Connectivity-Informed Adaptive Regularization for Generalized Outcomes

Damian Brzyski[a], Marta Karas[b], Beau Ances[c], Mario Dzemidzic[d], Joaquin Goni[e], Timothy W Randolph[f], Jaroslaw Harezlak[a]

[a]*Department of Epidemiology and Biostatistics, Indiana University, Bloomington, IN, USA*
[b]*Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA*
[c]*Washington University School of Medicine, St. Louis, MO, USA*
[d]*Indiana University School of Medicine, Indianapolis, IN, USA*
[e]*Purdue University, West Lafayette, IN, USA*
[f]*Fred Hutchinson Cancer Research Center, Seattle, WA, USA*

Abstract

One of the challenging problems in the brain imaging research is a principled incorporation of information from different imaging modalities in association studies. Frequently, data from each modality is analyzed separately using, for instance, dimensionality reduction techniques, which result in a loss of mutual information. We propose a novel regularization method, griPEER (generalized ridgified Partially Empirical Eigenvectors for Regression) to estimate the association between the brain structure features and a scalar outcome within the generalized linear regression framework. griPEER provides a principled approach to use external information from the structural brain connectivity to improve the regression coefficient estimation. Our proposal incorporates a penalty term, derived from the structural connectivity Laplacian matrix, in the penalized generalized linear regression. We address both theoretical and computational issues and show that our method is robust to the incomplete information about the structural brain connectivity. We also provide a significance testing procedure for performing inference on the estimated coefficients in this model. griPEER is evaluated in extensive simulation studies and it is applied in classification of the HIV+ and HIV- individuals.

Key words: Generalized Linear Regression, Penalized regression, Structured penalties, Laplacian matrix, Brain connectivity, Brain structure

1. Introduction

In brain imaging applications researchers often collect multiple data types, but in the majority of cases the analysis is performed separately for each of them. Implicit in the work of Randolph et al. (2012) is a framework for simultaneously utilizing multiple data types. For instance, structural and/or functional connectivity measures may serve as useful prior knowledge regarding the structure of dependencies between

brain regions when used in a linear model that aims to estimate the association of brain region properties (e.g., cortical thickness) with a scaler outcome. Karas et al. (2017) explicitly showed that using correct prior information significantly increases estimation accuracy. The statistical methodology, riPEER, developed by these authors allows for incorporating such predefined structure into a regression model a way that protects against using incorrect information. The derived estimation procedure, however, is limited by the assumption that the response variable is normally distributed. Such design excludes, for instance, a binary response that indicates the presence/absence of a condition such disease or phenotype.

To fill this gap, we developed a variant of riPEER, called *generalized ridgified Partially Empirical Eigenvectors for Regression* (griPEER), which handles the outcomes coming from the exponential family of distributions. In the context of brain imaging analysis, our approach allows the analysis to incorporate information such as that encoded in a structural or functional connectivity matrix. As with its precursor, griPEER is able to use the predefined information in a "soft" way — from full inclusion absence — depending on how well this information is confirmed by the data. To achieve this, griPEER employs a penalized optimization problem with a flexible, parameterized penalty term with parameters chosen in a fully automatic and data-driven manner.

We work with a generalized linear regression model where the ith scalar outcome, y_i, is assumed to be drawn from the exponential family of distribution with the parameter θ_i. We confine ourselves to the canonical link functions only and assume that $\theta = X\beta + Zb$. Here, X denotes a matrix of covariates (such as demographic data) for which the prior information is not used and the columns of Z correspond to variables having structure which is assumed to be at least partially known. In the analysis performed in this article, β includes the intercept and demographic data, while b represents the coefficients for the average thickness of 66 brain regions. These regions are assumed to be linked and this linkage is represented by a connectivity matrix; e.g., this matrix may encode a density of connections or the average Fractional Anisotropy (FA).

There is a wide literature on using structural information in image reconstruction and estimation (see, e.g., Bertero and Boccacci (1998), Engl et al. (2000), Phillips (1962)). In situations when the object of the interest is assumed to be a function belonging to a class of, say, differentiable functions, a differential operator-based penalty may be used to "regularize" or impose smoothness on the estimates (Huang et al., 2008). This may improve the prediction and interpretability and is "efficient and sometimes essential" in situations having many highly correlated predictors (Hastie et al., 1995). When the object of estimation is a vector, the penalties are very often constructed based on ℓ_1 and ℓ_2 norms. Examples include such methods as LASSO (Tibshirani, 1996), adaptive LASSO (Zou, 2006), ridge regression (Tikhonov, 1963) and elastic net (Zou and Hastie, 2005), to name just a few.

There is no the unique answer to the question of how to regularize a particular model and the final construction depends strongly on the context. If, for instance,

it is natural to assume sparsity in the coefficients or that they occur in blocks, then using the ℓ_1 norm to constrain them (as in the LASSO) or constrain the difference of adjacent coefficients (as in the fused LASSO) would be useful (Tibshirani et al., 2005). A more generalized fused lasso using two ℓ_1 norms could also be applied: one constraining on the coefficients and one constraining their pairwise differences (Xin et al., 2016).

When more intricate structure among the variables is expected and when some (possibly imprecise) knowledge of it is available, then less generic penalization schemes are more appropriate (Tibshirani and Taylor, 2011; Slawski et al., 2010). For example, a $p \times p$ adjacency matrix represents known connections, or "edges", between p nodes in a graph. This matrix can be used to inform a model that aims to estimate the relationship between an outcome an a vector of p values at the nodes in the graph. More specifically, the adjacency matrix is used to define the graph Laplacian matrix which represents differences between nodes (Chung, 2005), and may be used to penalize the process of estimating regression coefficients, b.

For any $p \times p$ matrix Q, defining a penalty of the form $\lambda b^\mathsf{T} Q b$, where λ is a non-negative regularization parameter constitutes the essence of the methods of Li and Li (2008) and Karas et al. (2017). Using a penalty of this form also serves to link the optimization problem with theory of mixed effects models in which b is assumed to be a random effects vector with distribution $\mathcal{N}(0, \sigma_b^2 Q^{-1})$, for some $\sigma_b^2 > 0$. This, in turn, reveals a connection with the Bayesian approach, where the distribution is treated as a prior on b; see e.g., Maldonado (2009).

Problems with such an interpretation include the fact that Q may not be invertible, as is the case when Q is defined as Laplacian or normalized Laplacian (Chung, 2005). Second, a single multiplicative parameter, λ, adjusts the trade-off between model fit and penalty terms but it can not change the regularization pattern, i.e., the shape of the set $\{b : \ penalty(b) = const\}$ is preserved. When Q is misspecified (is not informative) this lack of adaptivity may significantly degrade performance to be even worse than a uninformed penalty such as ridge regression or LASSO (Karas et al., 2017).

Both of these issues were considered in (Karas et al., 2017) which does not assume Q is exactly the true signal precision matrix, $\mathbb{Q}$, but is merely "close", in some sense; i.e., Q contains some amount of true information which can be exploited. Therefore, by considering a family of transformations of Q, and selecting the optimal member by applying a data-adaptive procedure, one may obtain a modified matrix which reflects $\mathbb{Q}$ better and improves prediction accuracy. Transformations of the form $Q + a\mathbf{I}_p$ $(a > 0)$ are used by Karas et al. (2017). Any such modification of Q is invertible and could be directly used in the estimation procedure. The resulting penalty term, $\lambda b^\mathsf{T}(Q + a\mathbf{I}_p)b$, has an equivalent form $\lambda_Q b^\mathsf{T} Q b + \lambda_R \|b\|_2^2$ and the connection with a specific linear mixed model enables the optimal selection of λ_Q and λ_R.

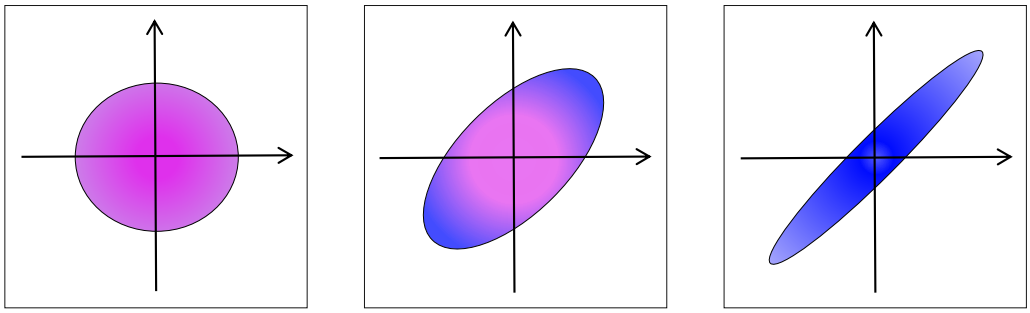
The approach by Karas et al. (2017) assumes the response variable is normally distributed and hence not suitable for categorical outcomes. In this presentation we extenbd the concept of riPEER's penalty function to the case when the distri-

bution of the response variable is a member of one-parameter exponential family of distributions. The proposed estimation method, griPEER, is of the form:

$$\begin{bmatrix} \hat{\beta}^{\mathrm{gP}} \\ \hat{b}^{\mathrm{gP}} \end{bmatrix} := \underset{\beta,b}{\operatorname{argmin}} \Big\{ -2\, loglik\big(\beta, b|y\big) \;+\; \lambda_Q b^T Q b \;+\; \lambda_R \|b\|_2^2 \Big\}, \tag{1}$$

where $loglik\big(\beta, b|y\big)$ is a log-likelihood. Here, the term $-2loglik\big(\beta, b|y\big)$ is used to fit the model to the response distribution while the parameters λ_Q and λ_R are chosen based on the connection between the optimization problem and the generalized linear mixed model; this is formulated explicitly in Section 2. It is important to emphasize that these parameters not only determine the trade-off between the model fit and the penalty term, but also on the form of the penalty, which determines the structure that the estimate is encouraged to have. More precisely, if λ_Q is large relative to λ_R, then the connectivity information has a large role in the estimation process. Conversely, when λ_Q is small relative to λ_R, the penalty is equal in all coordinates, as with ridge regression.

We illustrate this using a simple example with $p = 2$ variables and prior information that implies these variables are connected. Figure 1 shows how the shapes of contour sets of penalty, which decide on the solution structure, change for various lambdas. If the relationship between variables, as represented in Q, is reflected in data and if this is related to the outcome y, then griPEER will tend to choose relatively large λ_Q, which links the coefficients in b (see right plot in Figure 1). The other extreme is when the structure in Q is not informative for the relationship between y and Z. In this case, griPEER will select a relatively large λ_R inducing a ridge-like penalty that ignores Q (see the left panel in Figure 1).



(a) $\lambda_Q \ll \lambda_R$ (b) $\lambda_Q \approx \lambda_R$ (c) $\lambda_Q \gg \lambda_R$

Figure 1: Shapes of the set $\{b:\ \lambda_Q b^{\mathsf{T}} Q b + \lambda_R \|b\|_2^2 = 1\}$ for various pairs of regularization parameters: (a) the assumed strong connections between variables was neglected, (b) the moderate tendency for coefficients of the solution to be similar to each other (c) strong tendency for coefficients of the solution to be similar to each other

The remainder of this work is organized as follows. In Section 2, we formulate our statistical model, investigate the special case of binomial distribution and discuss the equivalence between GLMM and penalized optimization problems. We also describe

the penalty term construction from the graph-theory point of view. The estimation procedure we use to select the optimal regularization parameters is introduced in Section 3, while the next Section addresses the problem of the selection of response-relevant variables. The extensive simulations showing very good performance of griPEER in the context of estimation accuracy and variables selection (under various scenarios illustrating the impact of inaccurate prior information) are reported in Section 5. Finally, in Section 6, we apply our methodology to study the association of the brain's cortical thickness and HIV disease. The conclusions and a discussion are summarized in Section 7.

2. Statistical model

We address the problem of estimation in a penalized generalized linear model where the penalty term is derived from connectivity information. This information is represented by a $p \times p$ symmetric matrix having non-negative entries with zeros on the diagonal. This *adjacency matrix* or *connectivity matrix* and will be denoted by $\mathcal{A}$. The corresponding graph Laplacian matrix, Q, which defines the penalty term is defined next, followed by specific details about the statistical model in (1).

2.1. The graph Laplacian, Q

We are interested in modeling the association between a scaler outcome, y, and a set of p predictor variables that are measured at the nodes of graph. We assume that information about connections between the these variables — i.e., strengths of the connections between the nodes — can be summarized by a (symmetric) $p \times p$ adjacency matrix $\mathcal{A} = [a_{ij}]$, $1 \leq i, j \leq p$, having non-negative entries and zeros on the diagonal. We denote the *degree* of the jth node as $d_j := \sum_j a_{jj}$, and define the degree matrix as $D := diag(d_1, \ldots, d_p)$.

Following Chung (2005), we define the unnormalized Laplacian, Q_u, corresponding to $\mathcal{A}$ simply as $Q_u := D - \mathcal{A}$. This matrix is always positive semidefinite. It is also singular, since for the vector of ones, $\mathbf{1} := [1, \ldots, 1]^\mathsf{T}$ we have $\mathbf{1}^\mathsf{T} Q_u \mathbf{1} = \mathbf{1}^\mathsf{T} D \mathbf{1} - \mathbf{1}^\mathsf{T} \mathcal{A} \mathbf{1} = \sum_i d_i - \sum_i d_i = 0$.

Intuition on the role of a penalty of the form $b^T Q_u b$, as in (1), is gained by the following simple formula: for any adjacency matrix, $\mathcal{A}$, and its unnormalized Laplacian, Q_u, then

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

That is, the term $b^T Q_u b$ in the optimization problem (1) penalizes the squared differences of coefficients in a manner that is proportional to the strengths of connections between them. Consequently, coefficients corresponding to nodes having many strong connections (nodes with large degree) are constrained more than others. T

In order to allow a small number of nodes with large d_i to have more extreme values, we employ the normalized Laplacian, Q, which is obtained by by dividing

each column and row of Q_u by a square root of corresponding node's degree. As a result, the property (2), with Q instead of Q_u, takes the form

$$b^T Q b = \sum_{i,j:a_{ij}\neq 0} a_{ij} \left(\frac{b_i}{\sqrt{d_i}} - \frac{b_j}{\sqrt{d_j}} \right)^2 .$$

Q has ones one the diagonal and, as with the unnormalized Laplacian, it is a symmetric, positive semidefinite and singular matrix.

