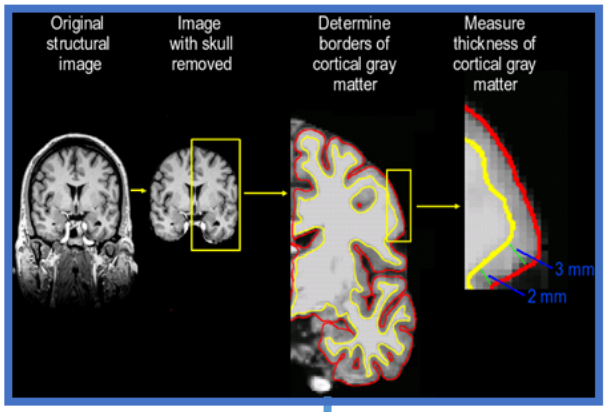
$$A = a_{ij}$$

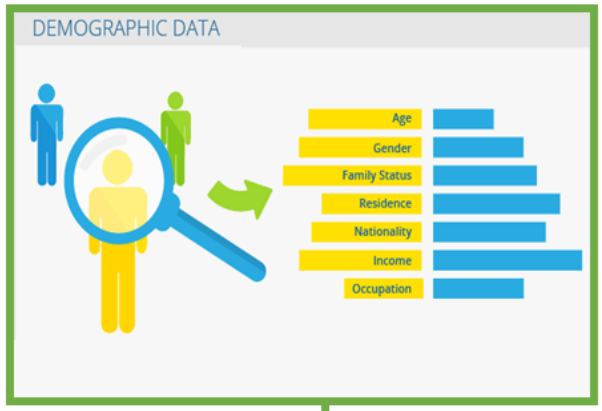


Questions

1. Scientific
   - Are changes in the brain structure associated with the HIV infection?
   - Is there any additional information provided by the structural connectivity?
2. Statistical
   - How to deal with the highly correlated predictors in the regression models?
   - How to incorporate the structural connectivity information in the regression models?

Statistical model

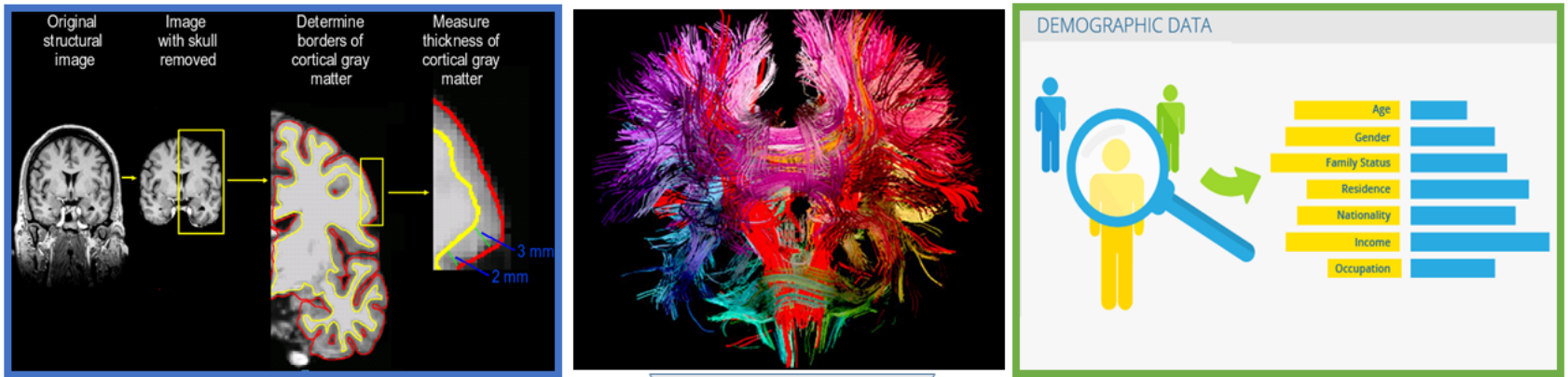




$$y = Zb + X\beta + \varepsilon$$

1. y – n-dimensional response (e.g. NP domain score)
2. $Z \in \mathbb{R}^{n \times 66}$ and $X \in \mathbb{R}^{n \times m}$
3. $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ for some unknown $\sigma^2 > 0$

Statistical model



$$y = Zb + X\beta + \varepsilon$$

1. y – n-dimensional response (e.g. NP domain score)
2. $Z \in \mathbb{R}^{n \times 66}$ and $X \in \mathbb{R}^{n \times m}$
3. $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ for some unknown $\sigma^2 > 0$

