Testing for Differences in Gaussian Graphical Models: Applications to Brain Connectivity

Eugene Belilovsky, Gaël Varoquaux, Matthew B. Blaschko

Hypothesis Testing for Differences in Gaussian Graphical Models: Applications to Brain Connectivity

Eugene Belilovsky Gaël Varoquaux Matthew B. Blaschko

***Abstract*—Functional brain networks are well described and estimated from data with Gaussian Graphical Models (GGMs), e.g. using sparse inverse covariance estimators. Comparing functional connectivity of subjects in two population calls for comparing these estimated GGMs. We study the problem of identifying differences in Gaussian Graphical Models (GGMs) known to have similar structure. We aim to characterize the uncertainty of differences with confidence intervals obtained using a parametric distribution on parameters of a sparse estimator. Sparse penalties enable statistical guarantees and interpretable models even in high-dimensional and low-number-of-samples settings. Quantifying the uncertainty of the parameters selected by the sparse penalty is an important question in applications such as neuroimaging or bioinformatics. Indeed, selected variables can be interpreted to build theoretical understanding or to make therapeutic decisions. Characterizing the distributions of sparse regression models is inherently challenging since the penalties produce a biased estimator. Recent work has shown how one can invoke the sparsity assumptions to effectively remove the bias from a sparse estimator such as the lasso. These distributions can be used to give us confidence intervals on edges in GGMs, and by extension their differences. However, in the case of comparing GGMs, these estimators do not make use of any assumed joint structure among the GGMs. Inspired by priors from brain functional connectivity we focus on deriving the distribution of parameter differences under a joint penalty when parameters are known to be sparse in the difference. This leads us to introduce the debiased multi-task fused lasso. We show that we can debias and characterize the distribution in an efficient manner. We then go on to show how the debiased lasso and multi-task fused lasso can be used to obtain confidence intervals on edge differences in Gaussian graphical models. We validate the techniques proposed on a set of synthetic examples as well as neuro-imaging dataset created for the study of autism.**

I. Introduction

Interactions in many real-world systems are well described via Gaussian Graphical Models. For instance, brain interactions between distant regions are revealed by correlations in brain activity, a process know as *functional connectivity*. Functional connectivity is an interesting probe on brain mechanisms as it persists in the absence of tasks (the so-called "resting-state") [3] and is thus applicable to study populations of impaired subjects, as in neurologic or psychiatric diseases [16]. From a formal standpoint, Gaussian Graphical models are well suited to estimate brain connections from functional Magnetic Resonance Imaging (fMRI) signals [37, 43]. A set of brain regions and related functional connections is then called a functional *connectome* [6]. Its variation across subjects can capture cognition [33, 35] or pathology [23, 41, 8]. However, the effects of pathologies are often very small, as resting-state fMRI is a weakly-constrained and noisy imaging modality, and the number of subjects in a study is often small given the cost of imaging. Statistical power is then a major concern [7]. In terms of development of statistical methods, the challenge is then to increase the power to detect differences between Gaussian Graphical models in the small-sample regime.

For such applications, estimation and comparison of Gaussian graphical models fall in the range of high-dimensional statistics: the number of degrees of freedom in the data is small compared to the dimensionality of the model. In this regime, sparsity-promoting ℓ_1-based penalties can make estimation well-posed and recover good estimation performance despite the scarcity of data [38, 14, 28, 11, 5, 1]. These encompass sparse regression methods such as the lasso or recovery methods such as basis pursuit, and can be applied to estimation of Gaussian graphical models with approaches such as the graphical lasso. There is now a wide body of literature which demonstrates the statistical properties of these methods [5]. Crucial to applications in medicine or neuroscience, a more recent line of work is interested in characterizing the uncertainty, in the form of confidence intervals and p-values, of the parameters selected by these methods [21, 22, 4, 25, 17]. These works focus primarily on the lasso and graphical lasso.

Approaches to estimate statistical significance on sparse models fall into several general categories: (a) re-sampling based methods which are inherently expensive and have difficult limiting distributions [5, 31, 10], (b) characterizations of the distribution of new parameters that enter a model along the solution path [25, 17], or (c) for a particular regularization parameter, debiasing the solution to obtain a new estimator for which the asymptotic distribution of the estimates is controlled [22, 21, 39].

Here, we consider the problem setting of two datasets known to have very similar underlying signals, but which individually may not be very sparse. A motivating example is determining the difference in brain networks of subjects from different groups: population analysis of connectomes [41, 23]. Recent literature in neuroscience [26] has suggested functional

E. Belilovsky is with CentraleSupélec, Université Paris-Saclay and Inria, Grande Voie des Vignes, 92295 Châtenay-Malabry, France

G. Varoquaux is with the Parietal team, Inria Saclay, 91191 Gif-Sur-Yvette, France

M. B. Blaschko is with the Center for Processing Speech and Images, Dept. Elektrotechniek, KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium

networks are not sparse. On the other hand, differences in connections across subjects should be sparse. Indeed the link between functional and anatomical brain networks [18] suggests they should not differ drastically from one subject to another. From a neuroscientific standpoint we are interested in determining which edges between two populations (e.g. autistic and non-autistic) are different. Furthermore we want to provide confidence-intervals on our results. We particularly focus on the setting where one dataset is larger than the other. In many applications it is more difficult to collect one group (e.g. individuals with specific pathologies) than another.

We introduce an estimator taylored to this goal: the debiased multi-task fused lasso. We show that, when the underlying parameter differences are indeed sparse, we can obtain a tractable Gaussian distribution for the parameter difference. This closed-form distribution underpins accurate hypothesis testing and confidence intervals.Using the relationship between nodewise regression and the inverse covariance matrix we show how our estimator can be used to learn the difference of Gaussian graphical models.

Although there is some recent work in providing graphical model selection with confidence intervals we are not aware of applications to the important problem of identifying differences in networks. In the setting of testing for differences, the confidence on the result is even more critical, as the differences are the direct outcome used for neuroscience research or medical practice, and it is important to provide the practitioner a measure of the uncertainty. Confidence intervals are standard practices of brain imaging, used e.g. in univariate analysis. Hence high dimensional extensions, as studied here, are easily understandable in this community.

The paper is organized as follows. In Section II we review previous work on joint parameter learning, the debiased lasso, and learning of GGMs. Section III discusses a joint debiasing procedure that specifically debiases the difference estimator. In Section IV we introduce the debiased multi-task fused lasso and show how it can be used to learn parameter differences in linear models. In Section V, we show how these results can be used for GGMs. In Section VI we validate our approach on synthetic and fMRI data.

II. Background and Related Work

In this section we discuss the general problem of joint parameter learning versus difference parameter learning. We review the common techniques for learning sparse GGMs. Finally we discuss the debiased lasso which we extend for the difference estimator problem in the sequel.

A. Joint Learning of Related Parameters and their Difference

In many applications one may be interested in learning multiple linear models from data that share many parameters. Situations such as this arise often in neuroimaging and bioinformatics applications. For example we may have brain networks from different subjects or groups of subjects which we assume *a priori* to have many shared parameter values. We can often improve the learning procedure of such models by incorporating fused penalties that penalize the $\|\cdot\|_1$ norm of the parameter differences or $\|\cdot\|_{1,2}$ which encourages groups of parameters to shrink together. These methods have been shown to substantially improve the learning of the joint models. However, the differences between model parameters are often pointed out only in passing [9, 11, 19].

On the other hand, in many situations we might be interested in actually understanding and identifying the differences between elements of the support. For example when considering brain networks of patients suffering from a pathology and healthy control subjects, the difference in brain connectivity may be of great interest. At the same time we might have very little data for certain populations. We argue that learning parameter differences under nodewise regression can have a high sample complexity, particularly when the true number of parameter differences is high relative to the support of the individual parameters.