2.2. Statistical model in general form

Consider the general setting where y is an $n \times 1$ vector of observations, and the design matrices, X and Z, are $n \times p$ and $n \times m$ matrices, respectively. The columns of X represent the p covariates and the rows are denoted by X_i. Similarly, the columns of Z correspond to m variables, or nodes in a graph, for which some connectivity information may be available; the rows are denoted by Z_i. We assume there exists(unknown) vectors b and β such that, for each $i \in \{1, \ldots, n\}$, y_i is the member of one-parameter exponential family of distributions of the form

$$f(y_i) = \exp\big\{y_i\theta_i - \psi(\theta_i) + c(y_i, \varphi)\big\}, \tag{3}$$

where $\theta_i := X_i\beta + Z_i b$ is a subject-specific parameter. The formula in 3 includes exponential, binomial, Poisson and Laplace densities.

It can be shown that for the exponential family of distributions, the mean of y_i is simply given by the first derivative of ψ in the point θ_i, while the variance could be expressed as the second derivative of ψ, i. e.

$$\mathbb{E}\big(y_i\big) = \psi'(\theta_i), \qquad \text{var}\,\big(y_i\big) = \psi''(\theta_i). \tag{4}$$

Moreover, the log-likelihood function is

$$loglik_{\psi,c}\big(\beta, b|\, y\big) = \sum_{i=1}^{n} \big\{y_i(X_i\beta + Z_i b) - \psi(X_i\beta + Z_i b) + c(y_i)\big\} \tag{5}$$

and it provides a core for the methodology we propose in this presentation. Indeed, we define griPEER as a solution to the following optimization problem

$$\begin{bmatrix} \hat{\beta}^{\text{gP}} \\ \hat{b}^{\text{gP}} \end{bmatrix} := \underset{\beta,b}{\text{argmin}} \Big\{ -2\, l_\psi\big(\beta, b|\, y\big) \;+\; \lambda_Q b^{\mathsf{T}} Q b \;+\; \lambda_R \|b\|_2^2 \Big\}, \tag{6}$$

where $l_\psi\big(\beta, b|\, y\big) := \sum_{i=1}^{n} \big\{y_i(X_i\beta + Z_i b) - \psi(X_i\beta + Z_i b)\big\}$ consists of the terms of log-likelihood function (5) depending ond b and β. Here, λ_Q and λ_R are regularization parameters, which are selected automatically, as described in Section 3.

2.3. The special case – binomial distribution

To provide focus to our presentation we will concentrate on the setting of a binomial outcome in all simulations (Section 5) and the responses in the applications (Section 6) are modeled by the binomial distribution. So in this subsection we explicitly describe this special choice of density function.

As in the classical logistic regression theory, we assume that the response, y_i, takes the value 1 with probability $e^{\theta_i}/(e^{\theta_i}+1)$ and 0 with the probability $1/(e^{\theta_i}+1)$. Consequently, the density function, $f(y_i)$, is given by

$$f(y_i) = \exp\Big\{y_i\theta_i - \ln(1+e^{\theta_i})\Big\}, \tag{7}$$

which is a member of exponential family of distributions (3) with $\psi(\theta_i) = \ln(1+e^{\theta_i})$ and $c(y_i) = 0$. We also have

$$\begin{cases} \mathbb{E}\big(y_i\big) &= \psi'(\theta_i) &= e^{\theta_i}/(e^{\theta_i}+1) \\ \text{var}\,\big(y_i\big) &= \psi''(\theta_i) &= e^{\theta_i}/(e^{\theta_i}+1)^2 \end{cases} . \tag{8}$$

From this, $\theta_i = \ln\left(\frac{\mathbb{E}(y_i)}{1-\mathbb{E}(y_i)}\right)$ which, with the assumption $\theta = X\beta + Zb$ adopted in the manuscript, yields the canonical link for logistic regression—the logit function.

2.4. Equivalence between GLMM and two optimization problems

The optimization problem in (6) is strongly connected with the specific GLMM formulation. Indeed, consider the model defined by the following conditions

- **A.1** β is a vector of fixed and b is a vector of random effects,
- **A.2** $y_i|b$ are independent and, consequently, $f(y|\,b) = \prod_{i=1}^n f\big(y_i\,|b\big)$,
- **A.3** $f\big(y_i\,|b\big) = \exp\Big\{y_i(X_i\beta + Z_ib) - \psi(X_i\beta + Z_ib) + c(y_i)\Big\}$, for some (known) functions ψ, c and $i = 1,\ldots,n$,
- **A.4** $b \sim \mathcal{N}\big(0,\, \widetilde{Q}_\lambda^{-1}\big)$, where $\widetilde{Q}_\lambda := \lambda_Q Q + \lambda_R\,\mathrm{I}_p$ for some unknown, positive parameters λ_Q and λ_R.

To see this correspondence, assume the parameters λ_Q and λ_R have been estimated, say as $\widehat{\lambda} := [\widehat{\lambda}_Q, \widehat{\lambda}_R]^\mathsf{T}$, and these values are used to obtain β and b. One can proceed by treating both fixed and random effects as parameters and finding ML estimates by maximizing (with respect to β, b) the density function

$$\begin{aligned} f(y,b) \;&= f(y|\,b)\,f(b) = \prod_{i=1}^n \Big\{f(y_i|\,b)\Big\}f(b) \;\propto \\ &\exp\Big\{\sum_{i=1}^n \Big[y_i\theta_i - \psi(\theta_i)\Big] - \frac{1}{2}b^\mathsf{T}\,\widetilde{Q}_{\widehat{\lambda}}\,b\Big\}, \end{aligned} \tag{9}$$

where $\theta_i = X_i\beta + Z_i b$, for $i = 1, \ldots, n$. Taking the logarithm of the above leads directly to the objective in optimization problem (6).

We now derive a constrained optimization problem that is equivalent to (6) and reveals the role of the regularization parameters on the solution from a slightly different perspective. For this, suppose that $\begin{bmatrix}\hat{\beta}\\ \hat{b}\end{bmatrix}$ is the solution to (6) for given parameters λ_Q and λ_R. Then define $c := \lambda_Q \hat{b}^\mathsf{T} Q \hat{b} + \lambda_R \|\hat{b}\|_2^2 \geq 0$. One can check that $\begin{bmatrix}\hat{\beta}\\ \hat{b}\end{bmatrix}$ also solves the problem

$$\begin{array}{ll} \underset{\beta,b}{\text{argmin}} & \Big\{ -2\, l_\psi\big(\beta, b|\, y\big) \;+\; \lambda_Q b^\mathsf{T} Q b \;+\; \lambda_R \|b\|_2^2 \Big\} \\ \text{subject to} & \lambda_Q b^\mathsf{T} Q b \;+\; \lambda_R \|b\|_2^2 = c. \end{array} \tag{10}$$

The multiplicative factor may be neglected as well as the term $\lambda_Q b^\mathsf{T} Q b \;+\; \lambda_R \|b\|_2^2$, which is constant on the feasible set. This yields

$$\begin{array}{ll} \underset{\beta,b}{\text{argmax}} & l_\psi\big(\beta, b|\, y\big) \\ \text{subject to} & \lambda_Q b^\mathsf{T} Q b \;+\; \lambda_R \|b\|_2^2 = c. \end{array} \tag{11}$$

This formulation clarifies the intuition behind the example in the Introduction and the corresponding Figure 1. I.e., griPEER selects the estimates by taking the maximal likelihood value on a set whose shape is explicitly regularized by the parameters λ_Q and λ_R.

3. A new estimation algorithm

To select the optimal values of λ_Q and λ_R, we employ the corresponding GLMM formulation defined by A.1 – A.4. The likelihood function, $\mathcal{L}\big(\beta, \lambda|y\big)$, is given by

$$\mathcal{L}\big(\beta, \lambda|y\big) \;=\; \int_{\mathbb{R}^p} f_{\beta,\lambda}(y|\, b) f_{\beta,\lambda}(b)\, db \;=$$
$$\int_{\mathbb{R}^p} \left|2\pi \widetilde{Q}_\lambda\right|^{-\frac{1}{2}} \exp\left\{ \sum_{i=1}^{n} \Big[y_i(X_i\beta + Z_i b) - \psi(X_i\beta + Z_i b) - c(y_i) \Big] - \frac{1}{2} b^\mathsf{T}\, \widetilde{Q}_\lambda\, b \right\} db. \tag{12}$$

Unfortunately, obtaining the maximum of $\mathcal{L}$ with respect to β and λ is complicated by the fact that there is no closed-form solution to the multidimensional integral in (12). For this, several approaches have been proposed. Breslow and Clayton (1993) proposed a general method based on Penalised Quasi-Likelihood (PQL) for the estimation of the fixed and prediction of random effects. Wolfinger and O'connell (1993) investigated the pseudo-likelihood (PL) approach which is closely related to the Laplace's approximation of $\mathcal{L}$. Other proposals include the Adaptive Gaussian Quadrature to approximate integrals with respect to a given kernel (Pinheiro and Chao, 2006) and an MCMC-based procedure (Zeger and Karim, 1991).

In this article we focus on the Wolfinger PL approach which is recognized as being fast and computationally efficient. It relies on the first-order Taylor series approximation and uses the Linear Mixed Model (LMM) proxy in the iterative process: at each iteration, the updates of β and b are based on the variance-covariance parameters of random effects. The steps are repeated until convergence.

The procedure we derive here differs from (Wolfinger and O'connell, 1993) in how the updates of β and b are obtained. In contrast to the Wolfinger PL approach, we do not get them via the solution to the mixed-model equations, but instead we employ the correspondence between GLMM and griPEER optimization problem, as described in 2.4. Specfically, the $(k-1)$-step estimates of λ_Q and λ_R (i.e., $\overset{[k-1]}{\lambda}_Q$ and $\overset{[k-1]}{\lambda}_R$) are used to obtain the $(k-1)$-step estimates of β and b ($\overset{[k-1]}{\beta}$ and $\overset{[k-1]}{b}$) via the solution to (6). Consequently, we can define $\overset{[k-1]}{\theta}_i := X_i \overset{[k-1]}{\beta} + Z_i \overset{[k-1]}{b}$.

Details of our estimation procedure are as follows. Using the Taylor approximation of function ψ' at point $\overset{[k-1]}{\theta}_i$ we get

$$\psi'(\theta_i) \approx \psi'(\overset{[k-1]}{\theta}_i) + \psi''(\overset{[k-1]}{\theta}_i) \cdot (\theta_i - \overset{[k-1]}{\theta}_i) \tag{13}$$

and therefore from (4)

$$\left[\psi''(\overset{[k-1]}{\theta}_i)\right]^{-1} \cdot \left(\mathbb{E}\big(y_i|\beta, b\big) - \psi'(\overset{[k-1]}{\theta}_i)\right) + \overset{[k-1]}{\theta}_i \approx \theta_i. \tag{14}$$

We now define a random variable $\overset{[k]}{y}_i := \left[\psi''(\overset{[k-1]}{\theta}_i)\right]^{-1} \cdot \left(y_i - \psi'(\overset{[k-1]}{\theta}_i)\right) + \overset{[k-1]}{\theta}_i$. The main step now is the assumption that the distribution of $\overset{[k]}{y}_i$ can be well approximated by a normal density. Computation of mean and variance of $\overset{[k]}{y}_i$ immediately yields

$$\mathbb{E}\big(\overset{[k]}{y}_i|b\big) \approx \theta_i = X_i\beta + Z_i b, \quad \text{and} \quad \operatorname{var}\big(\overset{[k]}{y}_i|b\big) = \left[\psi''(\overset{[k-1]}{\theta}_i)\right]^{-2}\psi''(\theta_i) \approx \left[\psi''(\overset{[k-1]}{\theta}_i)\right]^{-1}. \tag{15}$$

The assumption that $\overset{[k]}{y}_i$ is approximately normally distributed allows for replacing the GLMM formulation in kth step by an LMM of the form

B.1 β is a vector of fixed and b is a vector of random effects,

B.2 $\overset{[k]}{y} = [X\ Z]\begin{bmatrix}\beta \\ b\end{bmatrix} + \overset{[k]}{\varepsilon}$,

B.3 $\overset{[k]}{\varepsilon} \sim \mathcal{N}\big(0, \overset{[k]}{W}\big)$, where $\overset{[k]}{W} := \operatorname{diag}\left(\left[\psi''(\overset{[k-1]}{\theta}_1)\right]^{-1}, \ldots, \left[\psi''(\overset{[k-1]}{\theta}_n)\right]^{-1}\right)$,

B.4 $b \sim \mathcal{N}\big(0, \widetilde{Q}_\lambda^{-1}\big)$, where $\widetilde{Q}_\lambda$ was defined in (A.4).

Denote by $\overset{[k]}{\mathcal{P}} := \mathrm{I} - X(X^\mathsf{T}\overset{[k]}{W}{}^{-1}X)^{-1}X^\mathsf{T}\overset{[k]}{W}{}^{-1}$ the $\overset{[k]}{W}$-weighted projection onto the orthogonal complement of the columns of X. Now, defining $\overset{[k]}{\widetilde{y}} := \overset{[k]}{\mathcal{P}}\overset{[k]}{y}$, $\overset{[k]}{X} := \overset{[k]}{\mathcal{P}}X$ and

$\overset{[k]}{Z} := \overset{[k]}{\mathcal{P}} Z$ we assume that

$$\overset{[k]}{\widetilde{y}} \sim \mathcal{N}\big(0,\ \overset{[k]}{V}_\lambda\big), \quad \text{for} \quad \overset{[k]}{V}_\lambda := \overset{[k]}{Z}\, \widetilde{Q}_\lambda^{-1}\, \overset{[k]}{Z}^{\mathsf{T}} + \overset{[k]}{W}. \tag{16}$$

Maximizing the log-likelihood for $\overset{[k]}{\widetilde{y}}$, i.e. the function $l(\overset{[k]}{\widetilde{y}}; \lambda) := -\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|\overset{[k]}{V}_\lambda| - \frac{1}{2}\overset{[k]}{\widetilde{y}}^{\mathsf{T}} V_\lambda^{-1} \overset{[k]}{\widetilde{y}}$, leads directly to the optimization problem

$$\begin{bmatrix} \overset{[k]}{\lambda}_Q \\ \overset{[k]}{\lambda}_R \end{bmatrix} := \underset{\lambda \succeq 0}{\operatorname{argmin}} \Big\{ \ln|\overset{[k]}{V}_\lambda| \ + \ \overset{[k]}{\widetilde{y}}^{\mathsf{T}} V_\lambda^{-1} \overset{[k]}{\widetilde{y}} \Big\}, \tag{17}$$

where $\lambda \succeq 0$ refers to $\big\{(\lambda_Q, \lambda_R) :\ \lambda_Q \geq 0,\ \lambda_R \geq 0\big\}$. The following proposition helps us to rewrite the objective of (17). A proof is provided in the Appendix.