Penalized estimation

To find the estimates of b and β, we consider the optimization problem of the form

$$\underset{b,\beta}{\operatorname{argmin}} \left\{ \underbrace{\|y - Zb - X\beta\|_2^2}_{\text{model fit term}} + \lambda \underbrace{g(b)}_{\text{penalty on } b} \right\}.$$

1. $g(b) = \sum_i b_i^2 \longrightarrow$ Ridge estimate
2. $g(b) = \sum_i |b_i| \longrightarrow$ LASSO estimate
3. $g(b) = ||Lb||_2^2 \longrightarrow$ Generalized ridge

T. W. Randolph, J. Harezlak, Z. Feng, Structured penalties for functional linear models – partially empirical eigenvectors for regression, Electronic Journal of Statistics (2012)

Desired property of the estimate, $\hat{b}$

"Stronger connections between the brain regions i and j result in more similar coefficients $\hat{b}_i$ and $\hat{b}_j$."

Penalty selection

The natural choice of the penalty is

$$g(b) = \sum_{i,j} a_{ij}\left(b_i - b_j\right)^2.$$

1. $d_i := \sum_k A_{ik}$.
2. $D := diag\left(d_1, \ldots, d_{66}\right)$
3. $Q := D - A$ **[Laplacian of A]**

Then:

$$\sum_{i,j} a_{ij}\left(b_i - b_j\right)^2 = b^T Q b.$$

Connections with the linear mixed models (LMM)

Our objective function becomes

$$\underset{b,\beta}{\operatorname{argmin}} \left\{ \left\| y - Zb - X\beta \right\|_2^2 + \lambda b^T Q b \right\}.$$

This optimization problem is "equivalent" to the LMM formulation

1. $y = Zb + X\beta + \varepsilon$, where β is a vector of fixed effects and b a vector of random effects,
2. $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I\right)$,
3. $b \sim \mathcal{N}\left(0, \ \sigma_b^2 Q^{-1}\right)$,
4. λ, σ and σ_b are connected via $\lambda = \sigma^2 / \sigma_b^2$.

Selection of the regularization parameter

IDEA: define $\hat{\lambda}$ as $\hat{\lambda} = \frac{\hat{\sigma}^2}{\hat{\sigma}_b^2}$ where $\hat{\sigma}^2$ and $\hat{\sigma}_b^2$ are maximum likelihood estimates from the corresponding linear mixed model

PROBLEM: Laplacian is not an invertible matrix, which is required in computation

The method

riPEER (ridgified Partially Empirical Eigenvectors for Regression)

$$\begin{bmatrix} \hat{b}^{\text{rP}} \\ \hat{\beta}^{\text{rP}} \end{bmatrix} := \underset{b,\beta}{\operatorname{argmin}} \left\{ \left\| y - Zb - X\beta \right\|_2^2 + \underbrace{\lambda_Q b^T Q b}_{\text{graph part}} + \underbrace{\lambda_R \|b\|_2^2}_{\text{ridge part}} \right\}$$