In this work we limit ourselves to the case of two tasks (e.g. two groups of subjects), but the analysis can be easily extended to general multi-task settings. Consider the problem setting of data matrices $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$, which are $n_1 \times p$ and $n_2 \times p$, respectively. We model them as producing outputs Y_1 and Y_2, corrupted by noise ϵ_1 and ϵ_2 as follows

$$Y_1 = \boldsymbol{X_1}\beta_1 + \epsilon_1, \quad Y_2 = \boldsymbol{X_2}\beta_2 + \epsilon_2 \tag{1}$$

Let S_1 and S_2 index the elements of the support of β_1 and β_2, respectively. Furthermore the support of $\beta_1 - \beta_2$ is indexed by S_d and finally the union of S_1 and S_2 is denoted S_a. Using a squared loss estimator

$$\min_{\beta_1,\beta_2} \frac{1}{n_1}\|Y_1 - \boldsymbol{X_1}\beta_1\|^2 + \frac{1}{n_2}\|Y_2 - \boldsymbol{X_2}\beta_2\|^2 \tag{2}$$

we can obtain a difference estimate $\hat{\beta}_d = \hat{\beta}_1 - \hat{\beta}_2$. In general if S_d is very small relative to S_a then we will have a difficult time to identify the support S_d. This can be seen if we consider each of the individual components of the prediction errors. The larger the true support S_a the more it will drown out the subset which corresponds to the difference support. This is true even if use an ℓ_1 regularizer over the parameter vectors. Consequently, one cannot rely on the straightforward strategy of learning two independent estimates as in Equation (2) and taking their difference.

B. Gaussian Graphical Model Structure Learning

A standard approach to estimating Gaussian graphical models in high dimensions is to assume sparsity of the precision matrix and have an ℓ_0-norm constraint which limits the number of non-zero entries of the precision matrix. Since this leads to a combinatorial problem the ℓ_0-norm is often approximated with a convex ℓ_1-norm regularizer. This has led to the popular Graphical Lasso,

$$\min_{\boldsymbol{\Theta}} \log|\boldsymbol{\Theta}| - \mathrm{Tr}\left(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}\right) + \|\boldsymbol{\Theta}\|_1 \tag{3}$$

Which has been heavily studied [14]. Many variants of this approach that incorporate further structural assumptions have been proposed in the literature [34, 20, 19, 11, 29].

An alternative solution to approximating an ℓ_0 constraint on the precision matrix is neighborhood ℓ_1 regression from [28].

Here the authors make use of a long known property that connects the entries of the precision matrix to the problem of regression of one variable on all the others [27]. This property is critical to our proposed estimation procedure and we will further discuss this in Section II-C. This allows us to relate the regression models described in the previous section to finding edges connected to specific nodes in the GGM. We see that the discussion in the previous section still holds and in cases of weak sparsity but strong similarity we may not achieve a good sample complexity for finding the differences between the edge sets of two Gaussian graphical models.

GGM have been found good at recovering the main brain networks from fMRI data [37, 43]. Yet, recent work in neuroscience has showed that the structural wiring of the brain did not correspond to a very sparse network [26], thus questioning the underlying assumption of sparsity often used to estimate brain network connectivity. On the other hand, for the problem of finding differences between networks in two populations, sparsity may be a valid assumption. Indeed, it is well known that anatomical brain connections tend to closely follow functional ones [18]. Since anatomical networks do not differ drastically we can surmise that two brain networks should not differ much even in the presence of pathologies. The statistical method that we present here leverages sparsity, not in the networks themselves, but in the difference between two networks, to yield well-behaved estimation and hypothesis testing in the low-sample regime.

C. Background on Debiased Lasso

We review the idea of the debiased lasso and describe how it can be used to learn network differences. We start with the linear regression model with data matrix $\boldsymbol{X}$ and output Y, corrupted by ϵ noise as follows

$$Y = \boldsymbol{X}\beta + \epsilon \tag{4}$$

where we model ϵ as $N(0, \sigma_\epsilon^2 I)$. The lasso estimator is formulated as follows:

$$\hat{\beta}^\lambda = \arg\min_\beta \frac{1}{n}\|Y - \boldsymbol{X}\beta\|^2 + \lambda\|\beta\|_1 \tag{5}$$

We note that the differentiating and using the KKT conditions gives

$$\hat{k}^\lambda = \frac{1}{n}\boldsymbol{X}^T(Y - \boldsymbol{X}\beta) \tag{6}$$

where $\hat{k}$ is the subgradient of $\lambda\|\beta\|_1$

The debiased lasso estimator [39, 22] is then formulated as

$$\hat{\beta}_u^\lambda = \hat{\beta}^\lambda + \boldsymbol{M}\hat{k}^\lambda \tag{7}$$

for some $\boldsymbol{M}$ that is constructed to give guarantees on the asymptotic distribution of $\hat{\beta}_u^\lambda$.

We note the estimator above is not strictly unbiased in the finite sample case, but has a bias that rapidly approaches zero (w.r.t. n) if $\boldsymbol{M}$ is chosen appropriately, the true regressor β is indeed sparse, and the design matrix satistifes a certain restricted eigenvalue property [40, 39, 22]. Note that we can decompose this debiased estimator as follows:

$$\hat{\beta}_u^\lambda - \beta = \frac{1}{n}\boldsymbol{M}\boldsymbol{X}^T\epsilon - (\boldsymbol{M}\hat{\Sigma} - I)(\hat{\beta} - \beta) \tag{8}$$

The first term is Gaussian and the second term is responsible for the bias. Using Holder's inequality the second term can be bounded by $\|\boldsymbol{M}\hat{\Sigma} - I\|_\infty\|\hat{\beta} - \beta\|_1$. The first part of which we can bound using an appropriate selection of $\boldsymbol{M}$ while the second part is bounded by our implicit sparsity assumptions coming from lasso theory [5]. Two approaches from the recent literature discuss how one can select $\boldsymbol{M}$ to appropriately debias this estimate. In [39] it suffices to use nodewise regression to learn an inverse covariance matrix which guarantees constraints on $\|\boldsymbol{M}\hat{\Sigma} - I\|_\infty$. A second approach by [22] proposes to solve a quadratic program to directly minimize the variance of the debiased estimator while constraining $\|\boldsymbol{M}\hat{\Sigma} - I\|_\infty$ to induce sufficiently small bias.

Intuitively the construction of $\hat{\beta}_u^\lambda$ allows us to trade variance and bias via the $\boldsymbol{M}$ matrix. This allows us to overcome a naive bias-variance tradeoff by leveraging the sparsity assumptions that bound $\hat{\beta} - \beta$. In the sequel we expand this idea to the case of debiased parameter difference estimates and sparsity assumptions on the parameter differences.

The debiased lasso gives us an estimator that asymptotically converges to the partial correlations. As highlighted by [44] we can thus use the debiased lasso to obtain difference estimators with known finite sample distributions. This allows us to obtain confidence intervals on edge differences between Gaussian graphical models. We discuss this further in the sequel.

III. Debiased Difference Estimation

We start with the problem setting of data matrices $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, which are $n_1 \times p$ and $n_2 \times p$, respectively. We model them as producing outputs Y_1 and Y_2, corrupted by noise terms ϵ_1 and ϵ_2 as follows

$$Y_1 = \boldsymbol{X_1}\beta_1 + \epsilon_1, \quad Y_2 = \boldsymbol{X_2}\beta_2 + \epsilon_2 \tag{9}$$

Moreover we take the problem setting of an imbalanced dataset where $n_1 > n_2$ and possibly $n_1 \gg n_2$. We are interested in estimating the difference $\hat{\beta}_d = \hat{\beta}_1 - \hat{\beta}_2$, where $\hat{\beta}_1$ and $\hat{\beta}_2$ are regularized least squares estimates. In our problem setting we wish to obtain confidence intervals on debiased versions of these estimates in a high-dimensional setting (in the sense that $n_2 < p$), we aim to leverage assumptions about the form of the true β_d, primarily that it is sparse, while the independent $\hat{\beta}_1$ and $\hat{\beta}_2$ are weakly sparse or not sparse. We first consider a general case of any joint regularized least squares estimation of $\hat{\beta}_1$ and $\hat{\beta}_2$

$$\min_{\beta_1,\beta_2} \frac{1}{n_1}\|Y_1 - \boldsymbol{X_1}\beta_1\|^2 + \frac{1}{n_2}\|Y_2 - \boldsymbol{X_2}\beta_2\|^2 + R(\beta_1, \beta_2) \tag{10}$$