Proposition 3.1. *Let* $\overset{[k]}{\Omega} := \overset{[k]}{Z}^{\mathsf{T}}\ \overset{[k]}{W}^{-1}\ \overset{[k]}{Z}$ *and* $\overset{[k]}{q} := \overset{[k]}{Z}^{\mathsf{T}}\ \overset{[k]}{W}^{-1}\ \overset{[k]}{\widetilde{y}}$. *Then*

C.1 $\quad \ln\det \overset{[k]}{V}_\lambda \quad = \quad \det\big(\widetilde{Q}_\lambda + \overset{[k]}{\Omega}\big) \ - \ \ln\det \widetilde{Q}_\lambda \ + \ \ln\det\big(\overset{[k]}{W}\big),$

C.2 $\quad \overset{[k]}{\widetilde{y}}^{\mathsf{T}}\ V_\lambda^{-1}\ \overset{[k]}{\widetilde{y}} \quad = \quad - \ \overset{[k]}{q}^{\mathsf{T}}\big(\widetilde{Q}_\lambda + \overset{[k]}{\Omega}\big)^{-1}\ \overset{[k]}{q} \ + \ \overset{[k]}{\widetilde{y}}^{\mathsf{T}} W^{-1} \overset{[k]}{\widetilde{y}}.$

This proposition makes it possible to reformulate (17) and define the kth step update, $\overset{[k]}{\lambda}_Q$ and $\overset{[k]}{\lambda}_R$, as

$$\underset{\lambda \succeq 0}{\operatorname{argmin}} \Big\{ \ln\det\Big\{\big(\widetilde{Q}_\lambda + \overset{[k]}{\Omega}\big)\, \widetilde{Q}_\lambda^{-1}\Big\} - \ \overset{[k]}{q}^{\mathsf{T}}\big(\widetilde{Q}_\lambda + \overset{[k]}{\Omega}\big)^{-1} \overset{[k]}{q}\Big\}. \tag{18}$$

It is important to use an efficient and accurate method to solve (18) since this problem appears in every step k and determines when the entire algorithm terminates (when $\|\overset{[k]}{\lambda} - \overset{[k\text{-}1]}{\lambda}\|$ is sufficiently small). To achieve this, we have analytically derived the gradient and the Hessian of the objective function. (Details are in the the Appendix.) The final algorithm for selecting the regularization parameters is outlined here:

Algorithm 1 Finding regularization parameters in griPEER

Input: matrices: Z, X and Q; vector: y; initial point: $\overset{[0]}{\lambda} := [\overset{[0]}{\lambda_Q}, \overset{[0]}{\lambda_R}]^{\mathsf{T}}$; stop criterion: $\delta > 0$; function which defines the density: ψ; $k := 1$

do

1. define $\overset{[k-1]}{\beta}$ and $\overset{[k-1]}{b}$ by solving:
$$\underset{\beta,b}{\operatorname{argmin}} \left\{ -2\sum_{i=1}^{n} \left[y_i(X_i\beta + Z_i b) - \psi(X_i\beta + Z_i b) \right] + \overset{[k-1]}{\lambda_Q} b^T Q b + \overset{[k-1]}{\lambda_R} \|b\|_2^2 \right\};$$
2. $\overset{[k-1]}{\theta} := X\overset{[k-1]}{\beta} + Z\overset{[k-1]}{b}$, $\quad \overset{[k]}{W} := \operatorname{diag}\left(\left[\psi''(\overset{[k-1]}{\theta_1})\right]^{-1}, \ldots, \left[\psi''(\overset{[k-1]}{\theta_n})\right]^{-1} \right)$;
3. define $\overset{[k]}{y}$ by putting $\overset{[k]}{y_i} := \left[\psi''(\overset{[k-1]}{\theta_i})\right]^{-1} \cdot \left(y_i - \psi'(\overset{[k-1]}{\theta_i})\right) + \overset{[k-1]}{\theta_i}$, $\quad$ for $i = 1, \ldots, n$;
4. $\overset{[k]}{\mathcal{P}} := \mathrm{I} - X(X^{\mathsf{T}}\overset{[k]}{W^{-1}}X)^{-1}X^{\mathsf{T}}\overset{[k]}{W^{-1}}$;
5. $\overset{[k]}{\widetilde{y}} := \overset{[k]}{\mathcal{P}}\overset{[k]}{y}$, $\quad \overset{[k]}{X} := \overset{[k]}{\mathcal{P}}X$, $\quad \overset{[k]}{Z} := \overset{[k]}{\mathcal{P}}Z$;
6. $\overset{[k]}{\Omega} := \overset{[k]\mathsf{T}}{Z}\ \overset{[k]-1}{W}\ \overset{[k]}{Z}$, $\quad \overset{[k]}{q} := \overset{[k]\mathsf{T}}{Z}\ \overset{[k]-1}{W}\ \overset{[k]}{\widetilde{y}}$;
7. $\overset{[k]}{\lambda} := \underset{\lambda \succeq 0}{\operatorname{argmin}} \left\{ \ln\left|(\lambda_Q Q + \lambda_R \mathrm{I}_p + \overset{[k]}{\Omega})(\lambda_Q Q + \lambda_R \mathrm{I}_p)^{-1}\right| - \overset{[k]\mathsf{T}}{q}(\lambda_Q Q + \lambda_R \mathrm{I}_p + \overset{[k]}{\Omega})^{-1}\overset{[k]}{q} \right\}$;
8. $k \leftarrow k + 1$;

while $\{\|\overset{[k]}{\lambda} - \overset{[k-1]}{\lambda}\|/\|\overset{[k-1]}{\lambda}\| > \delta\}$

4. Procedures for the significance testing

Unlike the lasso estimation procedure that produces a sparse set of regression coefficients but does not (without additional theory Zhao and Shojaie (2016)) provide statistical significance testing, we employ two methods to identify variables that are identified as statistically significantly related to the response. Two such approaches are implemented in our software and we introduce them in this section. They both use the knowledge about the optimal regularization parameters described in the previous section. The first takes advantage of asymptotic properties of generalized linear model (GLM) estimates and construct the estimate of asymptotic variance-covariance matrix in the similar fashion as proposed by Cessie and Houwelingen (1992) in the context of ridge-penalized logistic regression. The second applies the bootstrap method. When griPEER is used for variable selection, we will refer to these two approaches as griPEER$_{\text{asmp}}$ (the asymptotic-based approach) and griPEER$_{\text{boot}}$ (the bootstrap-based approach), respectively. The numerical experiments performed in Section 5 suggest that griPEER$_{\text{boot}}$ is able to achieve significantly larger power than griPEER$_{\text{asmp}}$ under the settings reflecting brain imaging design and connectivity matrices. Since the same experiment shows similar rates of false discoveries among variables labeled as relevant, griPEER$_{\text{boot}}$ was used in real data analysis (Section 6) to find brain regions associated with HIV.

4.1. Asymptotic variance-covariance matrix

We start by introducing notation. Denote $\mathcal{X} := [X,\ Z]$, let $\mathcal{B}$ be $p+m$ dimensional estimate given by (6), and $\theta := \mathcal{X}\mathcal{B}$. Moreover, we define a $(p+m) \times (p+m)$ penalty

matrix as

$$\mathcal{Q} := \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & \lambda_Q Q + \lambda_R \, \mathrm{I}_p \end{array}\right], \tag{19}$$

where the non-negative parameters λ_Q, λ_R are adjusted by the procedure 1. In summary, $\mathcal{B}$ is the solution to

$$\underset{B \in \mathbb{R}^{p+m}}{\operatorname{argmin}} \; \Big\{ 2 \sum\nolimits_i \psi(\mathcal{X}_i B) \; - \; 2 y^{\mathsf{T}} \mathcal{X} B \; + \; B^{\mathsf{T}} \mathcal{Q} B \Big\}, \tag{20}$$

with ψ being a given function indicating the member of the exponential family of distributions (3). Furthermore, the formulas we derive in this section include the diagonal matrix Ψ defined as $\Psi := \operatorname{diag}\big\{\psi''(\theta_1), \ldots, \psi''(\theta_n)\big\}$.

Using the first-order Taylor approximation, as well as asymptotic properties of GLM estimate, one can find that the estimate asymptotic variance for $\mathcal{B}$ has a form

$$\operatorname{var}_a(\mathcal{B}) = \big(\mathcal{X}^{\mathsf{T}} \Psi \mathcal{X} + \mathcal{Q}\big)^{-1} \mathcal{X}^{\mathsf{T}} \Psi \mathcal{X} \big(\mathcal{X}^{\mathsf{T}} \Psi \mathcal{X} + \mathcal{Q}\big)^{-1}. \tag{21}$$

The derivation is based on Cessie and Houwelingen (1992) and was described in detail in Appendix A.3. Based on the above formula, we propose a simple decision-making strategy in which we label the ith covariate as statistically relevant if 0 is not included in the 95% confidence interval for its respective regression coefficient, i.e.

$$0 \notin \Big[\mathcal{B}_i - 1.96 \cdot \sqrt{\operatorname{var}_a(\mathcal{B})_{ii}} \; , \quad \mathcal{B}_i + 1.96 \cdot \sqrt{\operatorname{var}_a(\mathcal{B})_{ii}} \Big]. \tag{22}$$

4.2. The Bootstrap based approach

In this approach the variances of coefficients in $\mathcal{B}$, the solution to (20), was estimated based on Bootstrap samples. Each such sample was created from n elements of y and n corresponding rows of Z and X, which indices were selected randomly by sampling with replacement. The dataset obtained in jth repetition, $X^{[j]}$, $Z^{[j]}$ and $y^{[j]}$, were then substituted to the objective in (6) with λ_Q and λ_R being selected by Algorithm 1 applied to the original dataset (i.e., λ_Q and λ_R were estimated only once). The percentile bootstrap confidence intervals, with the significance level $\alpha = 0.05$, were defined based on all estimates, $\mathcal{B}^{[1]}$, ..., $\mathcal{B}^{[s]}$. The default value of s was set to 500 in our software and this number of bootstrap samples was generated in simulations performed in subsection 5.3. Coefficients from the griPEER estimate whose confidence intervals do not contain zero are labeled as response-related discoveries.

5. Numerical experiments

We conduct a simulation study to investigate the performance of griPEER in the situation when responses are modeled by binomial distribution. Results are compared with the logistic ridge estimates.

5.1. Definitions

Matrix density. For a $p \times q$ matrix A define its *density* as a proportion of non-zero entries,

$$dens(A) := \frac{1}{pq} \sum_{i,j} \mathbb{I}\{|A(i,j)| > 0\}. \tag{23}$$

Matrix dissimilarity. To quantify a dissimilarity between two $p \times q$ matrices, A and B, with $dens(A) = dens(B)$, we define

$$diss(A,B) := \Big(\sum_{i,j} \mathbb{I}\{|A(i,j) - B(i,j)| > 0\}\Big) / \Big(2 \sum_{i,j} \mathbb{I}\{B(i,j) > 0\}\Big), \tag{24}$$

with values in the interval $[0,1]$. If $diss(A,B) = 0$ then $A = B$ and $diss(A,B) = 1$ means that the positions of non-zero entries do not overlap.

5.2. Model coefficient estimation

5.2.1. Settings

"Informativeness" of the penalty term. The simulation settings were designed to evaluate performance in a variety of situations ranging from an "observed" connectivity matrix (i.e., a prescribed matrix used in estimation) that is fully informative to one that is completely non-informative. Here "informativeness" refers to the amount of true dependencies among the variables that are represented in the connectivity matrix.

Denote By $\mathcal{A}^{true}$ a matrix representing true connections between variables and by $\mathcal{A}^{obs}$ one which is observed and used in an estimation via griPEER. To express "informativeness" of $\mathcal{A}^{obs}$ with respect to $\mathcal{A}^{true}$, we use a measure of dissimilarity, $diss(\mathcal{A}^{obs}, \mathcal{A}^{true})$, defined in (24). We have

- $diss(\mathcal{A}^{obs}, \mathcal{A}^{true}) = 0$ reflects a situation when $\mathcal{A}^{obs}$ is fully informative;
- $diss(\mathcal{A}^{obs}, \mathcal{A}^{true}) = 1$ reflects a situation when $\mathcal{A}^{obs}$ is non-informative;
- $diss(\mathcal{A}^{obs}, \mathcal{A}^{true}) \in (0,1)$ indicates $\mathcal{A}^{obs}$ is partially informative.

Connectivity in the context of brain regions. One may view $\mathcal{A}^{true}$ as an adjacency matrix of a graph representing the connections between brain regions, and our simulations scenarios are based on the following four interpretations regarding this structure.

1. $\mathcal{A}_1$: *"homologous regions"*. $\mathcal{A}_1$ represents a situation when brain regions, i and j, are connected (i.e., $\mathcal{A}_1(i,j) = 1$) if i and j are homologous brain regions from different hemispheres, and $\mathcal{A}_1(i,j) = 0$ otherwise. This matrix is shown in Figure 2, left plot.
2. $\mathcal{A}_2$: *"modularity"*. $\mathcal{A}_2$ represents a situation when brain regions i and j are connected if and only if they belong to the same *module* with $\mathcal{A}_2(i,j) = 1$ within the module and 0 otherwise. This matrix is shown in Figure 2, middle left plot.