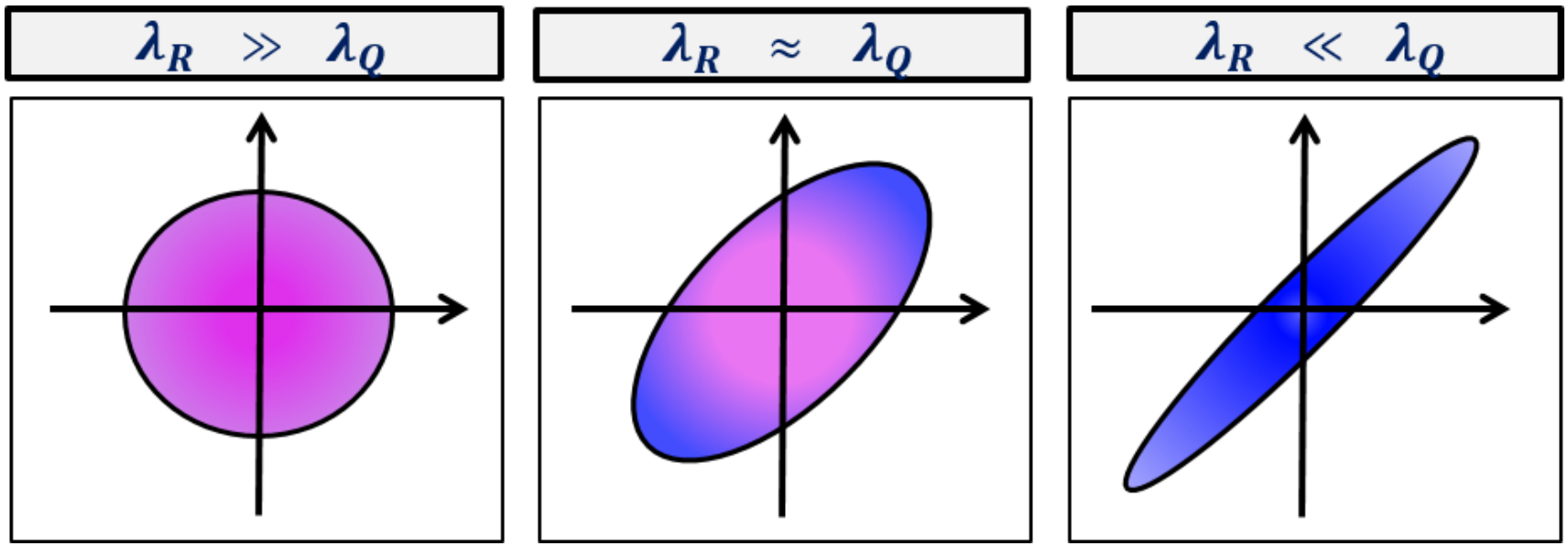


Figure 3: Different shapes of the set $\left\{ b : \lambda_Q b^T Q b + \lambda_R \|b\|_2^2 \leq 1 \right\}$ for $p = 2$.

Connections with the linear mixed models (LMM)

riPEER (ridgified Partially Empirical Eigenvectors for Regression)

$$\begin{bmatrix} \hat{b}^{\mathrm{rP}} \\ \hat{\beta}^{\mathrm{rP}} \end{bmatrix} := \underset{b,\beta}{\operatorname{argmin}} \left\{ \left\| y - Zb - X\beta \right\|_2^2 + b^T \left(\lambda_Q Q + \lambda_R I \right) b \right\}$$

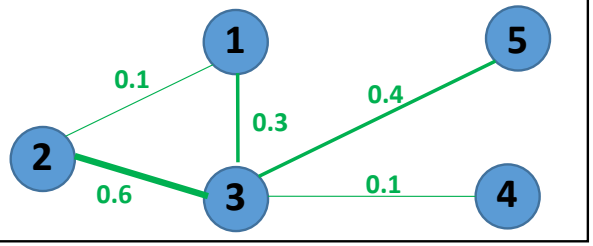
This problem is "equivalent" to the LMM formulation

1. $y = Zb + X\beta + \varepsilon$, where β is a vector of fixed effects and b a vector of random effects,
2. $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I\right)$,
3. $b \sim \mathcal{N}\left(0, \ \left(\sigma_Q^2 Q + \sigma_R^2 I\right)^{-1}\right)$,
4. λ_Q λ_R, σ, σ_Q and σ_R are connected via
$$\lambda_Q = \sigma^2/\sigma_Q^2, \ \ \lambda_R = \sigma^2/\sigma_R^2.$$

Simulation scheme

SIMULATED SIGNAL

Graph given by adjacency matrix A



Laplacian : Q_{true}

„Invertible Laplacian" :

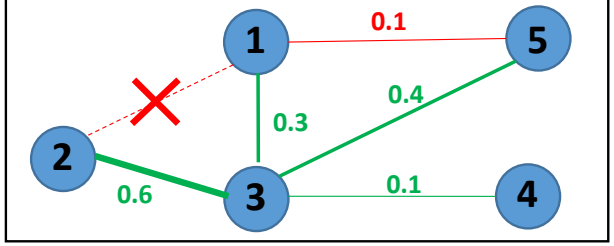
$$\tilde{Q}_{true} := Q_{true} + 0.001 \cdot I$$

True signal used in simulation:

$$b^{true} \sim N(0,\ \sigma_b^2 \tilde{Q}_{true}^{-1})$$

ESTIMATION

Distorted graph



Laplacian of distorted graph was used to find the estimate,

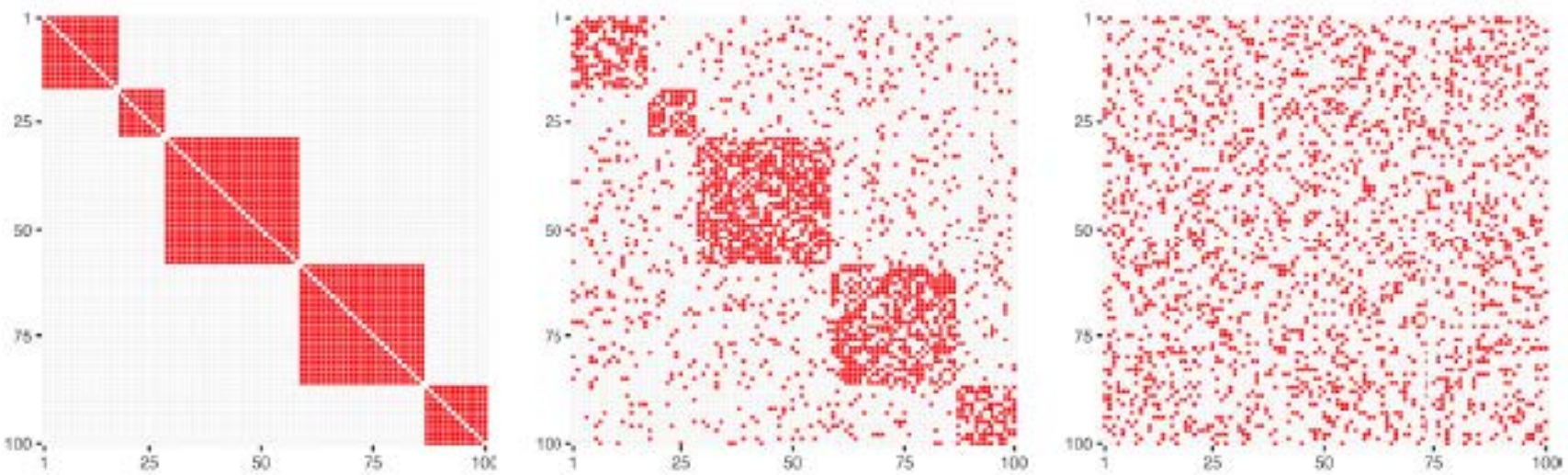
$$\hat{b}$$

MSEr **defined as**

$$\text{MSEr: } = \mathsf{E}\left[\frac{\|\hat{b} - b_{true}\|_2^2}{\|b_{true}\|_2^2}\right]$$

as a measure of estimation accuracy

Simulation scheme – distorted connectivity matrices



Simulation scheme

Three methods compared:

1. ridge: $\lambda_Q := 0$ (connectivity information is not used)
2. naive: $\lambda_R := 0$, $Q \to \widetilde{Q}$ (only λ_Q is selected)
3. riPEER (both lambdas are selected in an adaptive way)