We note that the differentiating and using the KKT conditions gives

$$\hat{k}^\lambda = \hat{k}_1 + \hat{k}_2 = \frac{1}{n_1}\boldsymbol{X_1}^T(Y - \boldsymbol{X_1}\beta_1) + \frac{1}{n_2}\boldsymbol{X_2}^T(Y - \boldsymbol{X_2}\beta_2) \tag{11}$$

where $\hat{k}^\lambda$ is the (sub)gradient of $R(\beta_1, \beta_2)$. Substituting Equation (9) we can now write

$$\begin{aligned}\hat{\Sigma}_1(\hat{\beta}_1 - \beta_1) + \hat{k}_1 &= \frac{1}{n_1}\boldsymbol{X_1}^T\epsilon_1 \\ \hat{\Sigma}_2(\hat{\beta}_2 - \beta_2) + \hat{k}_2 &= \frac{1}{n_2}\boldsymbol{X_2}^T\epsilon_2\end{aligned} \tag{12}$$

We would like to solve for the difference $\hat{\beta}_1 - \hat{\beta}_2$ but the covariance matrices may not be invertible. We introduce matrices $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$, which will allow us to isolate the relevant term. We will see that in addition these matrices will allow us to decouple the bias and variance of the estimators.

$$\begin{aligned}\boldsymbol{M}_1\hat{\Sigma}_1(\hat{\beta}_1 - \beta_1) + \boldsymbol{M}_1\hat{k}_1 &= \frac{1}{n_1}\boldsymbol{M}_1\boldsymbol{X}_{\boldsymbol{1}}^T\epsilon_1 \\ \boldsymbol{M}_2\hat{\Sigma}_2(\hat{\beta}_2 - \beta_2) + \boldsymbol{M}_2\hat{k}_2 &= \frac{1}{n_2}\boldsymbol{M}_2\boldsymbol{X}_{\boldsymbol{2}}^T\epsilon_2 \end{aligned} \tag{13}$$

subtracting these we obtain

$$\begin{aligned}&\boldsymbol{M}_1\hat{\Sigma}_1(\hat{\beta}_1 - \beta_1) - \boldsymbol{M}_2\hat{\Sigma}_2(\hat{\beta}_2 - \beta_2) + \boldsymbol{M}_1\hat{k}_1 - \boldsymbol{M}_2\hat{k}_2 \\ &\quad= \frac{1}{n_1}\boldsymbol{M}_1\boldsymbol{X}_{\boldsymbol{1}}^T\epsilon_1 - \frac{1}{n_2}\boldsymbol{M}_2\boldsymbol{X}_{\boldsymbol{2}}^T\epsilon_2\end{aligned}$$

We can now isolate the estimator plus a term we add back controlled by $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$

$$(\hat{\beta}_1 - \hat{\beta}_2) - (\beta_1 - \beta_2) + \boldsymbol{M}_1\hat{k}_1 - \boldsymbol{M}_2\hat{k}_2 \tag{14}$$

$$= \frac{1}{n_1}\boldsymbol{M}_1\boldsymbol{X}_{\boldsymbol{1}}^T\epsilon_1 - \frac{1}{n_2}\boldsymbol{M}_2\boldsymbol{X}_{\boldsymbol{2}}^T\epsilon_2 - \Delta \tag{15}$$

$$\Delta = (\boldsymbol{M}_1\hat{\Sigma}_1 - I)(\hat{\beta}_1 - \beta_1) - (\boldsymbol{M}_2\hat{\Sigma}_2 - I)(\hat{\beta}_2 - \beta_2) \tag{16}$$

$$\begin{aligned}\Delta =& \frac{(\boldsymbol{M}_1\hat{\Sigma}_1 - I + \boldsymbol{M}_2\hat{\Sigma}_2 - I)}{2}((\hat{\beta}_1 - \hat{\beta}_2) - (\beta_1 - \beta_2)) \\ &+ \frac{(\boldsymbol{M}_1\hat{\Sigma}_1 - \boldsymbol{M}_2\hat{\Sigma}_2)}{2}((\hat{\beta}_1 - \beta_1) + (\hat{\beta}_2 - \beta_2))\end{aligned} \tag{17}$$

Looking at this equation closer we see that Δ will control the bias of our estimator. At the same time we would like to minimize its variance given by

$$\frac{1}{n_1}\boldsymbol{M}_1\hat{\Sigma}_1\boldsymbol{M}_1\hat{\sigma}_1^2 + \frac{1}{n_2}\boldsymbol{M}_2\hat{\Sigma}_2\boldsymbol{M}_2\hat{\sigma}_2^2. \tag{18}$$

We can now overcome the limitations of simple bias variance trade-off by using an appropriate regularizer coupled with an assumption on the underlying signal β_1 and β_2. This will in turn make Δ asymptotically vanish while maximizing the variance.

Since we are interested in pointwise estimates, we can focus on bounding the infinity norm of Δ. Denoting $\beta_d := \beta_1 - \beta_2$ and $\beta_a := \beta_1 + \beta_2$

$$\begin{aligned}\|\Delta\|_\infty \leq& \frac{1}{2}\underbrace{\|\boldsymbol{M}_1\hat{\Sigma}_1 + \boldsymbol{M}_2\hat{\Sigma}_2 - 2I\|_\infty}_{\mu_1}\underbrace{\|\hat{\beta}_d - \beta_d\|_1}_{l_d} \\ &+\frac{1}{2}\underbrace{\|\boldsymbol{M}_1\hat{\Sigma}_1 - \boldsymbol{M}_2\hat{\Sigma}_2\|_\infty}_{\mu_2}\underbrace{\|\hat{\beta}_a - \beta_a\|_1}_{l_a}\end{aligned} \tag{19}$$

We can control the maximum bias by selecting $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ appropriately. If we use an appropriate regularizer coupled with sparsity assumptions we can bound the terms l_a and l_d and use this knowledge to appropriately select $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$.

We now see that if the same assumptions on the individual signals hold then we can apply the results of the debiased lasso and estimate $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ as in the previous section. In the case of interest where β_1 and β_2 share many weights we can do better by taking this as an assumption and applying a sparsity regularization on the difference by adding the term $\lambda_2\|\beta_1 - \beta_2\|_1$. Comparing the decoupled penalty to the fused penalty proposed we see that l_d would decrease at a given sample size.

The question now is how to estimate $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ so that $\|\Delta\|_\infty$ becomes negligible for a given n, p and sparsity assumption.

IV. Debiasing the Multi-Task Fused Lasso

Motivated by the inductive hypothesis from neuroscience described above we introduce a consistent low-variance estimator the debiased multi-task fused lasso. We propose to use the following regularizer $R(\beta_1, \beta_2) = \lambda_1\|\beta_1\|_1 + \lambda_1\|\beta_2\|_1 + \lambda_2\|\beta_1 - \beta_2\|_1$. This penalty has been referred to in some literature as the multi task fused lasso [9]. We propose to then debias this estimate as shown in (14). We estimate the $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ matrices by solving the following QP for each row m_1 and m_2 of the matrices $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$.

$$\begin{aligned}\min_{m_1,m_2} &\frac{1}{n_1}m_1^T\hat{\Sigma}_1 m_1 + \frac{1}{n_2}m_2^T\hat{\Sigma}_2 m_2 \\ s.t. \quad &\|\boldsymbol{M}_1\hat{\Sigma}_1 + \boldsymbol{M}_2\hat{\Sigma}_2 - 2I\|_\infty \leq \mu_1 \\ &\|\boldsymbol{M}_1\hat{\Sigma}_1 - \boldsymbol{M}_2\hat{\Sigma}_2\|_\infty \leq \mu_2\end{aligned} \tag{20}$$

We set the constraints as follows

$$\mu_1 \leq \frac{1}{c\lambda_2 s_d n_2^{0.01}} \tag{21}$$

$$\mu_2 \leq \frac{1}{a(\lambda_1 s_{1,2} + \lambda_2 s_d)n_2^{0.01}} \tag{22}$$

Where s_d is the difference sparsity and $c > 1$ and $a > 1$. $s_{1,2}$ is the parameter sparsity $|S_1| + |S_2|$. We next show that under certain assumptions and given the constraints above the $\|\Delta\|_\infty = o(1)$ w.h.p.