3. $\mathcal{A}_3$: *"density of connections, masked"*. $\mathcal{A}_3$ is defined based on the brain-imaging measure — *density of connections between brain regions* (see, Section 6) — and then is "masked" by modularity information. Here, $\mathcal{A}_3(i,j)$ equals the median of a density of connections between regions i and j if they belong to the same module. Otherwise, $\mathcal{A}_3(i,j) := 0$. Matrix $\mathcal{A}_3$ is shown in Figure 2, middle right plot.

4. $\mathcal{A}_4$: *"neighboring regions"*. $\mathcal{A}_4$ represents a situation when brain regions i and j are connected if they are "close" according to their spatial location ($\mathcal{A}_2(i,j) > 0$). Otherwise, they are not connected ($\mathcal{A}_4(i,j) := 0$). This matrix is shown in Figure 2, right plot.

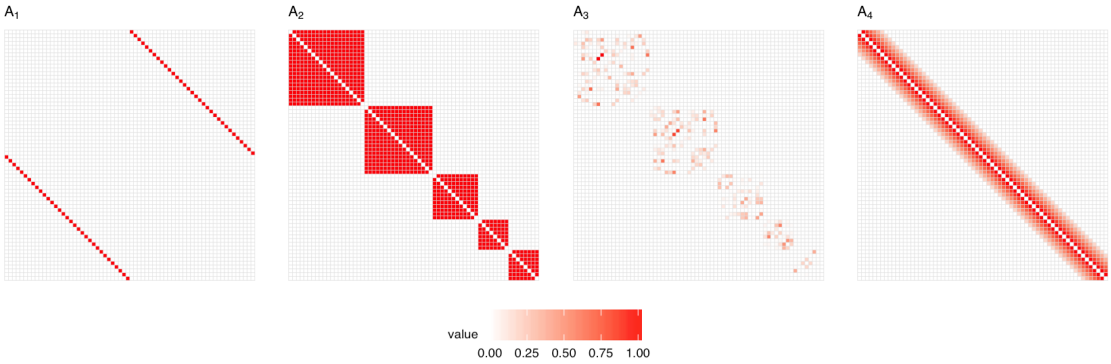


Figure 2: Matrices used in the simulation study to construct $\mathcal{A}^{true}$. Presented are variants for $p = 66$. Left plot: $\mathcal{A}_1$ *"homologous regions"*. Middle left plot: $\mathcal{A}_2$ *"modularity"*. Middle right plot: $\mathcal{A}_3$ *"density of connections, masked"*. Right plot: $\mathcal{A}_4$ *"neighboring regions"*.

A *homologous regions* matrix $\mathcal{A}_1$ reflects the situation where only homologous regions from two hemispheres are assumed to be connected. A *modularity* matrix $\mathcal{A}_2$, in turn, represents adjacency defining division of the brain cortical regions into five modules (Sporns, 2013; Cole et al., 2014; Sporns and Betzel, 2016). Next, a *"density of connections, masked"* matrix $\mathcal{A}_3$ is based on estimated density of connections between brain cortical regions, as described in Section 6). Finally, the *"neighboring regions"* matrix $\mathcal{A}_4$ models the situation where brain regions are spatially connected; i.e., the strength of connection between brain regions depends on the physical distance between them.

Simulation scenarios. We run three simulation scenarios to express different sources of "uninformativeness" $\mathcal{A}^{obs}$ which loosely reflect real-life scenarios. For each scenario, we tested all four types of matrices, $\mathcal{A}_1, ..., \mathcal{A}_4$.

- **Scenario 1.** The observed connectivity matrix, $\mathcal{A}^{obs}$, represents connections (partially) permuted with respect to connections represented by $\mathcal{A}^{true}$. Based on one of four considered matrices, the corresponding $\mathcal{A}^{obs}$ matrix is constructed by randomizing edges of a graph given by A^{true} until a desired dissimilarity,

$diss(\mathcal{A}^{obs}, \mathcal{A}^{true})$, is achieved (see: Fig. 3). The randomization technique preserves graph size, density, strength and graph degree-sequence (and hence degree distribution).

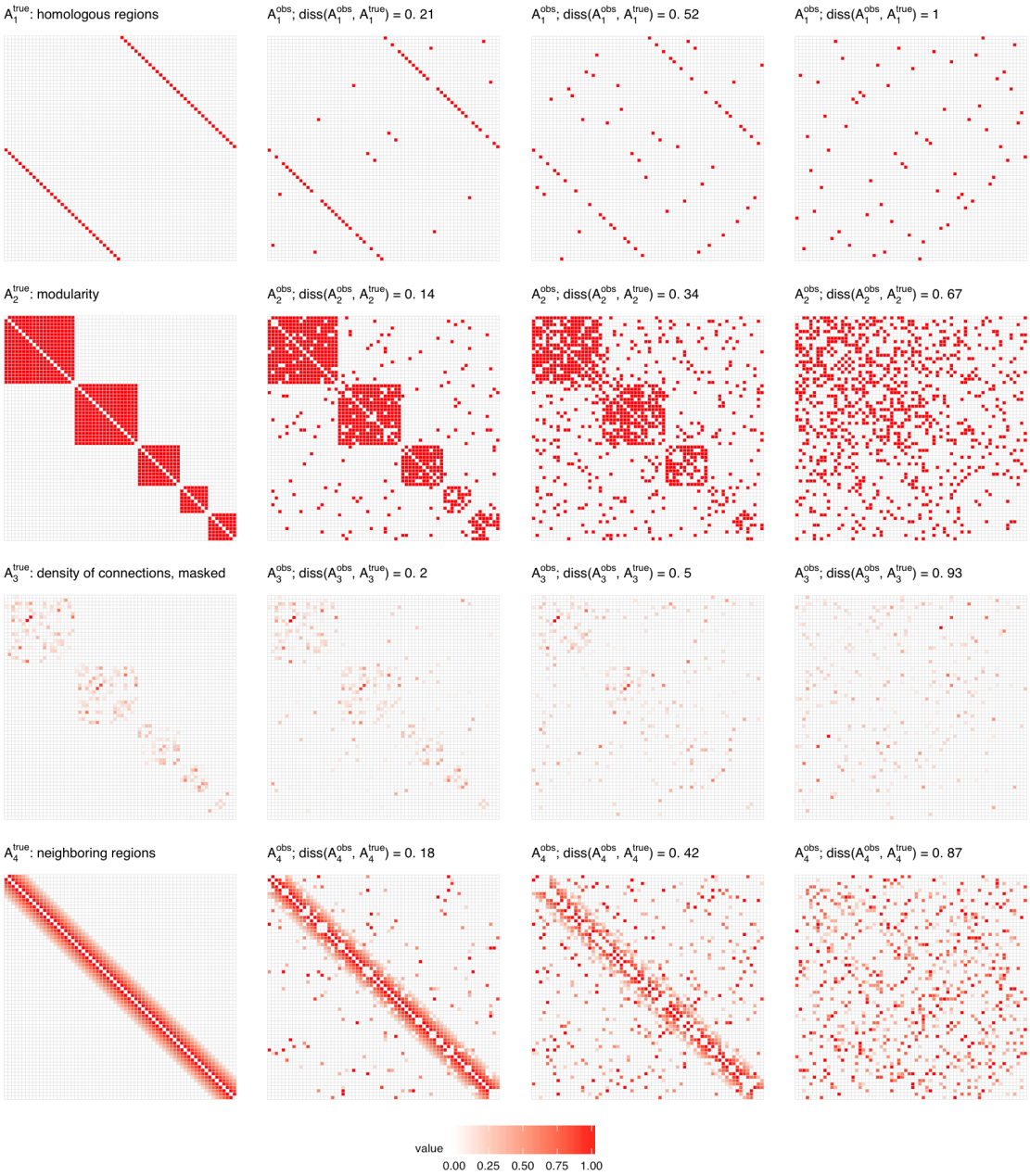


Figure 3: $\mathcal{A}^{true}$ connectivity graph adjacency matrices (1st column panel) and $\mathcal{A}^{obs}$ connectivity graph adjacency matrices (2nd-4th column panels) used in Scenario 1. $\mathcal{A}^{obs}$ matrix is constructed by randomizing $\mathcal{A}^{true}$ until a desired dissimilarity, $diss(\mathcal{A}^{obs}, \mathcal{A}^{true})$, is achieved ($diss$ is growing when moving from left to right side of each row plot panel).

- **Scenario 2.** We investigate the impact of using inaccurate information by labeling the negative connections between variables as the positive. For $\mathcal{A}_i$, with $i \in \{1, \ldots, 4\}$, the true signal was generated after changing the structure of variables dependencies by allowing some negative connections first. Specifically, $\mathcal{A}^{true}$ was defined by turning entries of columns $k \in \{1, 4, 7, 10\}$ and corresponding rows of $\mathcal{A}_i$ into their negative values. Here, $\mathcal{A}^{obs}_{i,j} = |\mathcal{A}^{true}_{i,j}|$ and hence $\mathcal{A}^{obs}$ contains only non-negative values.
- **Scenario 3.** The observed connectivity matrix $\mathcal{A}^{obs}$ is of lower or higher matrix density than $\mathcal{A}^{true}$. For $\mathcal{A}^{true}$ defined based on one of four considered matrices, the corresponding $\mathcal{A}^{obs}$ is constructed by randomly removing, respectively adding, edges to the graph of connections represented by $\mathcal{A}^{true}$ until the desired ratio of matrix densities, $dens(\mathcal{A}^{obs})/dens(\mathcal{A}^{true})$, is obtained (see Fig. 5).

Simulation procedure. In each numerical experiment, we perform the following steps.

1. For graph adjacency $\mathcal{A}^{true}$, compute its normalized Laplacian, Q^{true} (in Scenario 2 the node's degree is defined as $d_i := \sum_j |a_{ij}|$; see subsection 2.1).
2. Replace the zero singular values of Q^{true} by $0.01 \cdot s$, where s is the smallest nonzero singular value of Q^{true} $\big($to get an invertible matrix required in 6. (a)$\big)$.
3. For graph adjacency matrix, $\mathcal{A}^{obs}$, compute its normalized Laplacian, Q^{obs}.
4. Generate $Z \in \mathbb{R}^{n \times p}$, where the rows are independently distributed by $\mathcal{N}_p(0, \Sigma)$, where Σ is variance-covariance matrix estimated from a real data study (see: Sect. 6); standarize columns of Z so as they have mean 0 and unit ℓ_2 norm.
5. Generate X as n-dimensional column of ones.
6. Run the following steps 100 times:
   (a) generate $b \in \mathbb{R}^p$ as $b \sim \mathcal{N}\big(0, \sigma_b^2 (Q^{true})^{-1}\big)$; set $\beta = 0$,
   (b) define $\theta := X\beta + Zb$,
   (c) define $pr^{Binom} := \big[\, e^{\theta_1}/(1+e^{\theta_1}), \ldots, e^{\theta_n}/(1+e^{\theta_n}) \,\big]^{\mathsf{T}}$,
   (d) generate $y \sim Binom(pr^{Binom})$, $y \in \mathbb{R}^{n \times 1}$,
   (e) estimate model coefficients b, β with the two methods: (1) griPEER, assuming the binomial distribution of y and using Q^{obs} in a penalty term, (2) logistic ridge estimator,
   (f) compute b estimation error, $MSEr := \|\hat{b} - b\|_2^2 / \|b\|_2^2$, for two b estimates, (1) $\hat{b}^{\,griPEER}$ and (2) $\hat{b}^{\,l.\,ridge}$.
7. Compute mean MSEr out of the 100 runs from (5), for the two estimation methods.

Importantly, a "true" coefficient vector b obtained as $b \sim \mathcal{N}(0, \sigma_b^2 (Q^{true})^{-1})$ reflects the connectivity structure represented by $\mathcal{A}^{true}$. Examplary vectors b generated based on $\mathcal{A}_1, ..., \mathcal{A}_4$ are presented in Figure 11 in Appendix B.

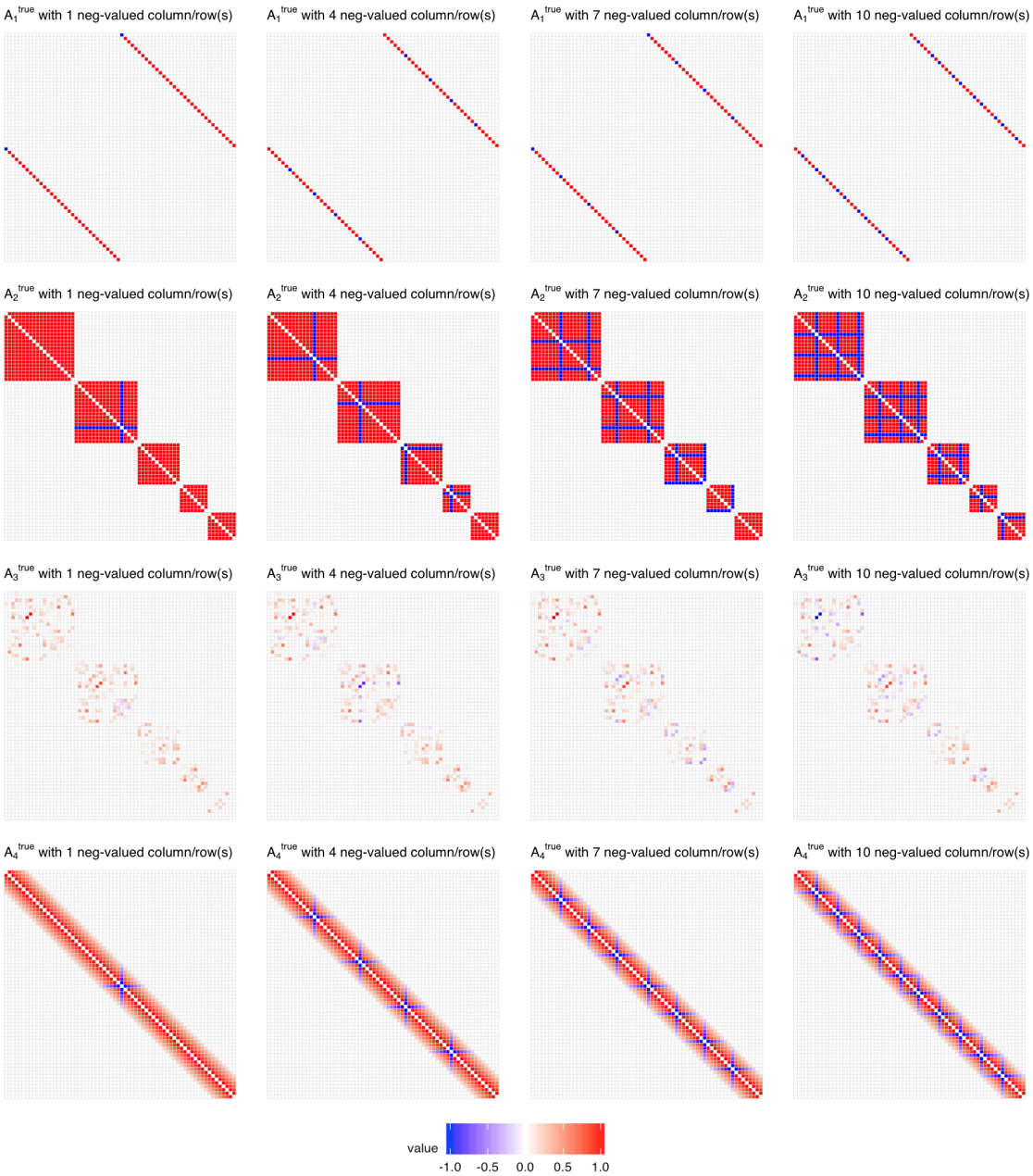


Figure 4: $\mathcal{A}^{true}$ connectivity graph adjacency matrices used in Scenario 2. $\mathcal{A}^{true}$ matrix is constructed from $\mathcal{A}_1, ..., \mathcal{A}_4$ matrices (1st-4th row panels, respectively) by turning entries of k, $k \in \{1, 4, 7, 10\}$, columns (and corresponding rows) of this matrix into their negative values (k is growing when moving from left to right side of each row plot panel).