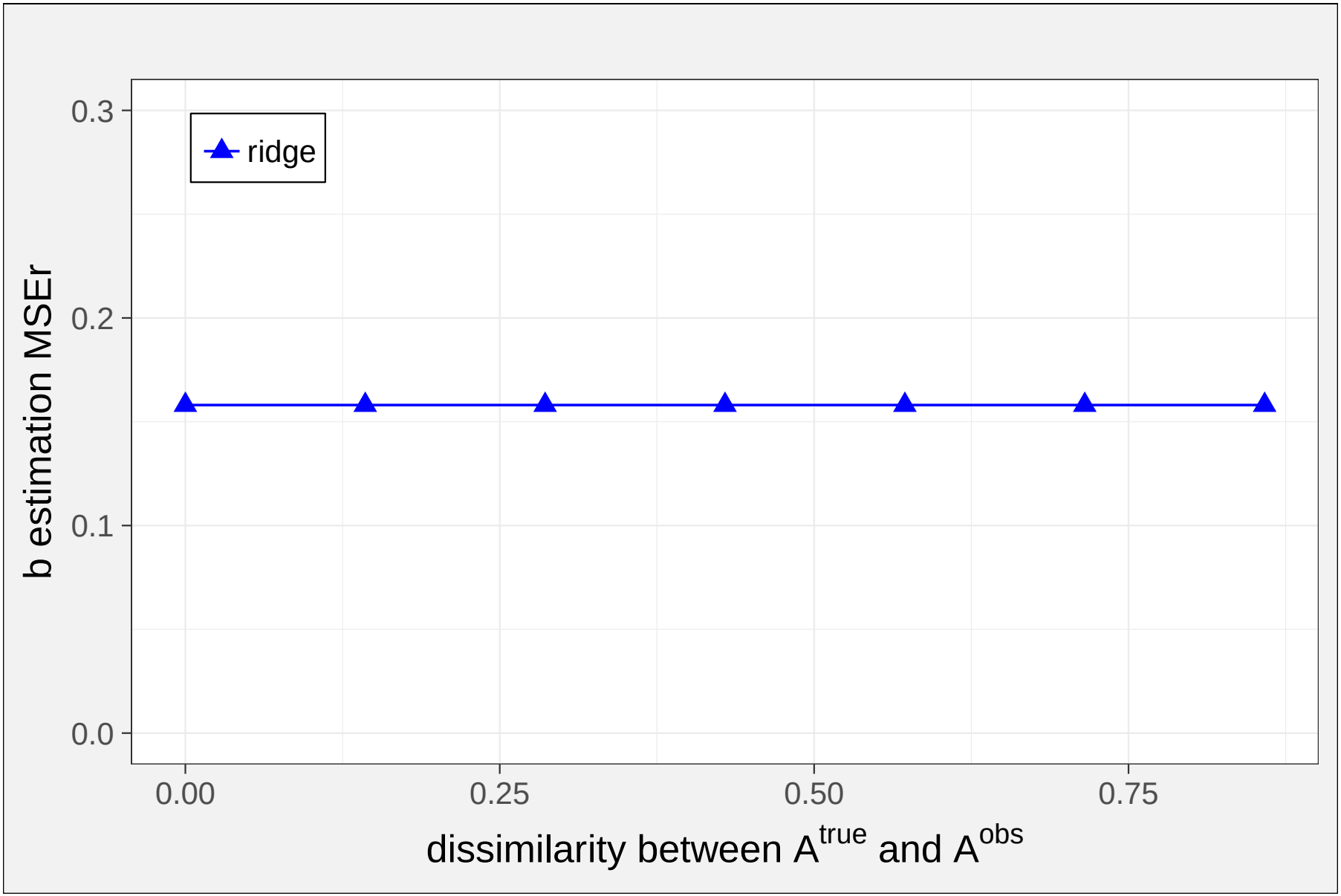
Axis of the plot

1. X axis:

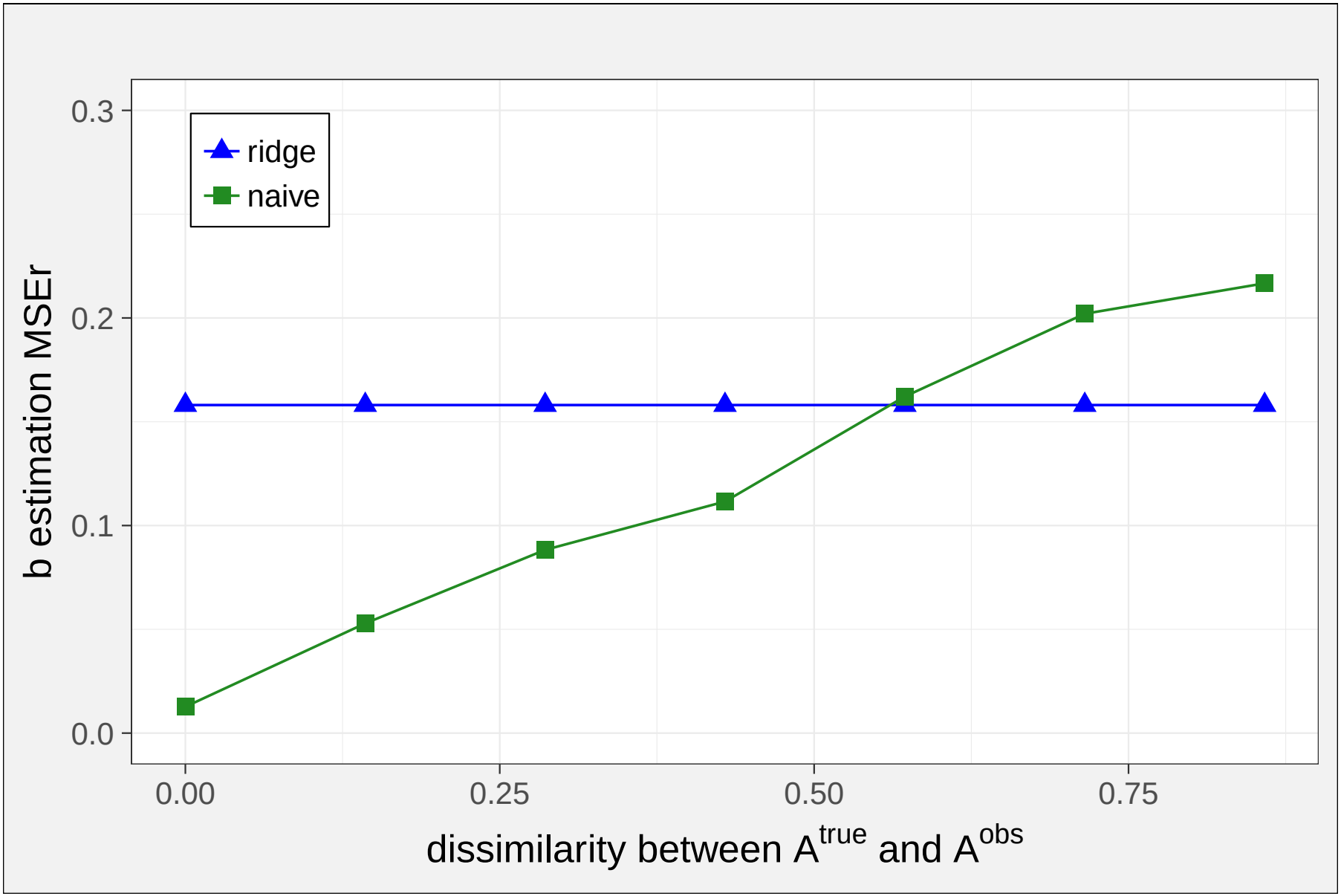
$$diss(A^{true}, A^{obs}) := \frac{\text{number of removed/added connections}}{\text{number of all nonzero connections in } A^{true}}$$