Let $\beta = [\beta_1; \beta_2]$,$\beta_d = \beta_1 - \beta_2$,$\beta_a = \beta_1 + \beta_2$. Let $S_{1,2}$ bet the support of β. Define $\boldsymbol{X}_N = [\boldsymbol{X}_1/\sqrt{n_1}, 0; 0, \boldsymbol{X}_2/\sqrt{n_2}]$

Lemma 1. *(Basic Inequality)* $\|\boldsymbol{X}_N(\hat{\beta} - \beta)\|_2^2 + \lambda_1\|\hat{\beta}\|_1 + \lambda_2\|\hat{\beta}_d\| \leq 2\epsilon^T\boldsymbol{X}_N(\hat{\beta} - \beta) + \lambda_1\|\beta\|_1 + \lambda_2\|\beta_d\|_1$

This follows from the fact that $\hat{\beta}$ is the minimizer of the fused lasso objective.

The term, $\epsilon^T\boldsymbol{X}_N(\hat{\beta} - \beta)$, commonly known as the empirical process term [5] can be bound as follows:

$$2|\epsilon^T\boldsymbol{X}_N(\hat{\beta} - \beta)| = \tag{23}$$

$$2|\epsilon_1^T\boldsymbol{X}_1(\hat{\beta}_1 - \beta_1)/n_1 + \epsilon_2^T\boldsymbol{X}_2(\hat{\beta}_2 - \beta_2)/n_2| \leq \tag{24}$$

$$2\|\hat{\beta}_1 - \beta_1\|_1 \max_{1\leq j\leq p}|\epsilon_1^T\boldsymbol{X}_1^{(j)}/n_1|$$

$$+2\|\hat{\beta}_2 - \beta_2\|_1 \max_{1\leq j\leq p}|\epsilon_2^T\boldsymbol{X}_2^{(j)}/n_2| \tag{25}$$

Where we utilize holder's inequality in the last line.

We define the random event $\mathcal{F}$ for which the following holds: $\max_{1\leq j\leq p}|\epsilon_1^T\boldsymbol{X}_1^{(j)}/n_1| \leq \lambda_0$ and $\max_{1\leq j\leq p}|\epsilon_1^T\boldsymbol{X}_1^{(j)}/n_1| \leq \lambda_0$. furthermore we can select $2\lambda_0 \leq \lambda_1$

Lemma 2. *Suppose $\hat{\Sigma}_{j,j} = 1$ for both $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ then we have for all $t > 0$ and $n_1 > n_2$*

$$\lambda_0 = 2\sigma_2\sqrt{\frac{t^2 + \log p}{n_2}} \tag{26}$$

$$P(\mathcal{F}) = 1 - 2\exp(-t^2/2) \tag{27}$$

Proof. This follows directly from the [5, Lemma 6.2] and taking $n_1 > n_2$. □

This allows us to get rid of the empirical process term on $\mathcal{F}$, with an appropriate choice of λ_1.

Given a set, S, denote β_S the vector of equal size to β but all elements not in S set to zero. We can now show the following

Lemma 3. *We have on $\mathcal{F}$ with $\lambda_1 \geq 2\lambda_0$*

$$\begin{aligned} &2\|\boldsymbol{X}_N(\hat{\beta} - \beta)\|_2^2 + \lambda_1\|\hat{\beta}_{S_{1,2}^c}\|_1 + 2\lambda_2\|\hat{\beta}_{d,S_d^c}\|_1 \\ &\leq 3\lambda_1\|\hat{\beta}_{S_{1,2}} - \beta_{S_{1,2}}\|_1 + 2\lambda_2\|\hat{\beta}_{d,S_d} - \beta_{d,S_d}\|_1 \end{aligned} \tag{28}$$

Proof. Following [5, Lemma 6.3] we start with the basic inequality on $\mathcal{F}$. Which gives

$$\begin{aligned} &2\|\boldsymbol{X}_N(\hat{\beta} - \beta)\|_2^2 + 2\lambda_1\|\hat{\beta}\|_1 + 2\lambda_2\|\hat{\beta}_d\| \\ &\leq \lambda_1\|\hat{\beta} - \beta\|_1 + 2\lambda_1\|\beta\|_1 + 2\lambda_2\|\beta_d\|_1 \end{aligned} \tag{29}$$

Since we assume the truth is in fact sparse,

$$\|\hat{\beta}_d - \beta_d\|_1 = \|\hat{\beta}_{d,S_d} - \beta_{d,S_d}\|_1 + \|\hat{\beta}_{d,S_d^C}\|_1 \tag{30}$$

$$\|\hat{\beta} - \beta\|_1 = \|\hat{\beta}_{S_{1,2}} - \beta_{S_{1,2}}\|_1 + \|\hat{\beta}_{S_{1,2}^C}\|_1 \tag{31}$$

Furthermore,

$$\|\hat{\beta}\|_1 \geq \|\beta_{S_{1,2}}\|_1 - \|\hat{\beta}_{S_{1,2}} - \beta_{S_{1,2}}\|_1 + \|\hat{\beta}_{S_{1,2}^C}\|_1 \tag{32}$$

$$\|\hat{\beta}_d\|_1 \geq \|\beta_{d,S_d}\|_1 - \|\hat{\beta}_{d,S_d} - \beta_{d,S_d}\|_1 + \|\hat{\beta}_{d,S_d^C}\|_1 \tag{33}$$

Substituting (32), (33), and (31) into (29) and rearranging completes the proof. □

From the lemma above we can now justify the bounds in (21) and (22)

$$\begin{aligned} &\lambda_1\|\hat{\beta}_a - \beta_a\|_1 \leq \lambda_1\|\hat{\beta}_{S_{1,2}} - \beta_{S_{1,2}}\|_1 + \lambda_1\|\hat{\beta}_{S_{1,2}^c}\|_1 \leq \\ &4\lambda_1\|\hat{\beta}_{S_{1,2}} - \beta_{S_{1,2}}\|_1 + 2\lambda_2\|\hat{\beta}_{S_d} - \beta_{S_d}\|_1 - 2\|\boldsymbol{X}_N(\hat{\beta} - \beta)\|_2^2 \end{aligned} \tag{34}$$

$$\begin{aligned} &\leq 4\lambda_1\sqrt{s_{1,2}}\|\hat{\beta}_{S_{1,2}} - \beta_{S_{1,2}}\|_2 + 2\lambda_2\sqrt{s_d}\|\hat{\beta}_{S_d} - \beta_{S_d}\|_2 \\ &\quad -2\|\boldsymbol{X}_N(\hat{\beta} - \beta)\|_2^2 \end{aligned} \tag{35}$$

Invoking the compatibility assumption [5, 22, 15] with compatibility constant $\phi_{\min}$

$$\begin{aligned} &\leq \frac{4\lambda_1\sqrt{s_{1,2}}}{\phi_{min}}\|\boldsymbol{X}_N(\hat{\beta} - \beta)\|_2 + \frac{2\lambda_2\sqrt{s_d}}{\phi_{min}}\|\boldsymbol{X}_N(\hat{\beta} - \beta)\|_2 \\ &\quad -2\|\boldsymbol{X}_N(\hat{\beta} - \beta)\|_2^2 \end{aligned} \tag{36}$$

$$\leq \frac{4\lambda_1^2 s_{1,2}}{\phi_{min}^2} + \frac{2\lambda_2^2 s_d}{\phi_{min}^2} \tag{37}$$

Proposition 1. *When the compatability condition [5, 15] holds and with $\lambda_1 > 2\lambda_0$ as per Lemma 2. Taking $\lambda_2 = O(\lambda_1)$ and use of the bound prescribed* (22) *gives $l_a u_2 = o(1)$.*

This follows by inverting the bound shown in (37).