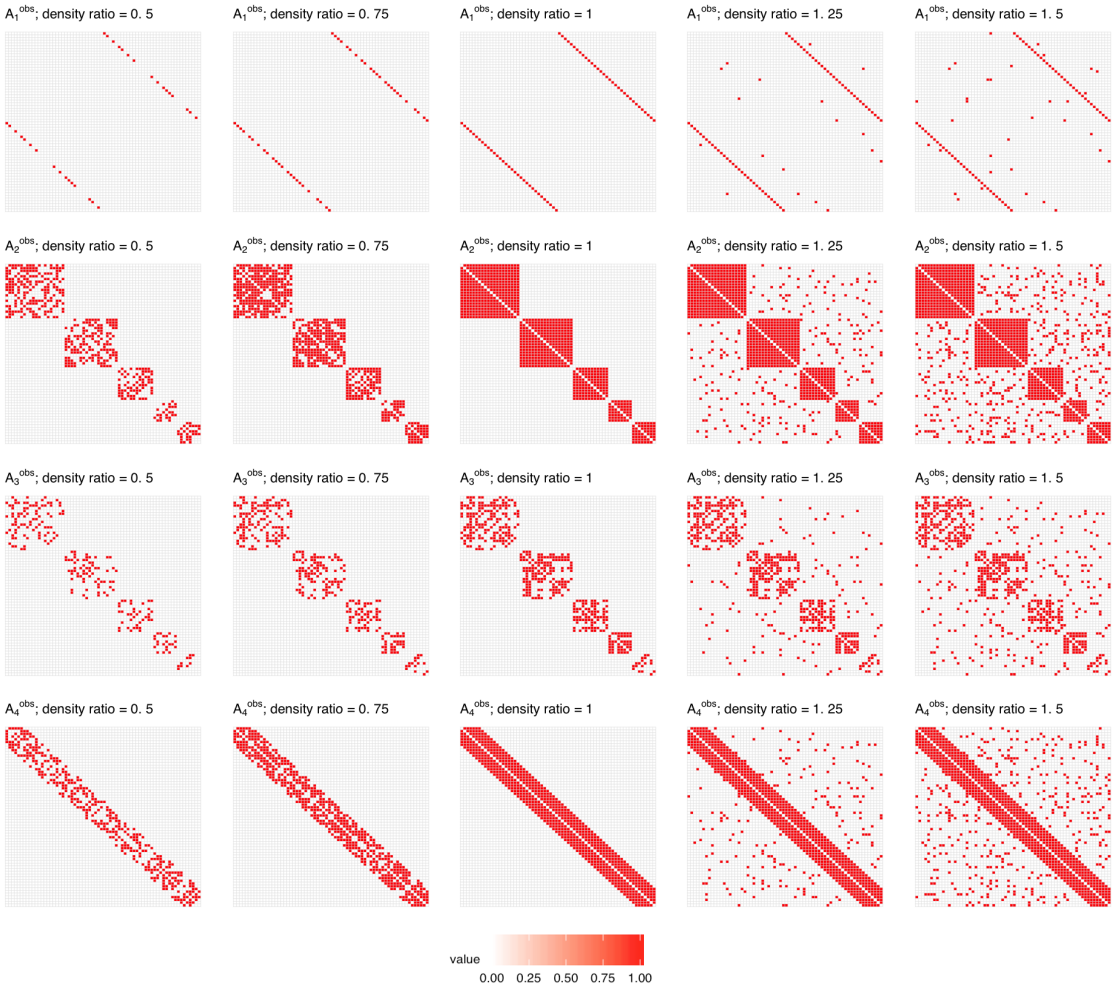


Figure 5: $\mathcal{A}^{true}$ and $\mathcal{A}^{obs}$ connectivity graph adjacency matrices used in Scenario 3. $\mathcal{A}^{true}$ matrix is defined as one of $\mathcal{A}_1, ..., \mathcal{A}_4$ matrices (1st-4th row panels, respectively). Corresponding $\mathcal{A}^{obs}$ is constructed by randomly removing / adding edges to the graph of connections represented by $\mathcal{A}^{true}$ until desired density ratio, $dens(\mathcal{A}^{obs})/dens(\mathcal{A}^{true})$, is obtained (ratio is growing from 0.5 to 1.5 when moving from left to right side of each row plot panel).

Simulation parameters. We consider the following choices of the experimental settings:

1. number of predictors: $p \in \{66, 198\}$,
2. number of observations: $n \in \{100, 200\}$,
3. $\mathcal{A}^{true}$ matrix constructed based on $\mathcal{A}_i \in \{\mathcal{A}_1, ..., \mathcal{A}_4\}$,
4. (Scenario 1.) dissimilarity between $\mathcal{A}^{obs}$ and $\mathcal{A}^{true}$: $diss(\mathcal{A}^{obs}, \mathcal{A}^{true}) \in [0, 1]$,
5. (Scenario 2.) number of columns (and corresponding rows) of $\mathcal{A}^{true}$ that have switched signs: $k \in \{0, 1, 4, 7, 10\}$,
6. (Scenario 3.) density ratio: $dens(\mathcal{A}^{obs})/dens(\mathcal{A}^{true}) \in [0.5, 1.5]$.

The number of predictors, $p = 66$, is motivated by the brain imaging analysis described in Section 6, where 66 brain regions were considered. To investigate the situations with larger number of predictors for ith type of connectivity pattern, we created block-diagonal adjacency matrices with $\mathcal{A}_i$'s as blocks. The adjacency matrix in the case with $p = 198$ was therefore defined as $diag\{\mathcal{A}_i, \mathcal{A}_i, \mathcal{A}_i\}$.

5.2.2. Results

Scenario 1. In Scenario 1, we compare griPEER and logistic ridge estimation methods in a situation when an observed connectivity matrix $\mathcal{A}^{obs}$ contains connections that are permuted with respect to connections represented by $\mathcal{A}^{true}$. We consider combinations of simulation parameter values: number of predictors $p \in \{66, 198\}$, number of observations $n \in \{100, 200\}$, $\mathcal{A}^{true}$ base matrix $\mathcal{A}_1, ..., \mathcal{A}_4$, dissimilarity between $\mathcal{A}^{obs}$ and $\mathcal{A}^{true}$ $diss(\mathcal{A}^{obs}, \mathcal{A}^{true}) \in [0, 1]$. Fig. 6 displays the aggregated (mean) values of the relative estimation error based on 100 simulation runs.

We observe that in each case, MSEr of griPEER is lower or equal to MSEr of logistic ridge. The utility of griPEER is particularly apparent in cases with fully informative and largely informative $\mathcal{A}^{obs}$; these cases correspond to low values of dissimilarity $diss(\mathcal{A}^{obs}, \mathcal{A}^{true})$ (marked at x-axis). As $\mathcal{A}^{obs}$ gets less informative about the true connections between coefficients in a model, MSEr of griPEER approaches MSEr of logistic ridge; these cases correspond to high values of dissimilarity $diss(\mathcal{A}^{obs}, \mathcal{A}^{true})$. The result illustrates an important property of griPEER estimation method: adaptiveness to the amount of true information contained in an observed $\mathcal{A}^{obs}$ matrix. When $\mathcal{A}^{obs}$ is largely informative, incorporating $\mathcal{A}^{obs}$ into the estimation is clearly a benefits. When $\mathcal{A}^{obs}$ carries little or no information about the true connections between model coefficients, griPEER yields MSEr no larger than MSEr of logistic ridge estimator.

The performances of griPEER and logistic ridge depend on the structure of connections imposed by $\mathcal{A}^{true}$ on the true b. We can observe that a difference between MSEs for griPEER and logistic ridge is smaller when $\mathcal{A}^{true}$ is defined based on $\mathcal{A}_1$: *homologous regions* matrix (Fig. 6, left column panel). Indeed, $\mathcal{A}_1$ has smaller density than $\mathcal{A}_2$, $\mathcal{A}_3$, $\mathcal{A}_4$ matrices, and imposes fewer connections between true coefficients

in a model. Therefore, utilizing (full or partial) connectivity information $\mathcal{A}^{obs}$ in estimation for $\mathcal{A}_1$-based signals is less beneficial compared to other considered patterns of coefficients dependencies. Furthermore, when each node is connected with everty other by a path consisting of strong connections, as in a case when $\mathcal{A}^{true}$ is created based on $\mathcal{A}_4$ (4th column panel in Fig. 6), it is expected that all "true" model coefficients in a generated vector b are strongly dependent on each other; see Fig. 11 in Appendix B. In such a situation, even using even inaccurate information about the connections (high $diss(\mathcal{A}^{obs}, \mathcal{A}^{true})$ values) may be still be beneficial, as long as the correct message about strong coefficients' dependence is provided. Finally, if we compare the results within each column panel of Fig. 6, we observe, as expected, that the estimation error gets smaller as number of predictors p gets smaller and as number of observations n gets larger.

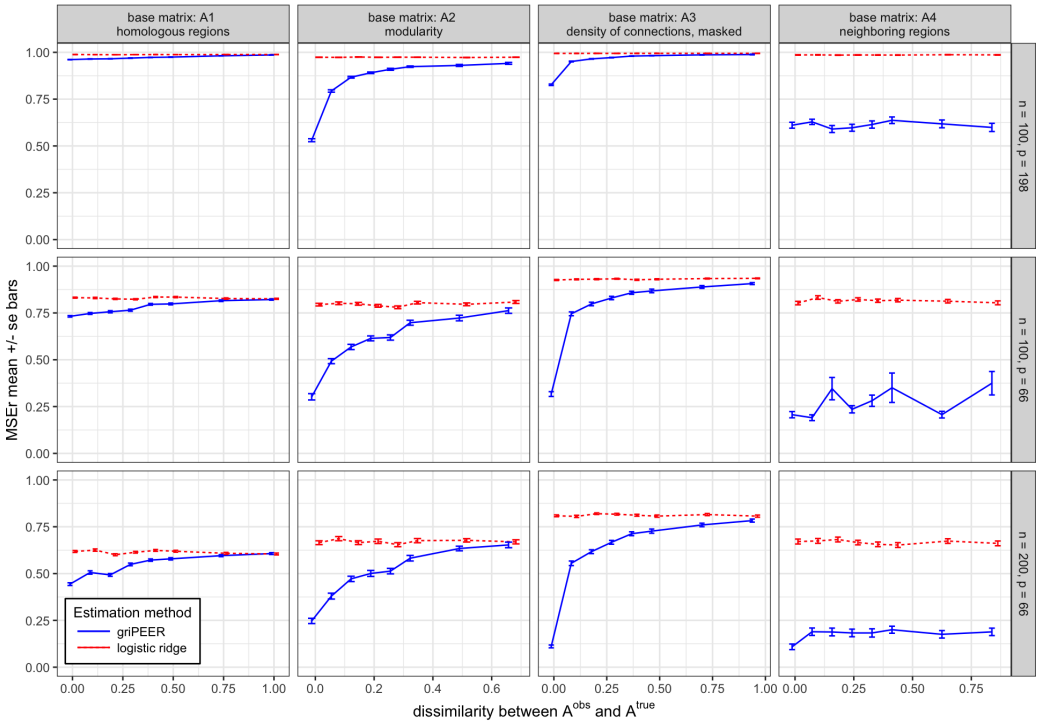


Figure 6: MSEr for estimation of b in Scenario 1. Results for griPEER (blue line) and logistic ridge (gray line). Presented are the average values of MSEr from 100 experiment runs for: $n \in \{100, 200\}$, $p \in \{66, 198\}$ and four true connectivity pattern inducing matrices, $\mathcal{A}_1, ..., \mathcal{A}_4$. Dissimilarity between $\mathcal{A}^{obs}$ and $\mathcal{A}^{true}$ measured by $diss(\mathcal{A}^{obs}, \mathcal{A}^{true})$ is represented by x-axis. Standard error of the mean bars are showed.