2. Y axis: $MSEr := \mathbb{E}\left[\frac{\|\hat{b} - b^{true}\|_2^2}{\|b^{true}\|_2^2}\right]$.

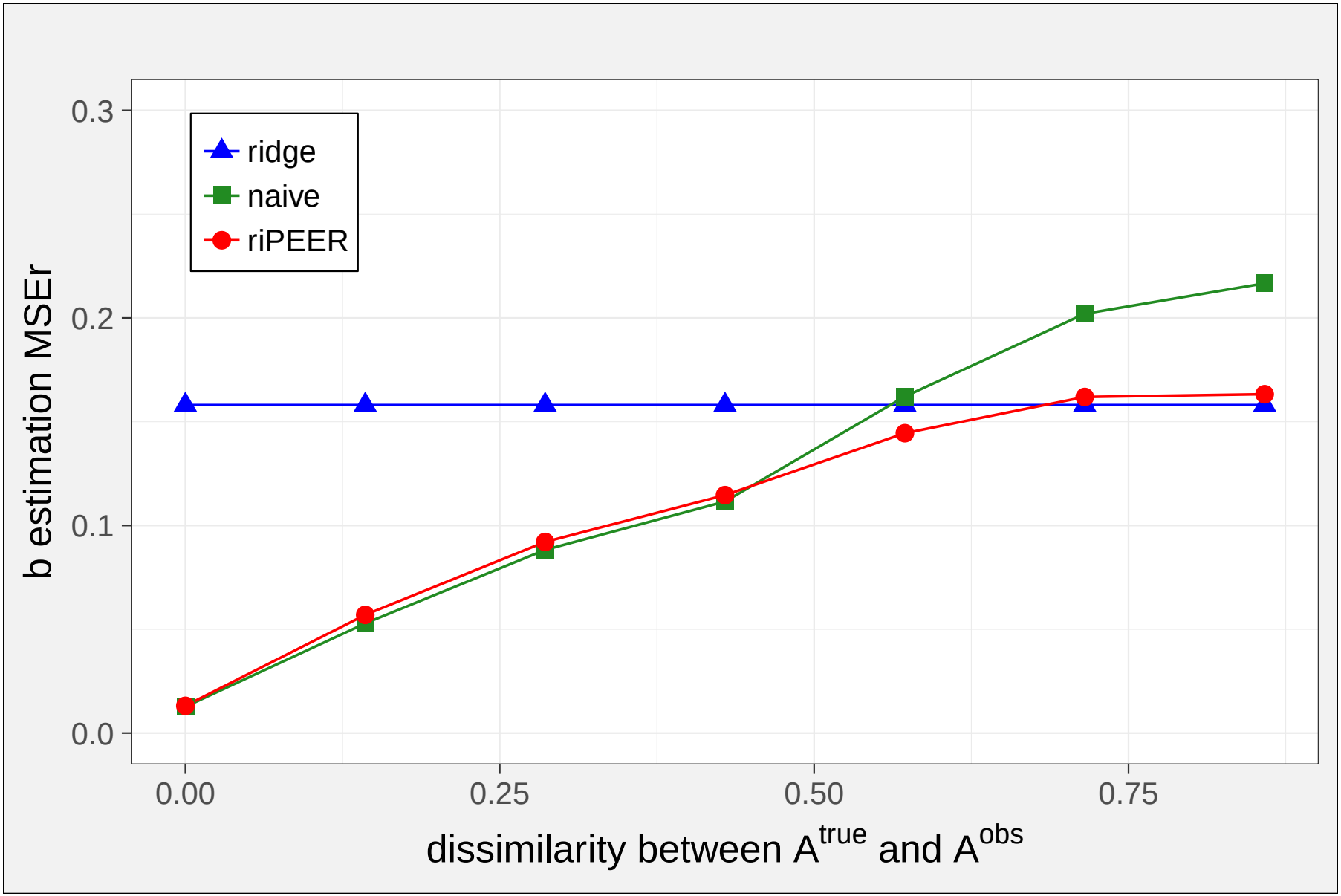
Simulation results



Simulation results



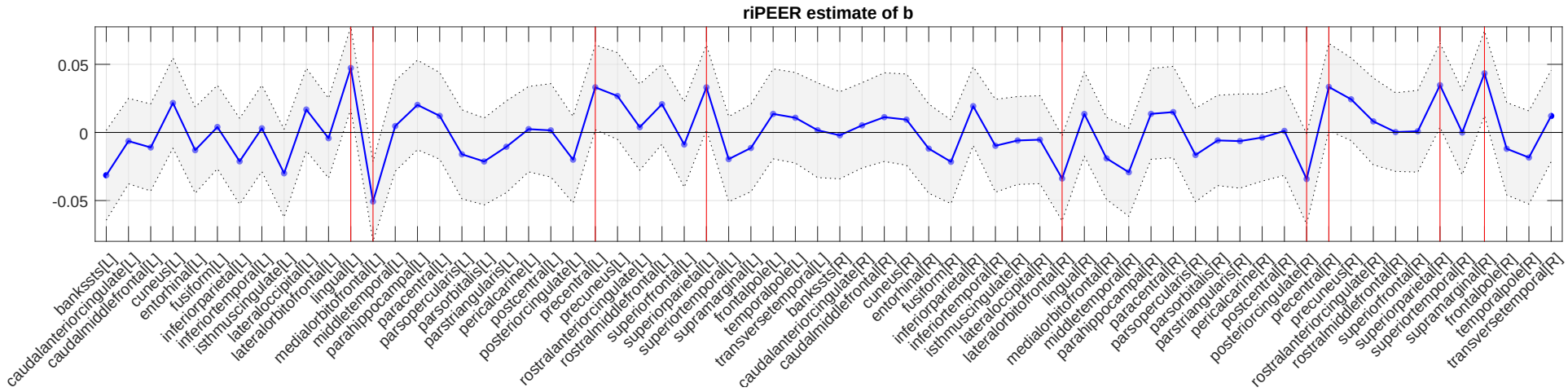
Simulation results



Results: HIV study

1. Association between cortical thickness and speed of information processing
2. 66 considered brain's regions
3. $N = 199$ individuals

Results: Speed of Information Processing



- Cortical regions: lingual[L], precentral[L], superiorparietal[L], lateralorbitofrontal[R], precentral[R], superiorparietal[R], supramarginal[R], medialorbitofrontal[L], posteriorcingulate[R]