Now we consider the bound for l_d.

$$\lambda_2\|\hat{\beta}_d - \beta_d\|_1 = \lambda_2\|\hat{\beta}_{d,S} - \beta_{d,S}\|_1 + \lambda_2\|\hat{\beta}_{d,S^c}\|_1 \tag{38}$$

$$\leq 2\lambda_2\|\hat{\beta}_{d,S} - \beta_{d,S}\|_1 + 3\lambda_1\|\hat{\beta}_{S_{1,2}} - \beta_{S_{1,2}}\|_1/2 \tag{39}$$

$$-\|\boldsymbol{X}_N(\hat{\beta} - \beta)\|_2^2 - \lambda_1\|\hat{\beta}_{S_{1,2}^c}\|_1/2 \tag{40}$$

In the domain of interest $n_1 \gg n_2$ if we select $\lambda_2 = O(\lambda_1)$ we can see the relevant terms related to the parameter support become small with respect to terms with $S_{1,2}$. Thus the error on the difference should dominate. In this region we can have $3\lambda_1\|\hat{\beta}_{S_{1,2}} - \beta_{S_{1,2}}\|_1/2 - \lambda_1\|\hat{\beta}_{S_{1,2}^c}\|_1/2 \leq c\lambda_2\|\hat{\beta}_{d,S} - \beta_{d,S}\|_1$ where $c > 0$.

$$\lambda_2\|\hat{\beta}_d - \beta_d\|_1 \leq 2\lambda_2\|\hat{\beta}_{d,S} - \beta_{d,S}\|_1 - \|\boldsymbol{X}_N(\hat{\beta} - \beta)\|_2^2 \tag{41}$$

$$\leq 2c\lambda_2\sqrt{s_d}\|\hat{\beta}_{d,S} - \beta_{d,S}\|_2 - \|\boldsymbol{X}_N(\hat{\beta} - \beta)\|_2^2 \tag{42}$$

Invoking the compatibility assumption [5]

$$\leq 2c\lambda_2\sqrt{s_d}\|\boldsymbol{X}_N(\hat{\beta} - \beta)\|_2/\phi_{min} - \|\boldsymbol{X}_N(\hat{\beta} - \beta)\|_2^2 \tag{43}$$

$$\leq \frac{c^2\lambda_2^2 s_d}{\phi_{min}^2} + \|\boldsymbol{X}_N(\hat{\beta} - \beta)\|_2^2 - \|\boldsymbol{X}_N(\hat{\beta} - \beta)\|_2^2 \tag{44}$$

Thus when the assumptions hold $\|\hat{\beta}_d - \beta_d\|_1 \leq \frac{c^2\lambda_2 s_d}{\phi_{min}^2}$ and use of the bound prescribed (21) gives $l_d u_1 = o(1)$.

Therefore using the prescribed $\boldsymbol{M}$s obtained with (20) we obtain an unbiased estimator given by (14) the difference with asymptotic variance (18)

V. GGM Difference Structure Discovery with Significance

In this section we discuss how the debiased lasso and debiased multi-task fused lasso proposed in the previous sections can be used to learn the structure of a difference of Gaussian graphical models and to provide significance results on the presence of edges within the difference graph. We refer to these two procedures as Difference of Neighborhoods Debiased Lasso Selection and Difference of Neighborhoods Debiased Fused Lasso Selection.

We recall that the conditional independence properties of a GGM are given by the zeros of the precision (inverse covariance) matrix. Furthermore it is known that these zeros correspond to the regression parameters when regressing one variable on the other. Thus a natural extension of the debiased lasso to estimating a de-biased precision matrix appears when we consider the relationship between neighborhood regressions and the precision matrices. By obtaining a debiased lasso estimate for each node in the graph [44] notes this leads to a sparse unbiased precision matrix estimate with a known asymptotic distribution. Subtracting these estimates for two different datasets gives us a difference estimate. This leads to an estimator whose zeros correspond to difference of graph edges in GGM with a known asymptotic distribution. We can similarly use the debiased multi-taks fused lasso described above and the joint debiasing procedure to obtain a test statistic for the difference of networks.

We now formalize this procedure. Given GGM's $j = 1, 2$. Let $\boldsymbol{X}_j$ denote the random variable $\mathbf{R}^p$ associated with GGM

j. We denote $\boldsymbol{X}_j, v$ the random variable associated with a node, v of the GGM and $\boldsymbol{X}_{j,v^c}$ all other nodes in the graph. For notational convenience we will often denote the node of interest Y_j, v instead of $\boldsymbol{X}_j, v$.

We denote $\hat{\beta}_{j,v}$ the lasso or multi-task debiased lasso estimate of $\boldsymbol{X}_{j,v^c}$ onto Y_j, v, then $\hat{\beta}_{j,dL,v}$ the debiased lasso. Finally let $\beta_{j,v}$ denote the unknown regression, $Y_j, v = \boldsymbol{X}_{j,v^c}\beta_{j,v} + \epsilon_j$ where $\epsilon_j \sim \mathbf{N}(0, \sigma_{\epsilon_j}\mathbf{I})$. We denote the row of the precision matrix $\boldsymbol{\Theta}_j$ corresponding to v as $\boldsymbol{\Theta}_{j,v}$.

Define $\beta^i_{D,v}$ the test statistic associated with the edge v, i in the difference of GGM's $j = 1, 2$,

$$\hat{\beta}^i_{D,s} = \hat{\beta}^i_{1,dL,s} - \hat{\beta}^i_{2,dL,s} \tag{45}$$

Proposition 2. *Given the $\hat{\beta}^i_{D,v}$, $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ computed as in [22] for the lasso or as in Section IV for the multi-task fused lasso. When the respective assumptions of these estimators are satisfied the following holds w.h.p.*

$$\hat{\beta}^i_{D,v} - \beta^i_{D,v} = W + o(1) \tag{46}$$

$$W \sim \mathbf{N}(0, [\hat{\sigma}^2_1 \boldsymbol{M}_1 \hat{\Sigma}_1 \boldsymbol{M}_1^T + \hat{\sigma}^2_2 \boldsymbol{M}_2 \hat{\Sigma}_2 \boldsymbol{M}_2^T]_{i,i}) \tag{47}$$

The proof of this follows directly from the asympotitc consistency of each individual component of the partial correlation matrix, $\boldsymbol{B}$. Zeros of $\boldsymbol{B}$ correspond to zeros of the precision matrix and indicate conditional independence amongst variables.

We can now define the the null hypothesis of interest as follows

$$H_0 : \boldsymbol{\Theta}_{1,(i,j)} = \boldsymbol{\Theta}_{2,(i,j)} \tag{48}$$

Obtaining a test statistic for each element $\beta^i_{D,v}$ allows us to perform hypothesis testing on individual edges, all the edges, or groups of edges (controlling for the FWER). We summarize the Neighbourhood Debiased Lasso Selection process in Algorithm 1 and the Neighbourhood Debiased Multi-Task Fused Lasso Selection in Algorithm 2 which can be used to obtain a matrix of all the relevant test statistics.

Algorithm 1 Difference Network Selection with Neighborhood Debiased Lasso

$V = \{1, ..., P\}$
NxP Data Matrices, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$
Px(P-1) Output Matrix, $\boldsymbol{B}$
for $v \in V$ **do**
  Estimate unbiased $\hat{\sigma}_1, \hat{\sigma}_2$ from $X_{1,v}, X_{2,v}$
  for $j \in \{1, 2\}$ **do**
    $\beta_j \leftarrow SolveLasso(\boldsymbol{X}_{j,v^c}, X_{j,v})$
    $\boldsymbol{M}_j \leftarrow MEstimator(\boldsymbol{X}_{j,v^c})$
    $\beta_{j,U} \leftarrow \beta_j$+$\boldsymbol{M}_j \boldsymbol{X}^T_{j,v^c}(X_{j,v} - \boldsymbol{X}_{j,v^c}\beta_j)$
  end for
  $Var \leftarrow diag(\frac{\hat{\sigma}^2_1}{n_1}\boldsymbol{M}_1^T\hat{\Sigma}_1\boldsymbol{M}_1 + \frac{\hat{\sigma}^2_2}{n_2}\boldsymbol{M}_2^T\hat{\Sigma}_2\boldsymbol{M}_2)$
  for $j \in v^c$ **do**
    $\boldsymbol{B}_{v,j} = (\beta_{1,U,j} - \beta_{2,U,j})/\sqrt{Var_j}$
  end for
end for
$\boldsymbol{B}$ is a matrix of test statistics, which are $N(0, 1)$ under H_0