Scenario 2. In Scenario 2., we compare griPEER and logistic ridge estimation methods in a situation when an observed connectivity matrix, $\mathcal{A}^{obs}$, represents only positive connections, whereas $\mathcal{A}^{true}$ represents both positive and negative connections. We run the simulation for number of observations, $n = 100$, number of variables,

$p = 66$, and for $\mathcal{A}^{true}$ created based on four connectivity pattern inducing matrices, $\{\mathcal{A}_1, ..., \mathcal{A}_4\}$. Matrix $\mathcal{A}^{true}$ was generated form $\mathcal{A}_i$ by switching signs in k columns (and corresponding rows), were $k \in \{1, 4, 7, 10\}$. Fig. 7 displays the aggregated (mean) values of the relative estimation error based on 100 simulation runs.

With increasing k, $\mathcal{A}^{obs}$ increasingly differs from the connectivity pattern used in the true signal generation and so the relative difference between MSEr for logistic ridge and griPEER decreases (for nearly all settings). Notably, MSEr for griPEER remains less than or equal to MSEr for logistic ridge. The results suggest that even using some incorrect information regarding the true connectivity structure (such as misspecifying negative dependencies as being positive) is not detrimental.

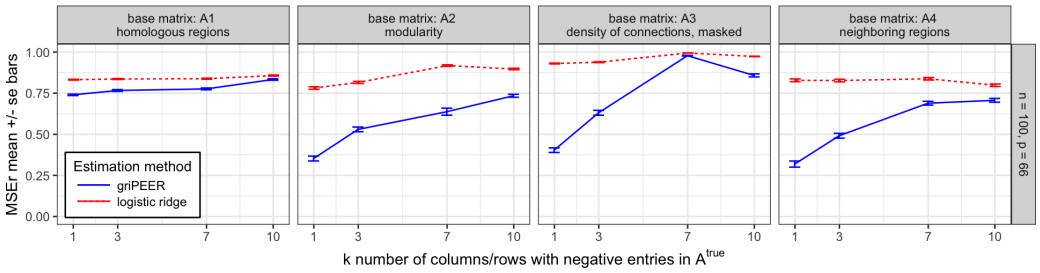


Figure 7: MSEr for estimation of b in Scenario 2. Results for griPEER (blue line) and logistic ridge (gray line). Presented are the average values of MSEr from 100 experiment runs for $n = 100$, $p = 66$ and four true connectivity pattern inducing matrices, $\mathcal{A}_1, ..., \mathcal{A}_4$. The number of columns (and corresponding rows) of $\mathcal{A}_i$, for which entries signs where switched in $\mathcal{A}^{true}$ construction is represented by x-axis. Standard error of the mean bars are showed.

Scenario 3. In this scenario, we compare griPEER and logistic ridge estimation methods in a situation when $\mathcal{A}^{obs}$ is of lower / higher matrix density than $\mathcal{A}^{true}$. As in Scenario 2, we consider $n = 100$ and $p = 66$. This time, we do not change the signs of $\{\mathcal{A}_1, ..., \mathcal{A}_4\}$ matrices but we generate $\mathcal{A}^{true}$ by adding/removing some connections to/from $\mathcal{A}_i$. This influences the density of resulting matrix. In the simulation we consider $dens(\mathcal{A}^{obs})/dens(\mathcal{A}^{true}) \in [0.5, 1.5]$ as a densities ratio range. Fig. 8 displays the mean values of the relative estimation error based on 100 simulation runs.

We can observe that, similar to Scenario 1, incorporating information on only a few connections ($\mathcal{A}_1$ case) yields the smallest gain in the estimation accuracy measured by MSEr among all considered connectivity patterns. If $\mathcal{A}^{true}$ is set to $\mathcal{A}_4$, then (again, analogously to Scenario 1) the information about strong coefficients' dependence is provided through $\mathcal{A}^{obs}$. This results in substantially lower MSEr for griPEER across all densities ratio range we considered. When $\mathcal{A}^{true}$ is equal to one of modules-based matrices, $\mathcal{A}_2$ or $\mathcal{A}_3$, we still benefit from using $\mathcal{A}^{obs}$ of lower density than $\mathcal{A}^{true}$, since $\mathcal{A}^{obs}$ contains unaffected information about five separated modules in connectivity structure (values smaller than 1 at x-axis). Including the false connections in $\mathcal{A}^{obs}$ (values greater than 1 at x-axis) disturbs the message about

the lack of dependencies between modules. A loss in griPEER's estimation accuracy is apparent at the transition point $x = 1$. It remains, however, significantly better than the estimation accuracy for logistic ridge over the entire range of considered densities ratios.

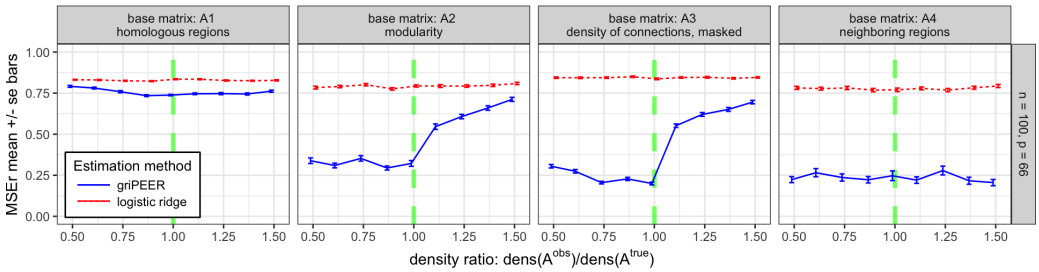


Figure 8: MSEr for estimation of b in Scenario 3. Results for griPEER (blue line) and logistic ridge (gray line). Presented are the average values of MSEr from 100 experiment runs for $n = 100$, $p = 66$ and four true connectivity pattern inducing matrices, $\mathcal{A}_1, ..., \mathcal{A}_4$. Ratio of densities, $dens(\mathcal{A}^{obs})/dens(\mathcal{A}^{true})$, is represented by x-axis and varies from 0.5 to 1.5. Standard error of the mean bars are showed. Green dashed vertical lines denote the cases when ratio of matrix densities equals 1; in these cases, $\mathcal{A}^{obs}$ is identical to $\mathcal{A}^{true}$.

5.3. Model coefficient significance testing

5.3.1. Settings

We design a simulation study to evaluate performance of the two procedures for coefficient significance testing for griPEER, introduced in Section 4: asymptotic variance-covariance matrix-based approach, griPEER$_{\text{asmp}}$, and Bootstrap-based approach, griPEER$_{\text{boot}}$.

Simulation scenario. We follow the simulation setting used in Scenario 1, described in subsection 5.2. Specifically, we assume that $\mathcal{A}^{obs}$ represents connections (partially) permuted with respect to connections represented by $\mathcal{A}^{true}$. I.e., the corresponding $\mathcal{A}^{obs}$ is constructed by randomizing entries in A^{true} until a desired dissimilarity, $diss(\mathcal{A}^{obs}, \mathcal{A}^{true})$, is achieved; see: Figure 3. The randomization technique preserves graph size, density, strength and graph degree-sequence (and hence degree distribution). Here, we confine ourselves to $p = 66$ and the case when $\mathcal{A}^{true}$ is based on $\mathcal{A}_3$; i.e., the median of a density of connections masked by modularity information, which corresponds to the construction of an adjacency matrix in the brain imaging analysis Section 4.

The adopted simulation scheme starts by generating the true signal and responses as in subsection 5.2. We generate large number of observations, $n = 1000$, but in the estimation we use only 150 records to emulate a real data setting. The large sample size is used only to label the variables which are "truly relevant" so that the performance of griPEER$_{\text{asmp}}$ and griPEER$_{\text{boot}}$ in the context of variables selection can be

assessed. Defining "truly relevant" variables is done through the asymptotic confidence interval for the logistic model estimate (non-regularized estimation), which is unbiased and asymptotic normal (Fahrmeir and Kaufmann, 1985). The details are described below.

Simulation study procedure. In the experiment, we perform the following steps.

1. Apply steps 1–5 from the simulation study procedure described in subsection 5.2 with $n = 1000$.
2. Run the following steps 100 times:
   (a) generate p-dimensional vector of true coefficients, b, as well as n-dimensional vectors, θ and y, by following steps 2(a)–2(d) described in subsection 5.2,
   (b) calculate the asymptotic standard deviations, $\delta_i := \sqrt{\left[\left(Z^\intercal \Psi Z\right)^{-1}\right]_{ii}}$, for $i = 1, \ldots, p$, where $\Psi := \operatorname{diag}\left\{\frac{e^{\theta_1}}{(e^{\theta_1}+1)^2}, \ldots, \frac{e^{\theta_n}}{(e^{\theta_n}+1)^2}\right\}$ (see, Appendix A.3),
   (c) divide the set of indices, $\{1, \ldots, p\}$, into two separated groups: I_T, corresponding to the variables defined as *relevant* and I_F, corresponding to the variables defined as *irrelevant*, by using the criterion
   $$i \in I_T \iff 0 \notin \left[b_i - 1.96\,\delta_i,\quad b_i + 1.96\,\delta_i\right],$$
   (d) generate the data for estimation, y^*, X^* and Z^*, by taking first 150 rows of y, X and Z; center and normalize the columns of Z^* to zero means and unit ℓ_2 norms,
   (e) apply griPEER$_{\text{asmp}}$ and griPEER$_{\text{boot}}$ on y^*, X^* and Z^* to indicate response-related variables defined by each of methods,
   (f) based on information about "truly relevant" and "truly irrelevant" variables; i.e., the known division into I_T and I_F, for each method identify: S — the number of true discoveries and V — the number of false discoveries,
   (g) for each method collect measures $pow^* := \frac{S}{|I_T|}$ and $fdr^* := \frac{V}{V+S}$,
3. Define the estimates of *power* and *FDR* as the averages of pow^* and fdr^* (across 100 repetitions of the step 2).

5.3.2. Results

Figure 9 displays the values of power (left plot) and FDR (right plot), estimated based on the simulation procedure described in subsection 5.3.1. As expected, for both methods power decreases as $\mathcal{A}^{obs}$ becomes less informative regarding the true connections between coefficients in a model. We observe however, that griPEER$_{\text{boot}}$ is able to reach substantially higher power than griPEER$_{\text{asmp}}$ under considered settings. The estimated FDRs are not very distinct for both methods and they tend to be very similar for less accurate connectivity information.

The results obtained in the simulation study suggest that one can potentially gain power by utilizing griPEER$_{\text{boot}}$ for coefficient significance testing, compared to griPEER$_{\text{asmp}}$ approach. In addition, power gain occurs without a substantial increase of FDR. consequently, we employ griPEER$_{\text{boot}}$ in the real data application in Section 6.

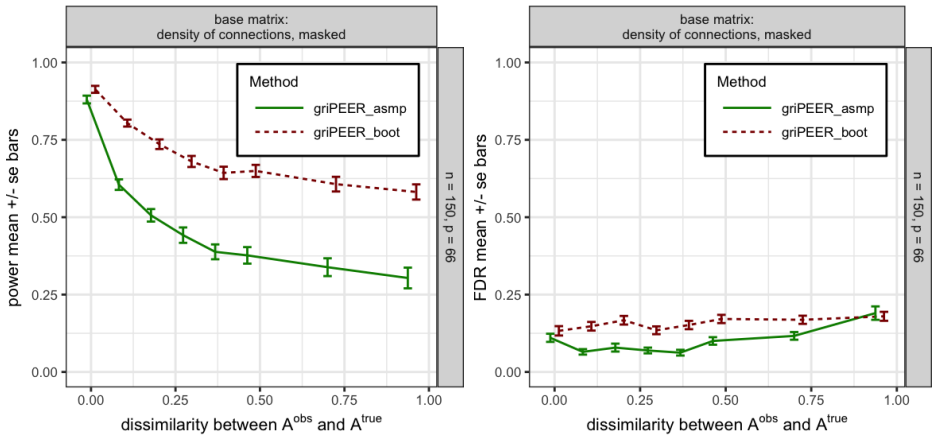


Figure 9: The estimated values for power (left plot) and FDR (right plot) obtained with asymptotic variance-covariance matrix-based approach (griPEER$_{\text{asmp}}$, blue line) and Bootstrap-based approach (griPEER$_{\text{boot}}$; gray line). Values are aggregated (mean) out of 100 experiment runs, for number of observations $n = 150$, number of variables $p = 66$ and $\mathcal{A}_3$ (Figure 2, middle right plot) as a true connectivity pattern. The dissimilarity between $\mathcal{A}^{obs}$ and $\mathcal{A}^{true}$ is represented by the x-axis. Standard error of the mean bars are showed.

5.4. The software used in simulations

The code used to generate the results was built in `Matlab` and is available at GitHub (https://github.com/dbrzyski/griPEER).

6. Imaging data application

We model the association between the presence/absence of HIV and the properties of the structural cortical brain imaging data. More specifically, we employ cortical thickness measurements obtained using the FreeSurfer software (Fischl, 2012) to classify the binary response indicating the status of HIV infection, where 0 indicates an HIV-negative individual and 1 an HIV-positive individual.

6.1. Data and preprocessing

Study sample. The analyzed sample consists of 162 young (age range: 18–42 years) males, where 108 were HIV-positive and 54 were HIV-negative. Study sample subjects' demographic and a HIV-related characteristics are summarized in Table 1.

Study variables	Min	Max	Mean	Median	StdDev
Age	18	41	25.80	23	6.47
Recent CD4	20	1179	461.83	446	243.81
Nadir CD4	15	690	289.13	293	158.31

Table 1: Study sample subjects' characteristics.

Cortical measurements. The FreeSurfer software package (version 5.1) was used to process the acquired structural MRI data, including gray-white matter segmentation, reconstruction of cortical surface models, labeling of regions on the cortical surface and analysis of group morphometry differences. The resulting dataset has cortical measurements for 68 cortical regions with parcellation based on Desikan-Killiany atlas (Desikan et al., 2006). The subset of 66 variables describing average gray matter thickness (in millimeters) of gray matter brain regions did not incorporate left and right insula due to their exclusion from the structural connectivity matrix.