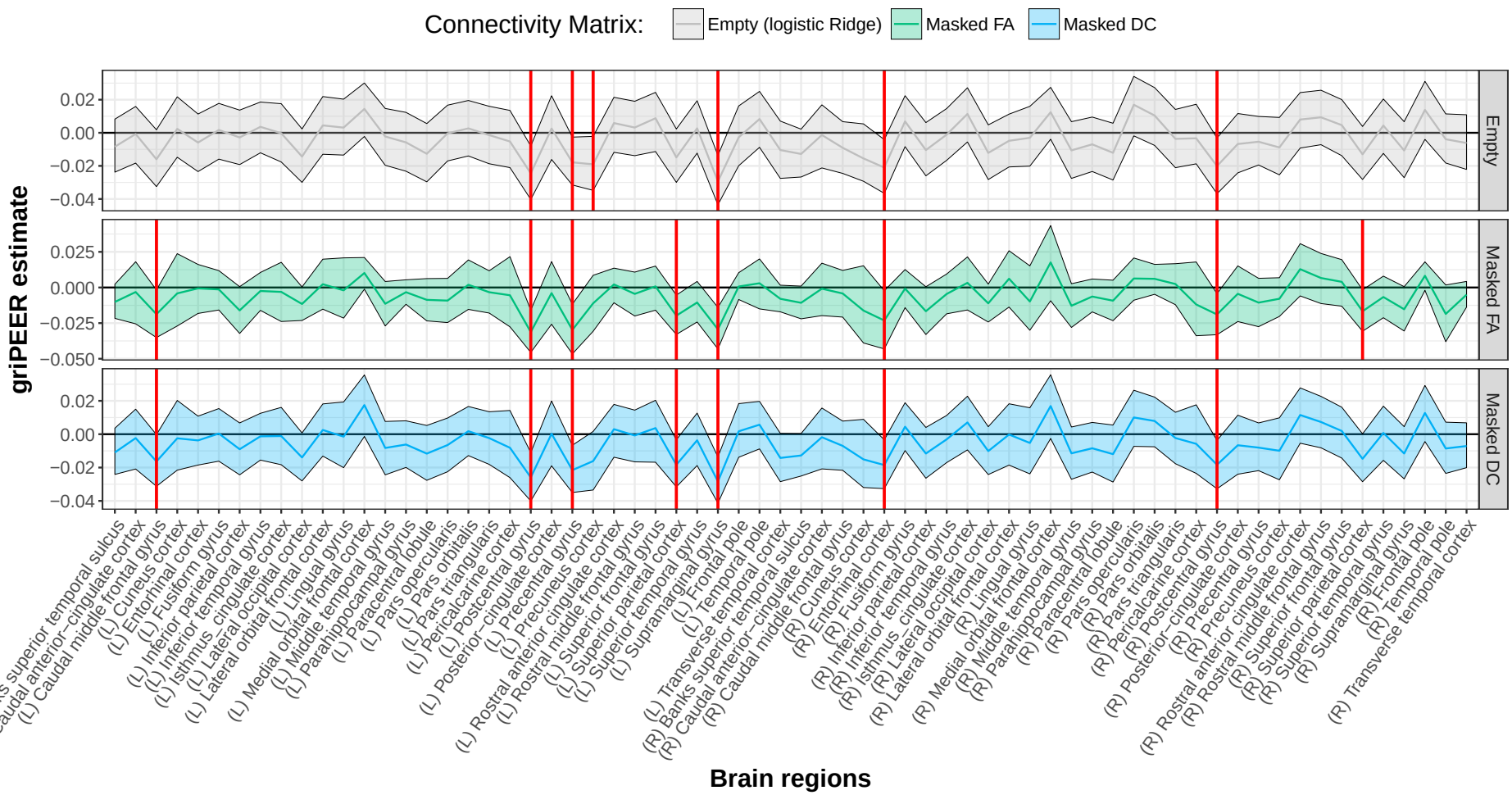
Non-Gaussian distributions

- $y_i \sim$ member of an Exponential family of distribution
- Consider an optimization problem of the form

$$\underset{b,\beta}{\operatorname{argmin}} \left\{ \underbrace{-2\, loglik(y;\, \beta, b)}_{\text{model fit term}} + \underbrace{g_\lambda(b)}_{\text{penalty on } b} \right\}.$$

- $g_\lambda(b) := \lambda_Q b^T Q b + \lambda_R \|b\|_2^2$
- λ_Q and λ_R are selected based on the equivalence with GLMM

HIV data: HIV+ vs. HIV-



Contributions: riPEER and griPEER

1. **Brain connectivity informed regularization methods for regression**, *Statistics in Biosciences*, April 2019, Volume 11, Issue 1, 47–90

   mdpeer R package is available online at https://CRAN.R-project.org/package=mdpeer

2. **Connectivity-Informed Adaptive Regularization Under non-Gaussian Design**, doi: https://doi.org/10.1101/322420

   griPEER Matlab software is available online at https://github.com/dbrzyski/griPEER

Conclusions and Discussion

1. Regularization methods using a priori information in brain imaging setting in a principled way.
2. Additional (external) information quantified via a Laplacian matrix.
3. Penalty parameters chosen in an adaptive way via ML/REML.
4. The more knowledge one has about informative structure, the more specific one can be in defining the penalty matrix.
5. Use of structural connectivity to inform the associations between cortical thickness and either NP outcomes (continuous) or HIV status (binary).

Funding: NIMH R01MH108467

Collaborators

Damian Brzyski
Technical Univ. of Wroclaw, Poland

Marta Karas
Johns Hopkins University

Timothy Randolph
Fred Hutchinson C. R. Center

Mario Dzemidzic
Indiana University

Joaquin Goni
Purdue University

- *Brain Connectivity-Informed Regularization Methods for Regression,* **Statistics in Biosciences (2017),** *https://doi.org/10.1007/s12561-017-9208-x*
- **mdpeer R package** is available online at *https://CRAN.R-project.org/package=mdpeer*

Beau Ances
Washington University

References

1. M. Karas, D. Brzyski, M. Dzemidzic, J. Goni, D. A. Kareken, T. W. Randolph, J. Harezlak, *Brain connectivity–informed regularization methods for regression*, Stat Biosci (2019). https://doi.org/10.1007/s12561-017-9208-x.
2. T. W. Randolph, J. Harezlak, Z. Feng, *Structured penalties for functional linear models – partially empirical eigenvectors for regression*, Electronic Journal of Statistics **6** (2012), 323–353.
3. C. Li, H. Li, *Network–constrained Regularization and Variable Selection for Analysis of Genomic Data*, JBioinformatics **24** (2008), no. 9, 1175–1182.
4. F. Chung, *Laplacians and the Cheeger Inequality for Directed Graphs*, Annals of Combinatorics **9** (2005), no. 1, 1–19.
5. E. Demidenko, *Mixed Models: Theory and Applications*, Wiley **9** (2004).

Pictures sources:

- FreeSurfer, http://www.opensourceimaging.org/project/freesurfer/
- Athinoula A. Martinos Center, http://www.martinos.org/neurorecovery/technology.htm
- PsyPost, http://www.psypost.org/2014/04/atypical-brain-connectivity-associated-with-autism-spectrum-disorder-24472