Algorithm 2 Difference Network Selection with Neighborhood Debiased Fused Lasso

$V = \{1, ..., P\}$
NxP Data Matrices, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$
Px(P-1) Output Matrix $\boldsymbol{B}$
for $v \in V$ **do**
  Estimate unbiased $\hat{\sigma}_1, \hat{\sigma}_2$ from $X_{1,v}, X_{2,v}$
  β_1,$\beta_2 \leftarrow SolveFusedLasso(\boldsymbol{X}_{1,v^c}, X_{1,v}, \boldsymbol{X}_{2,v^c}, X_{2,v})$
  $\boldsymbol{M}_1, \boldsymbol{M}_2 \leftarrow MEstimator(X_{1,v^c}, \boldsymbol{X}_{2,v^c})$
  for $j \in \{1, 2\}$ **do**
    $\beta_{j,U} \leftarrow \beta_j$+$\boldsymbol{M}_j \boldsymbol{X}^T_{j,v^c}(X_{j,v} - \boldsymbol{X}_{j,v^c}\beta_j)$
  end for
  $Var \leftarrow diag(\frac{\hat{\sigma}^2_1}{n_1}\boldsymbol{M}_1^T\hat{\Sigma}_1\boldsymbol{M}_1 + \frac{\hat{\sigma}^2_2}{n_2}\boldsymbol{M}_2^T\hat{\Sigma}_2\boldsymbol{M}_2)$
  for $j \in v^c$ **do**
    $\boldsymbol{B}_{v,j} = (\beta_{1,U,j} - \beta_{2,U,j})/\sqrt{Var_j}$
  end for
end for
$\boldsymbol{B}$ is a matrix of test statistics, which are $N(0, 1)$ under H_0

VI. Experiments

To numerically validate and compare our method we perform experiments on synthetic data generated from similar GGM as well as a real fMRI resting state dataset.

A. Simulations

We generate synthetic data based on two Gaussian graphical models with 75 vertices. Each of the individual graphs have a sparsity of 19% and their difference sparsity is 3%. We construct the models by taking two identical precision matrices and randomly removing some edges from both. We generate synthetic data using both precision matrices. We use $n_1 = 800$ samples for the first dataset and vary the second dataset $n_2 = 20, 30, ...150$.

We perform a regression using the debiased lasso and the debiased multi-task fused lasso on each node of the graphs. As an extra baseline we consider the projected ridge method from the R package "hdi" [12]. We use the debiased lasso of [22] where we set $\lambda = k\hat{\sigma}\sqrt{\log p/n}$. We select c by 3-fold cross validation $k = \{0.1, ..100\}$ and $\boldsymbol{M}$ as prescribed in [22] which we obtain by solving a quadratic program. $\hat{\sigma}$ is an unbiased estimator of the noise variance. For the debiased lasso we let both $\lambda_1 = k_1\hat{\sigma_2}\sqrt{\log p/n_2}$ and $\lambda_2 = k_2\hat{\sigma_2}\sqrt{\log p/n_2}$, and select based on 3-fold cross-validation from the same range as k. $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are obtained as in Equation (20) with the bounds (21) and (22) being set with $c = a = 2$, $s_d = 2, s_{1,2} = 15$, and the cross validated λ_1 and λ_2. In both debiased lasso and fused multi-task lasso cases we utilize the Mosek QP solver package to obtain M. For the projected ridge method we use the hdi package to obtain two estimates of β_1 and β_2 along with their upper bounded biases which are then used to obtain p-values for the difference.

We report the false positive rate, the power, the coverage and interval length as per [39] for the difference of graphs. In these experiments we aim to aggregate statistics to demonstrate power of the test statistic as such we consider each edge as a separate test and do not attempt to control the family-wise

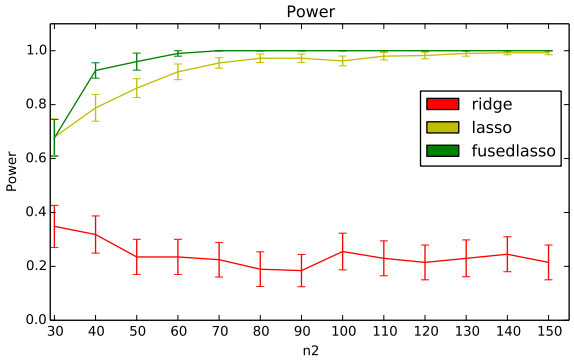

Fig. 1. With n_1 fixed at 800, we show the power of the test for different number of samples in the second data set. The statistical power of debiased fused lasso is highest

Method	FP	TP(Power)	Cov S	Cov S_d^c	len S	len S_d^c
Deb. Lasso	3.7%	80.6%	96.2%	92%	2.199	2.195
Deb. Fused Lasso	0.0%	93.3%	100%	98.6%	2.191	2.041
Ridge Projection	0.0%	18.6%	100%	100%	5.544	5.544

TABLE I
COMPARISON OF DEBIASED LASSO, DEBIASED FUSED LASSO, AND PROJECTED RIDGE REGRESSION FOR EDGE SELECTION IN DIFFERENCE OF GGM. THE SIGNIFICANCE LEVEL IS 5%, $n_1 = 800$ AND $n_2 = 60$. ALL METHODS HAVE FALSE POSITIVE BELOW THE SIGNIFICANCE LEVEL AND THE DEBIASED FUSED LASSO DOMINATES IN TERMS OF POWER. THE COVERAGE OF THE DIFFERENCE SUPPORT AND NON-DIFFERENCE SUPPORT IS ALSO BEST FOR THE DEBIASED FUSED LASSO, WHICH SIMULTANEOUSLY HAS SMALLER CONFIDENCE INTERVALS ON AVERAGE.

error rate. We show the numerical results for $n_2 = 60$ in Table I. It can be seen the power and coverage is substantially better for the debiased fused multi-task lasso, while at the same time the confidence interval lengths are shorter.

In Figure 1 we show the power of the test for different values of n_2. We can see that the fusedlasso outperforms the other methods substantially. Projected ridge regression is particularly weak as it uses a worst case p-value obtained using an estimate of an upper bound on the bias [12].

B. Autism Dataset

Correlations in brain activity measured via fMRI reveal functional interactions between remote brain regions [24, 30]. In population analysis, they are used to measure how connectivity varies between different groups. Such analysis of brain function is particularly important in psychiatric diseases, that have no known anatomical support: the brain functions in a pathological aspect, but nothing abnormal is clearly visible in the brain tissues. Autism spectrum disorder is a typical example of such ill-understood psychiatric disease. Resting-state fMRI is accumulated in an effort to shed light on this diseases mechanisms: comparing the connectivity of autism patients versus control subjects. The ABIDE (Autism Brain Imaging Data Exchange) dataset [13] is a large scale effort sharing rest fMRI for this purpose. It gathers brain scans from 1 112 subjects across 16 sites, with 539 individuals suffering from autism spectrum disorder and 573 typical controls. Readily preprocessed and curated data is available[1]: the data have been slice-time interpolated, motion corrected, and realigned across subjects using C-PAC[36].

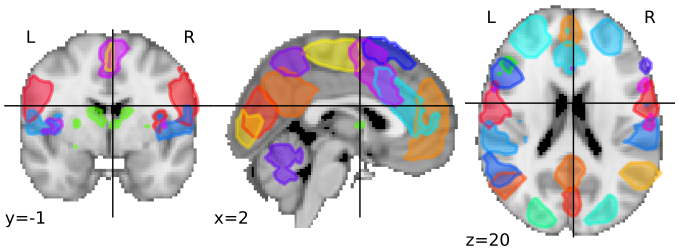

Fig. 2. Outlines of the regions of the MSDL atlas: the MSDL altas is comprised of 39 maps segmenting different resting-state networks.

In a connectome analysis approach [41, 33], each subject is described by a Gaussian graphical model measuring functional connectivity between a set of regions. A first step to build a connectome is thus to isolate regions of interest from the brain. We use a multi-subject atlas of functional regions derived from resting-state fMRI using multi-subject dictionary-learning (MSDL) [42]. The corresponding MSDL atlas (displayed in Fig. 2) is available from the internet[2].