Structural connectivity information. In the analysis we used two adjacency matrices, which were incorporated in the estimation with griPEER through the normalized Laplacian matrix. The adjacency matrices were created based on two structural connectivity information types: density of connections (DC) and fractional anisotropy (FA). For each of them, two steps were performed to achieve the final adjacency matrix, $\mathcal{A}$. In the first step, we computed the entry-wise median (across subjects) of DC or FA connectivity matrices. The second step relied on "masking by modularity partition", i.e. limiting the information achieved in the first step only to the connections between brain regions being in the same modules (i.e. we set $\mathcal{A}_{ij} := 0$, if regions i and j were not in the same module). For this purpose, we used the modularity connectivity matrix (see Sporns (2013); Cole et al. (2014); Sporns and Betzel (2016)), which defines the division of the brain into five separated communities. The modularity matrix was obtained by using Louvain method (Blondel et al., 2008) and based on model proposed in Hagmann et al. (2008). More details on this construction can be found in Karas et al. (2017).

6.2. Estimation methods

We employed logistic ridge and griPEER$_{\text{boot}}$ to classify the HIV-infected and non-infected individuals based on the estimate cortical thickness measurements. All analyses were adjusted for *Age* with its respective coefficient non-penalized. Consequently, X was an n by 2 matrix containing the column of ones (representing the intercept) and the column corresponding to subjects' age. Columns of design matrices (other than intercept) were centered to zero mean and normalized to unit standard deviation before the estimation. The selection of regularization parameter in logistic ridge was done within the GLMM framework. For all methods we used

Bootstrap-based approach with 50,000 samples, to define the subset of statistically significant variables.

6.3. Results

The estimates obtained from the logistic ridge and the griPEER$_{\text{boot}}$ for considered groups of subjects are presented in Figure 10. Brain regions labeled as response-related are marked with solid red vertical lines. In Table 2, we summarize the estimated values corresponding to brain regions being labeled as response-related by at least one considered approach. Note that all significant associations are negative, indicating thinner cortical areas are indicative of HIV-positive status. Significant estimates obtained from the griPEER for both types of connectivity matrices (FA- and DC-based) agree in 7 out of 8 cortical brain regions, while the logistic ridge significant findings disagree with the FA-based griPEER estimates in 4 regions and with the DC-based griPEER estimates in 3 regions.

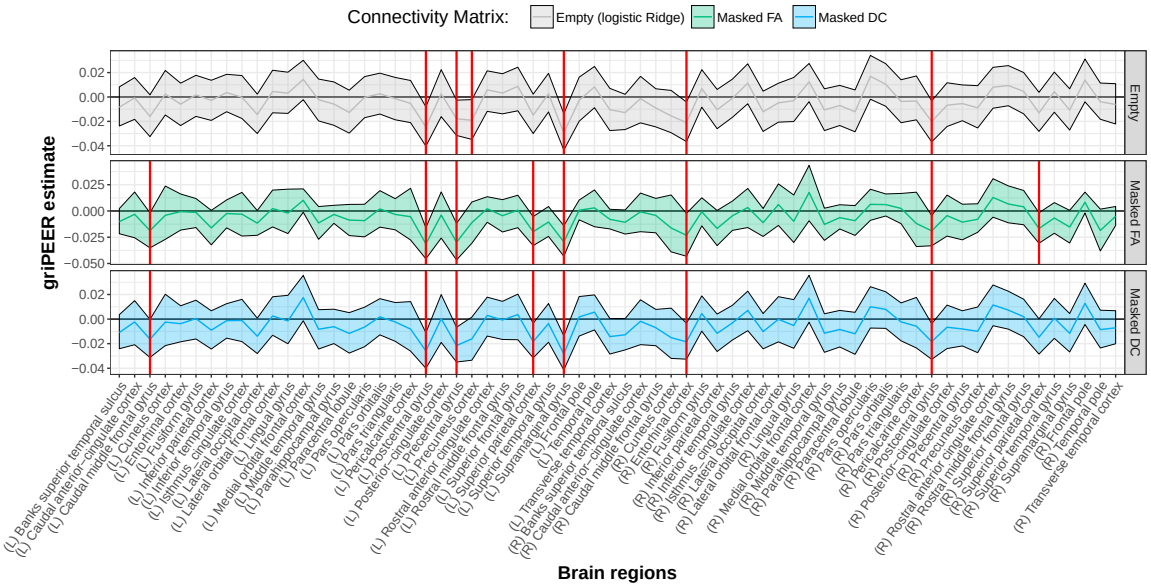


Figure 10: Results obtained by griPEER$_{\text{boot}}$ on 162 subjects (with 108 HIV-infected). Here, the response variable was defined as the disease indicator and 66 cortical brain regions were considered – 33 from the left and 33 from the right hemisphere. Regions labeled as response-related were marked by red vertical lines. Confidence intervals were calculated based on 50,000 bootstrap samples.

7. Discussion

We have provided a rigorous and computationally feasible method which incorporates additional information to estimate regression parameters in the generalized linear model setting. Our proposed method, griPEER, extends our work performed

Connectivity type	Caudal middle frontal [L]	Post central [L]	Pre central [L]	Pre cuneus [L]	Superior parietal [L]	Supra marginal [L]	Entorhinal [R]	Post central [R]	Superior parietal [R]
Empty	-0.016	**-0.025**	**-0.018**	**-0.019**	-0.015	**-0.029**	**-0.021**	**-0.020**	-0.013
Masked FA	**-0.019**	**-0.031**	**-0.030**	-0.011	**-0.020**	**-0.029**	**-0.023**	**-0.019**	**-0.017**
Masked DC	**-0.016**	**-0.026**	**-0.022**	-0.016	**-0.018**	**-0.028**	**-0.018**	**-0.018**	-0.015

Table 2: Estimates of the cortical brain regions coefficients obtained by the logistic ridge and griPEER$_{\text{boot}}$ with two different connectivity matrices – fractional anisotropy (Masked FA) and density of connections (Masked DC). Both matrices were masked by the modularity matrix before the analysis. Values corresponding to regions being labeled as response-related are shown in **bold, green font** and non-significant findings are shown using the red font. We show the results for all regions being selected by at least one method as response-related.

in the linear model setting Karas et al. (2017). We utilize the structural connectivity information obtained from the DTI to inform the association between the cortical covariates and a generalized outcome (e.g. binary indicator of HIV-infection). The structural connectivity information is used to create a Laplacian matrix, which in turn is used to specify the regularization penalty.

The simulation study shows that in each scenario considered, the proposed method, griPEER, outperforms logistic ridge in a binomial model coefficient estimation – griPEER yields smaller or similar estimation relative error $MSEr = \|\hat{b} - b\|_2^2 / \|b\|_2^2$ compared to the logistic ridge. Performance of griPEER is significantly better when the observed connectivity information is fully or largely informative about the true connectivity structure between model coefficients. Notably, even in cases when observed connectivity information is only partially informative or completely non-informative, the proposed method yields MSEr no larger than the logistic ridge estimator.

Application of griPEER to classify the individuals as HIV-infected and non-infected resulted in discovery of 3 additional cortical regions, namely Left Caudal Middle Frontal Gyrus, Left Superior Parietal Lobule and Right Superior Parietal Lobule, that were thinner in the HIV-infected individuals.

Our future work will incorporate both structural and functional brain connectivity information in the regularized estimation procedure. We will also include other properties of the cortex, namely the cortical area and its curvature.

Declaration of interest

The authors confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Acknowledgements

Research support was partially supported by the NIMH grants R01MH108467.

References

Bertero, M., Boccacci, P., 1998. Introduction to Inverse Problems in Imaging. Institute of Physics, Bristol, UK.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008 (10).

Breslow, N. E., Clayton, D. G., 1993. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88 (421), 9–25.

Cessie, S. L., Houwelingen, J. C. V., 1992. Ridge estimators in logistic regression. Journal of the Royal Statistical Society. Series C (Applied Statistics) 41 (1), 191–201.

Chung, F., 2005. Laplacians and the cheeger inequality for directed graphs. Annals of Combinatorics 9 (1), 1–19.

Cole, M. W., Bassett, D. S., Power, J. D., Braver, T. S., Petersen, S. E., 2014. Intrinsic and task-evoked network architectures of the human brain. Neuron 83 (1), 238–251.

Desikan, R. S., Segonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., Killiany, R. J., 2006. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. NeuroImage 31 (3), 968–80.

Engl, H. W., Hanke, M., Neubauer, A., 2000. Regularization of inverse problems. Kluwer, Dordrecht, Germany.

Fahrmeir, L., Kaufmann, H., 1985. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. Ann. Statist. 13 (1), 342–368.

Fischl, B., Aug. 2012. FreeSurfer. Neuroimage 62 (2), 774–81.
URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3685476/`

Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C. J., Wedeen, V. J., Sporns, O., 2008. Mapping the structural core of human cerebral cortex. PLoS Biol 6 (7), e159.

Hastie, T., Buja, A., Tibshirani, R., 1995. Penalized discriminant analysis. The Annals of Statistics 23 (1), 73–102.

Huang, J., Shen, H., Buja, A., 2008. Functional principal components analysis via penalized rank one approximation. Electronic Journal of Statistics 2, 678–695.

Karas, M., Brzyski, D., Dzemidzic, M., Goni, J., Kareken, D. A., Randolph, T. W., Harezlak, J., 2017. Brain connectivity–informed regularization methods for regression. Preprint available on the bioXiv: 10.1101/117945.

Li, C., Li, H., 2008. Network–constrained regularization and variable selection for analysis of genomic data. Bioinformatics 24 (9), 1175–1182.

Maldonado, Y. M., 2009. Mixed models, posterior means and penalized least-squares. Optimality 57, 216–236.

Phillips, D., 1962. A technique for the numerical solution of certain integral equations of the first kind. Journal of the ACM 9 (1), 84–97.

Pinheiro, J. C., Chao, E. C., 2006. Efficient laplacian and adaptive gaussian quadrature algorithms for multilevel generalized linear mixed models. Journal of Computational and Graphical Statistics 15 (1), 58–81.

Randolph, T. W., Harezlak, J., Feng, Z., 2012. Structured penalties for functional linear models – partially empirical eigenvectors for regression. Electronic Journal of Statistics 6, 323–353.

Slawski, M., Castell, W. Z., Tutz, G., 2010. Feature selection guided by structural information. Annals of Applied Statistics 4 (2), 1056–1080.

Sporns, O., 2013. Network attributes for segregation and integration in the human brain. Current opinion in neurobiology 23 (2), 162–171.

Sporns, O., Betzel, R. F., 2016. Modular brain networks. Annual review of psychology 67, 613.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B 58 (1), 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K., 2005. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B 67 (1), 91–108.

Tibshirani, R., Taylor, J., 2011. The solution path of the generalized lasso. The Annals of Statistics 39 (3), 1335–1371.

Tikhonov, A., 1963. Solution of incorrectly formulated problems and the regularization method. Soviet Math 4 (4), 1035–1038.

Wolfinger, R., O'connell, M., 1993. Generalized linear mixed models a pseudo-likelihood approach. Journal of Statistical Computation and Simulation 48 (3), 233–243.

Xin, B., Kawahara, Y., Wang, Y., Gao, W., 2016. Efficient generalized fused lasso and its applications. ACM Transactions on Intelligent Systems and Technology (TIST) 7 (4).

Zeger, S., Karim, M. R., 1991. generalized linear models with random effects - a gibbs sampling approach. Journal of the American Statistical Association 86 (413), 79–86.

Zhao, S., Shojaie, A., 2016. A significance test for graph-constrained estimations. Biometrics 72 (2), 484–493.

Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101 (476), 1418–1429.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B 67 (2), 301–320.

A. Appendix

A.1. Proof of proposition 3.1

The claim quickly follows from two well known linear algebra theorems: Woodbury identity and matrix determinant lemma. We recall the both results below as Theorem A.1.

Theorem A.1. *Suppose that A and C are invertible n by n matrices and U, V are n by p matrices. Then*

$$\textbf{\textit{I.1}} \quad \det\left(A+UV^{\mathsf{T}}\right) = \det\left(\mathrm{I}_p + V^{\mathsf{T}}A^{-1}U\right)\det\left(A\right),$$

$$\textbf{\textit{I.2}} \quad \left(A+UCV\right)^{-1} = A^{-1} - A^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}.$$

We will start with the proof of (C.1). Thanks to I.1,

$$\begin{aligned}\det\left(\overset{[\mathrm{k}]}{V_\lambda}\right) &= \det\left(\overset{[\mathrm{k}]}{W} + \overset{[\mathrm{k}]}{Z}\,\widetilde{Q}_\lambda^{-1}\,\overset{[\mathrm{k}]\mathsf{T}}{Z}\right) = \det\left(\mathrm{I}_p + \overset{[\mathrm{k}]\mathsf{T}}{Z}\,\overset{[\mathrm{k}]}{W}^{-1}\overset{[\mathrm{k}]}{Z}\,\widetilde{Q}_\lambda^{-1}\right)\det\left(\overset{[\mathrm{k}]}{W}\right)\\ &= \det\left(\widetilde{Q}_\lambda + \overset{[\mathrm{k}]\mathsf{T}}{Z}\,\overset{[\mathrm{k}]}{W}^{-1}\overset{[\mathrm{k}]}{Z}\right)\det\left(\widetilde{Q}_\lambda\right)^{-1}\det\left(\overset{[\mathrm{k}]}{W}\right).\end{aligned} \tag{25}$$

Now

$$\ln\det(\overset{[\mathrm{k}]}{V_\lambda}) = \ln\det\left(\lambda_Q Q + \lambda_R\,\mathrm{I}_p + \overset{[\mathrm{k}]}{\Omega}\right) - \ln\det\left(\lambda_Q Q + \lambda_R\,\mathrm{I}_p\right) + \ln\det\left(\overset{[\mathrm{k}]}{W}\right), \tag{26}$$

which finishes the proof (C.1). To show the second claim, we will rewrite $\overset{[\mathrm{k}]}{V_\lambda^{-1}}$ as

$$\overset{[\mathrm{k}]}{V_\lambda^{-1}} = \overset{[\mathrm{k}]}{W^{-1}} - \overset{[\mathrm{k}]}{W^{-1}}\overset{[\mathrm{k}]}{Z}\left(\lambda_Q Q + \lambda_R\,\mathrm{I}_p + \overset{[\mathrm{k}]}{\Omega}\right)^{-1}\overset{[\mathrm{k}]\mathsf{T}}{Z}\,\overset{[\mathrm{k}]}{W}^{-1}, \tag{27}$$

thanks to (I.2). Therefore

$$\overset{[\mathrm{k}]}{\widetilde{y}}^{\mathsf{T}}\,\overset{[\mathrm{k}]}{V_\lambda^{-1}}\,\overset{[\mathrm{k}]}{\widetilde{y}} = -\,\overset{[\mathrm{k}]\mathsf{T}}{q}\left(\lambda_Q Q + \lambda_R\,\mathrm{I}_p + \overset{[\mathrm{k}]}{\Omega}\right)^{-1}\overset{[\mathrm{k}]}{q} + \overset{[\mathrm{k}]}{\widetilde{y}}^{\mathsf{T}}\overset{[\mathrm{k}]}{W}^{-1}\overset{[\mathrm{k}]}{\widetilde{y}}. \tag{28}$$