We are interested in determining edge differences between the autism group and the control group. We use this data to show how our parametric hypothesis test can be used to determine differences in brain networks. Since no ground truth exists for this problem, we will focus the evaluation of our testing on comparison with a permutation testing framework [32, 10]. Here we permute the two conditions (e.g. autism and control group) and evaluate what the true p-value is and recompute our test statistics for each situation. This provides us with a finite sample strict control on the error rate. In essence this allows us to do a brute force validation of our parametric test.

For our experiments we take 2000 randomly chosen volumes from the control group subjects and 100 volumes from the autism group subjects. We perform permutation testing using the de-biased lasso, de-biased multi-task fused lasso, and projected ridge regression. Parameters for the de-biased fused lasso are chosen as in the previous section. For the de-biased lasso we use the exact settings for λ and constraints on M provided in the experimental section of [22]. Projected ridge regression is evaluated as in the previous section.

In Figure 3 we can see a comparison of three parametric approaches versus their analogue obtained with a permutation test. The chart plots the permutation p-values of each entry in the 38×39 $\boldsymbol{B}$ matrix against the expected parametric p-value. We see that for all the methods the points are above the line indicating the tests are not breaching the expected false positive rates, however the de-biased lasso and ridge projecting are very conservative and produce few values which actually detect. The de-biased multi-task fused lasso yields far more detections on the same dataset, almost all within the expected false positive rate or near it.

[1]http://preprocessed-connectomes-project.github.io/abide/

[2]https://team.inria.fr/parietal/research/spatial_patterns/spatial-patterns-in-resting-state/

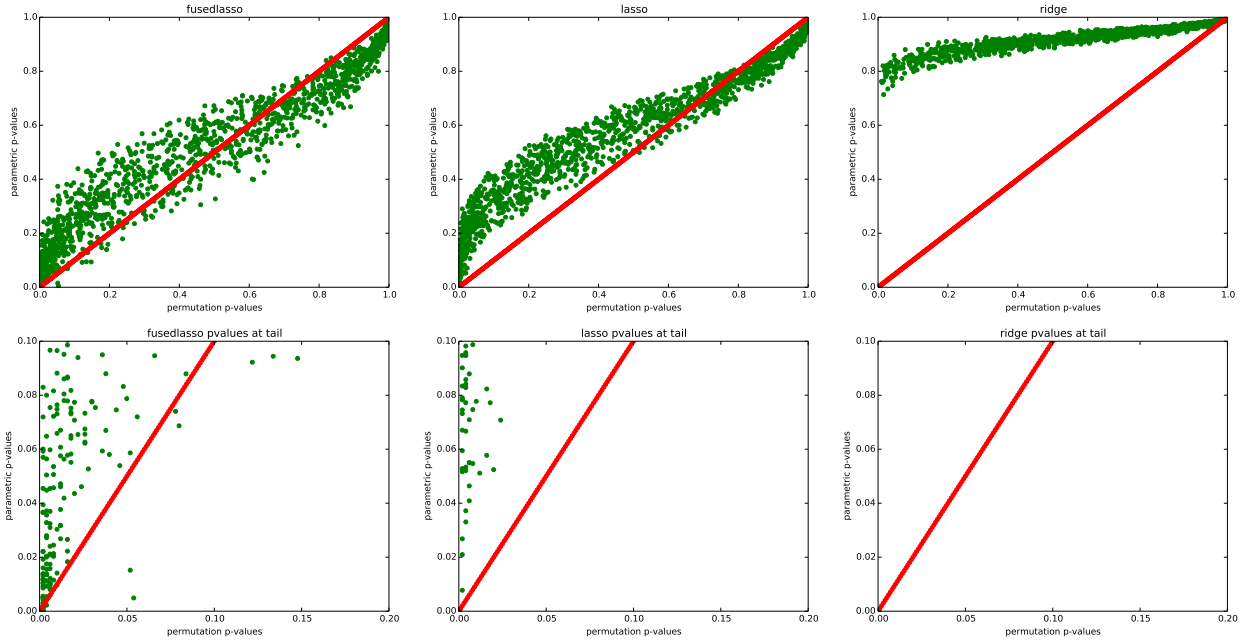

Fig. 3. Permutation testing comparing debiased fused lasso, debiased lasso, and projected ridge regression on the ABIDE dataset. We perform a resampling procedure on the test statistic of each to obtain a p-value, we expect this to p-value to align with the parametric p-value obtained without re-sampling. The chart plots the permutation p-values of each entry in the 38×39 $\boldsymbol{B}$ matrix against the expected parametric p-value. We see that for all 3 the points are above the line indicating the tests are not breaching the expected false positive rates, however debiased lasso and ridge projecting are very conservative and produce few values which actually detect. The fused lasso yields far more detections on the same dataset, almost all within the expected false positive rate or near it.

We now analyse the reproducibility of the results by repeatedly sampling 100 subsets of the data (with the same proportions $n_1 = 2000$ and $n_2 = 100$), obtaining the matrix of test statistics, adjusting the p-values to control the family wise error rate and selecting edges that fall below the 5% significance level. We correct for multiple comparisons using the Benjamini-Hochberg FDR procedure [2] which controls the family wise error rate (FWER). In Figure 4 we see the precentage of edges selected at least once are selected again. For the debiased-lasso procedure we obtain no edges which fall below the 5% significance level at the given FWER, thus for illustrative purposes we show the same chart without correcting the p-values in Figure 5.

In Figure 6 we show the edges which were frequently selected by the de-biased fused multi-task lasso selection in the 100 trials.

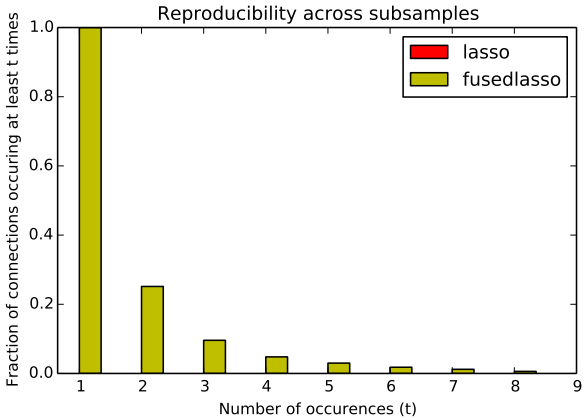

Fig. 4. Reproducibility of results from subsampling, debiased lasso does not produce any significant edge differences that correspond to a 5% family wise error rate

VII. Conclusions

We have shown how to characterize the distribution of differences of sparse estimators and how to use this distribution to obtain confidence intervals and p-values on GGM network differences. To address this specific problem, we have introduced the de-biased multi-task fused lasso. We have demonstrated on synthetic and real data that this approach can provide accurate p-values and a sizable increase on statistical power compared to standard procedures. The settings match very well those of population analysis for functional brain connectivity, and the increase of statistical power is direly needed to tackle the low sample sizes [7].

Future work calls for expanding the analysis to cases with more than two groups as well as considering a $\ell_{1,2}$ penalty sometimes used at the group level [43]. Additionally the squared loss objective optimizes excessively the prediction and could potentially be modified to lower further the sample complexity in terms of parameter estimation.

Acknowledgements

This work is partially funded by Internal Funds KU Leuven, ERC Grant 259112, FP7-MC-CIG 334380, and DIGITEO 2013-0788D - SOPRANO. GV also acknowledges funding from ANR-11-BINF-0004 NiConnect.

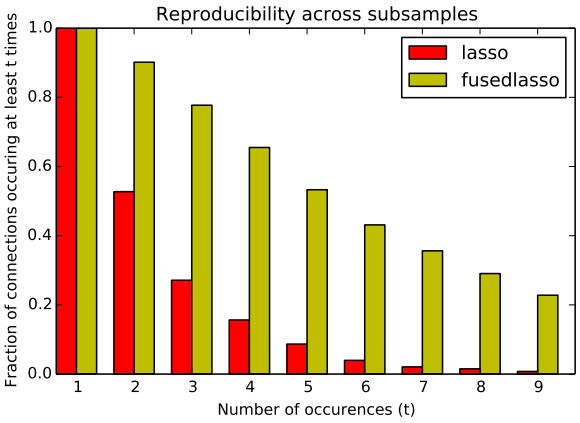

Fig. 5. Reproducibility of results from sub-sampling using uncorrected error rate shown for illustrative purposes. We see that the fused lasso is much more likely to detect edges and produce stable results.

REFERENCES

[1] Onureena Banerjee, Laurent El Ghaoui, Alexandre d'Aspremont, and Georges Natsoulis. Convex optimization techniques for fitting sparse Gaussian graphical models. In *International Conference on Machine learning*, pages 89–96. ACM, 2006.

[2] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, pages 289–300, 1995.

[3] Bharat Biswal, F Zerrin Yetkin, Victor M Haughton, and James S Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic resonance in medicine*, 34:537, 1995.

[4] Peter Bühlmann, Markus Kalisch, and Lukas Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1:255–278, 2014.

[5] Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 1st edition, 2011.

[6] Edward T Bullmore and Danielle S Bassett. Brain graphs: graphical models of the human brain connectome. *Annual review of clinical psychology*, 7:113, 2011.

[7] Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14:365, 2013.

[8] F Xavier Castellanos, Adriana Di Martino, R Cameron Craddock, Ashesh D Mehta, and Michael P Milham. Clinical applications of the functional connectome. *Neuroimage*, 80:527, 2013.

[9] Xi Chen, Seyoung Kim, Qihang Lin, Jaime G Carbonell, and Eric P Xing. Graph-structured multi-task regression and an efficient optimization method for general fused lasso. *arXiv preprint arXiv:1005.3579*, 2010.

[10] Benoit Da Mota, Virgile Fritsch, Gaël Varoquaux, Tobias Banaschewski, Gareth J Barker, Arun LW Bokde, Uli Bromberg, Patricia Conrod, et al. Randomized parcellation based inference. *NeuroImage*, 89:203–215, 2014.

[11] Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B*, 76(2):373–397, 2014.

[12] Ruben Dezeure, Peter Bühlmann, Lukas Meier, and Nicolai Meinshausen. High-dimensional inference: Confidence intervals, p-values and R-software hdi. *arXiv preprint arXiv:1408.4026*, 2014.

[13] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, et al. The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19:659, 2014.

[14] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[15] Apratim Ganguly and Wolfgang Polonik. Local neighborhood fusion in locally constant Gaussian graphical models. *arXiv preprint arXiv:1410.8766*, 2014.

[16] Michael Greicius. Resting-state functional connectivity in neuropsychiatric disorders. *Current opinion in neurology*, 21(4):424–430, 2008.

[17] Max Grazier G'Sell, Jonathan Taylor, and Robert Tibshirani. Adaptive testing for the graphical lasso. *arXiv preprint arXiv:1307.4765*, 2013.

[18] CJ Honey, O Sporns, Leila Cammoun, Xavier Gigandet, Jean-Philippe Thiran, Reto Meuli, and Patric Hagmann. Predicting human resting-state functional connectivity from structural connectivity. *Proceedings of the National Academy of Sciences*, 106(6):2035–2040, 2009.

[19] Jean Honorio and Dimitris Samaras. Multi-task learning of Gaussian graphical models. In *International Conference on Machine Learning*, pages 447–454, 2010.

[20] Jean Honorio, Dimitris Samaras, Nikos Paragios, Rita Goldstein, and Luis E Ortiz. Sparse and locally constant Gaussian graphical models. In *Advances in Neural Information Processing Systems*, pages 745–753, 2009.

[21] Jana Jankova and Sara van de Geer. Confidence intervals for high-dimensional inverse covariance estimation. *arXiv preprint arXiv:1403.6752*, 2014.

[22] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[23] Clare Kelly, Bharat B Biswal, R Cameron Craddock, F Xavier Castellanos, and Michael P Milham. Characterizing variation in the functional connectome: promise and pitfalls. *Trends in cognitive sciences*, 16:181, 2012.

[24] Martin A Lindquist et al. The statistical analysis of fMRI data. *Statistical Science*, 23(4):439–464, 2008.

[25] Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Annals of Statistics*, 42(2):413, 2014.

[26] Nikola T Markov, Mária Ercsey-Ravasz, David C Van Es-

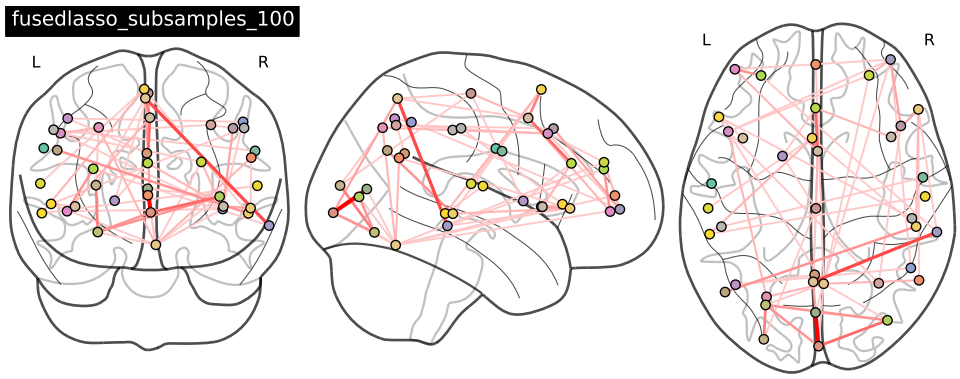

Fig. 6. Connectome of repeatedly picked up edges in 100 trials. Only shows edges which were selected more than once. Darker red indicates the edge was selected more frequently.

sen, Kenneth Knoblauch, Zoltán Toroczkai, and Henry Kennedy. Cortical high-density counterstream architectures. *Science*, 342(6158):1238406, 2013.

[27] George Marsaglia. Conditional means and covariances of normal variables with singular covariance matrix. *Journal of the American Statistical Association*, 59(308):1203–1204, 1964.

[28] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

[29] Karthik Mohan, Mike Chung, Seungyeop Han, Daniela Witten, Su-In Lee, and Maryam Fazel. Structured learning of Gaussian graphical models. In *Advances in Neural Information Processing Systems*, pages 620–628, 2012.

[30] Ricardo Pio Monti, Peter Hellyer, David Sharp, Robert Leech, Christoforos Anagnostopoulos, and Giovanni Montana. Estimating time-varying brain connectivity networks from functional MRI time series. *NeuroImage*, 103:427–443, 2014.

[31] Manjari Narayan and Genevera I Allen. Mixed effects models to find differences in multi-subject functional connectivity. *bioRxiv*, page 027516, 2015.

[32] Thomas E Nichols and Andrew P Holmes. Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, 15(1):1–25, 2002.

[33] Jonas Richiardi, Hamdi Eryilmaz, Sophie Schwartz, Patrik Vuilleumier, and Dimitri Van De Ville. Decoding brain states from fMRI connectivity graphs. *Neuroimage*, 56:616–626, 2011.

[34] Srikanth Ryali, Tianwen Chen, Kaustubh Supekar, and Vinod Menon. Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty. *NeuroImage*, 59(4):3852–3861, 2012.

[35] WR Shirer, S Ryali, E Rykhlevskaia, V Menon, and MD Greicius. Decoding subject-driven cognitive states with whole-brain connectivity patterns. *Cerebral cortex*, 22(1):158–165, 2012.

[36] S Sikka, B Cheung, R Khanuja, S Ghosh, C Yan, Q Li, J Vogelstein, R Burns, S Colcombe, et al. Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (C-PAC). In *5th INCF Congress of Neuroinformatics*, 2014.

[37] Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for FMRI. *Neuroimage*, 54:875, 2011.

[38] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, pages 267–288, 1996.

[39] Sara Van de Geer, Peter Bühlmann, Yaacov Ritov, Ruben Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[40] Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[41] Gaël Varoquaux and R Cameron Craddock. Learning and comparing functional connectomes across subjects. *NeuroImage*, 80:405–415, 2013.

[42] Gaël Varoquaux, Alexandre Gramfort, Fabian Pedregosa, Vincent Michel, and Bertrand Thirion. Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In *Information Processing in Medical Imaging*, page 562, 2011.

[43] Gaël Varoquaux, Alexandre Gramfort, Jean-Baptiste Poline, and Bertrand Thirion. Brain covariance selection: better individual functional connectivity models using population prior. In *Advances in Neural Information Processing Systems*, pages 2334–2342, 2010.

[44] Lourens Waldorp. Testing for graph differences using the desparsified lasso in high-dimensional data.