A.2. Gradient and Hessian for the objective in (18)

Denote by $h(\lambda_Q,\lambda_R)$ the objective function of interest, i.e.

$$h(\lambda_Q,\lambda_R) := \ln\det\left(\lambda_Q Q+\lambda_R\,\mathrm{I}_p+\Omega\right)-\ln\det\left(\lambda_Q Q+\lambda_R\,\mathrm{I}_p\right)-q^{\mathsf{T}}\left(\lambda_Q Q+\lambda_R\,\mathrm{I}_p+\Omega\right)^{-1}q, \tag{29}$$

where Ω and q were defined in the statement of proposition 3.1 ("[k]"s symbols were omitted for clarity). After using notations $D_\lambda := \left(\lambda_Q Q + \lambda_R\,\mathrm{I}_p + \Omega\right)^{-1}$ and $\widetilde{Q}_\lambda := \lambda_Q Q + \lambda_R\,\mathrm{I}_p$, this function takes the short form

$$h(\lambda_Q,\lambda_R) = \ln\det D_\lambda^{-1} - \ln\det\widetilde{Q}_\lambda - q^{\mathsf{T}}D_\lambda q. \tag{30}$$

To find the gradient and Hessian of h we will use the following well known formulas

Proposition A.2. *Suppose that A and B are p by p, symmetric, positive semi-definite matrices, ν is p is p-dimensional vector and $tA+sB$ is positive definite. Then it holds*

$$\textbf{\textit{II.1}} \quad \frac{\partial}{\partial t}\Big\{\ln\det\big(tA+sB\big)\Big\} = \operatorname{tr}\Big[\big(tA+sB\big)^{-1}A\Big],$$

$$\textbf{\textit{II.2}} \quad \frac{\partial^2}{\partial t\partial s}\Big\{\ln\det\big(tA+sB\big)\Big\} = -\operatorname{tr}\Big[\big(tA+sB\big)^{-1}A\big(tA+sB\big)^{-1}B\Big],$$

$$\textbf{\textit{II.3}} \quad \frac{\partial^2}{\partial t^2}\Big\{\ln\det\big(tA+sB\big)\Big\} = -\operatorname{tr}\Big[\Big(\big(tA+sB\big)^{-1}A\Big)^2\Big],$$

$$\textbf{\textit{II.4}} \quad \frac{\partial}{\partial t}\Big\{-\nu^{\mathsf{T}}\big(tA+sB\big)^{-1}\nu\Big\} = \nu^{\mathsf{T}}\big(tA+sB\big)^{-1}A\big(tA+sB\big)^{-1}\nu,$$

$$\textbf{\textit{II.5}} \quad \frac{\partial^2}{\partial t\,\partial s}\Big\{-\nu^{\mathsf{T}}\big(tA+sB\big)^{-1}\nu\Big\} = -\nu^{\mathsf{T}}\big(tA+sB\big)^{-1}A\big(tA+sB\big)^{-1}B\big(tA+sB\big)^{-1}\nu - \nu^{\mathsf{T}}\big(tA+sB\big)^{-1}B\big(tA+sB\big)^{-1}A\big(tA+sB\big)^{-1}\nu,$$

$$\textbf{\textit{II.6}} \quad \frac{\partial^2}{\partial t^2}\Big\{-\nu^{\mathsf{T}}\big(tA+sB\big)^{-1}\nu\Big\} = -2\nu^{\mathsf{T}}\big(tA+sB\big)^{-1}A\big(tA+sB\big)^{-1}A\big(tA+sB\big)^{-1}\nu.$$

Thanks to the above, we quickly get

$$\nabla h\big|_{\lambda=\lambda_0} = \begin{bmatrix} \operatorname{tr}\big[(D_{\lambda_0}-\widetilde{Q}_{\lambda_0}^{-1})Q\big] + q^{\mathsf{T}}D_{\lambda_0}QD_{\lambda_0}q \\ \operatorname{tr}\big[D_{\lambda_0}-\widetilde{Q}_{\lambda_0}^{-1}\big] + q^{\mathsf{T}}D_{\lambda_0}^2q \end{bmatrix}. \tag{31}$$

and

$$\mathbf{H}\big(h\big)\big|_{\lambda=\lambda_0} = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{bmatrix}, \tag{32}$$

where

$$\begin{aligned} \mathbf{H}_{11} &:= -\operatorname{tr}\big[D_{\lambda_0}QD_{\lambda_0}Q - \widetilde{Q}_{\lambda_0}^{-1}Q\widetilde{Q}_{\lambda_0}^{-1}Q\big] - 2q^{\mathsf{T}}D_{\lambda_0}QD_{\lambda_0}QD_{\lambda_0}q, \\ \mathbf{H}_{22} &:= -\operatorname{tr}\big[D_{\lambda_0}^2 - \widetilde{Q}_{\lambda_0}^{-2}\big] - 2q^{\mathsf{T}}D_{\lambda_0}^3q, \\ \mathbf{H}_{12} = \mathbf{H}_{21} &:= -\operatorname{tr}\big[(D_{\lambda_0}^2 - \widetilde{Q}_{\lambda_0}^{-2})Q\big] - q^{\mathsf{T}}D_{\lambda_0}QD_{\lambda_0}^2q - q^{\mathsf{T}}D_{\lambda_0}^2QD_{\lambda_0}q. \end{aligned}$$

A.3. Asymptotic confidence interval

We start with the optimization problem equivalent to 20, with the objective multiplied by $\frac{1}{2}$,

$$\operatorname*{argmin}_{B\in\mathbb{R}^{p+m}} \Big\{ \underbrace{\textstyle\sum_i \psi(\mathcal{X}_iB) - y^{\mathsf{T}}\mathcal{X}B + \frac{1}{2}B^{\mathsf{T}}\mathcal{Q}B}_{\ell(B)} \Big\}. \tag{33}$$

Calculating the derivatives of ℓ yields

$$\frac{\partial\ell}{\partial B}(B) = \mathcal{X}^{\mathsf{T}}\psi'(\mathcal{X}B) - \mathcal{X}^{\mathsf{T}}y + \mathcal{Q}B \quad \text{and} \quad \frac{\partial^2\ell}{\partial B^2}(B) = \mathcal{X}^{\mathsf{T}}\Psi_{\mathcal{X}B}\mathcal{X} + \mathcal{Q}, \tag{34}$$

where

$$\begin{cases} \psi'(\mathcal{X}B) := \left[\, \psi'(\mathcal{X}_1 B), \ldots, \psi'(\mathcal{X}_n B) \,\right]^{\mathsf{T}} \\ \Psi_{\mathcal{X}B} := \operatorname{diag}\left\{ \psi''(\mathcal{X}_1 B), \ldots, \psi''(\mathcal{X}_n B) \right\} \end{cases} . \tag{35}$$

Denote by B_{T} the true signal and consider the Taylor series expansion of $\frac{\partial \ell}{\partial B}$ about B_{T}. If we consider the value of Taylor polynomial in the solution of (33), $\hat{B}$, this yields the following expression

$$\frac{\partial \ell}{\partial B}(\hat{B}) = \frac{\partial \ell}{\partial B}(B_{\mathrm{T}}) \; + \; (\hat{B} - B_{\mathrm{T}})^{\mathsf{T}} \frac{\partial^2 \ell}{\partial B^2}(B_{\mathrm{T}}) \; + \; o\big(\|\hat{B} - B_{\mathrm{T}}\|_2^2\big) \tag{36}$$

Since the left-hand side of the above equals zero, using (34) we get the first-order approximation of $\hat{B}$

$$\begin{aligned} \hat{B} = \;\; & B_{\mathrm{T}} - \left[\tfrac{\partial^2 \ell}{\partial B^2}(B_{\mathrm{T}})\right]^{-1} \tfrac{\partial \ell}{\partial B}(B_{\mathrm{T}}) = \\ & B_{\mathrm{T}} - \left[\mathcal{X}^{\mathsf{T}} \Psi_{\mathcal{X}B_{\mathrm{T}}} \mathcal{X} + \mathcal{Q}\right]^{-1} \left[\mathcal{X}^{\mathsf{T}} \psi'(\mathcal{X}B_{\mathrm{T}}) - \mathcal{X}^{\mathsf{T}} y + \mathcal{Q}B\right] = \\ & \left[\mathcal{X}^{\mathsf{T}} \Psi_{\mathcal{X}B_{\mathrm{T}}} \mathcal{X} + \mathcal{Q}\right]^{-1} \left[\big(\mathcal{X}^{\mathsf{T}} \Psi_{\mathcal{X}B_{\mathrm{T}}} \mathcal{X} + \mathcal{Q}\big) B_{\mathrm{T}} - \mathcal{X}^{\mathsf{T}} \psi'(\mathcal{X}B_{\mathrm{T}}) + \mathcal{X}^{\mathsf{T}} y - \mathcal{Q}B_{\mathrm{T}}\right] = \\ & \left[\mathcal{X}^{\mathsf{T}} \Psi_{\mathcal{X}B_{\mathrm{T}}} \mathcal{X} + \mathcal{Q}\right]^{-1} \left[\mathcal{X}^{\mathsf{T}} \Psi_{\mathcal{X}B_{\mathrm{T}}} \mathcal{X} B_{\mathrm{T}} - \mathcal{X}^{\mathsf{T}} \psi'(\mathcal{X}B_{\mathrm{T}}) + \mathcal{X}^{\mathsf{T}} y\right] = \\ & \left[\mathcal{X}^{\mathsf{T}} \Psi_{\mathcal{X}B_{\mathrm{T}}} \mathcal{X} + \mathcal{Q}\right]^{-1} \mathcal{X}^{\mathsf{T}} \Psi_{\mathcal{X}B_{\mathrm{T}}} \mathcal{X} \hat{B}^0, \end{aligned} \tag{37}$$

where $\hat{B}^0 := B_{\mathrm{T}} + \left[\mathcal{X}^{\mathsf{T}} \Psi_{\mathcal{X}B_{\mathrm{T}}} \mathcal{X}\right]^{-1} \Big(\mathcal{X}^{\mathsf{T}} y - \mathcal{X}^{\mathsf{T}} \psi'(\mathcal{X}B_{\mathrm{T}})\Big)$ is the first-order approximation of the generalized linear model estimate, i.e. for $\mathcal{Q} = 0$. It was shown that, under some regularity conditions, this estimate is unbiased and asymptotic normal (Fahrmeir and Kaufmann, 1985). The corresponding asymptotic variance is $\left[\mathcal{X}^{\mathsf{T}} \Psi_{\mathcal{X}B_{\mathrm{T}}} \mathcal{X}\right]^{-1}$. Consequently, the asymptotic variance , var_a, of $\hat{B}$ is given by

$$var_a(\hat{B}) = \left[\mathcal{X}^{\mathsf{T}} \Psi_{\mathcal{X}B_{\mathrm{T}}} \mathcal{X} + \mathcal{Q}\right]^{-1} \mathcal{X}^{\mathsf{T}} \Psi_{\mathcal{X}B_{\mathrm{T}}} \mathcal{X} \left[\mathcal{X}^{\mathsf{T}} \Psi_{\mathcal{X}B_{\mathrm{T}}} \mathcal{X} + \mathcal{Q}\right]^{-1}. \tag{38}$$

B. Appendix

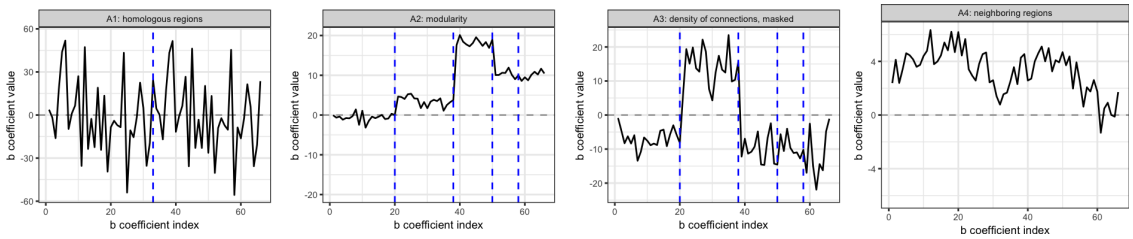


Figure 11: Examplary vectors of b model coefficients generated as $b \sim \mathcal{N}(0, \sigma_b^2(Q^{true})^{-1})$, where Q^{true} is Laplacian matrix of $\mathcal{A}^{true}$ graph adjacency matrix. Clearly, b coefficient values reflect the connectivity structure represented by $\mathcal{A}^{true}$ matrices assumed in the simulation study; left plot: $\mathcal{A}_1$ *"homologous regions"*, middle left plot: $\mathcal{A}_2$ *"modularity"*, middle right plot: $\mathcal{A}_3$ *"density of connections, masked"*, right plot: $\mathcal{A}_4$ *"neighboring regions"* (see: Fig. 2). In the left plot, vertical dashed line marks the separation between coefficients corresponding to left hemisphere brain regions and right remishpere brain regions assumed in $\mathcal{A}_1$ *"homologous regions"* construction. In the middle left plot, vertical dashed lines mark the separation between connectivity modules assumed in $\mathcal{A}_2$ *"homologous regions"* construction. In the middle right plot, vertical dashed lines mark the separation between connectivity modules assumed in $\mathcal{A}_3$ *"homologous regions"* construction.