HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2019 December 06.

Published in final edited form as:

Biometrics. 2019 December ; 75(4): 1299–1309. doi:10.1111/biom.13075.

Incorporating Prior Information with Fused Sparse Group Lasso: Application to Prediction of Clinical Measures from Neuroimages

Joanne C. Beer[1,*], **Howard J. Aizenstein**[2], **Stewart J. Anderson**[3], **Robert T. Krafty**[3]

[1]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, U.S.A.

[2]Department of Psychiatry, University of Pittsburgh, Pittsburgh, Pennsylvania, U.S.A.

[3]Department of Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania, U.S.A.

Summary:

Predicting clinical variables from whole-brain neuroimages is a high-dimensional problem that can potentially benefit from feature selection or extraction. Penalized regression is a popular embedded feature selection method for high-dimensional data. For neuroimaging applications, spatial regularization using the l_1 or l_2 norm of the image gradient has shown good performance, yielding smooth solutions in spatially contiguous brain regions. Enormous resources have been devoted to establishing structural and functional brain connectivity networks that can be used to define spatially distributed yet related groups of voxels. We propose using the fused sparse group lasso penalty to encourage structured, sparse, and interpretable solutions by incorporating prior information about spatial and group structure among voxels. We present optimization steps for fused sparse group lasso penalized regression using the alternating direction method of multipliers (ADMM) algorithm. With simulation studies and in application to real fMRI data from the Autism Brain Imaging Data Exchange, we demonstrate conditions under which fusion and group penalty terms together outperform either of them alone.

Keywords

Autism; Neuroimaging; Penalized regression; Predictive model; Regularization; Structured sparsity

1. Introduction

Since the earliest functional magnetic resonance imaging (fMRI) studies of the human brain were carried out in the early 1990s, there have been relatively few translations of basic neuroscience findings to clinical applications in psychiatry, such as the use of biomarkers for determining diagnosis, prognosis, or predicting treatment response (Kapur et al., 2012; Woo et al., 2017). The traditional mass univariate approach in neuroimaging, which fits a model to each voxel independently, has been successful at characterizing group-level brain structure and function. However, a predictive model approach, where neuroimage features

* joanne.beer@pennmedicine.upenn.edu.

serve as predictors and a clinical variable is modeled as the outcome, may be better suited to clinical application. Predictive models are able to exploit dependencies between brain regions and thus can potentially explain more variability in the outcome than a mass univariate approach. Moreover, predictive models can yield individual-level out-of-sample predictions with potential clinical utility.

In this paper, we show how the fused sparse group lasso, a structured, sparse estimator, can incorporate prior information into a predictive model, thereby allowing researchers to harness results from the extensive recent research on brain structural and functional connectivity. Our goals include not only predictive accuracy, but also interpretable parameter estimates, as based on the following criteria: First, model structure entails that parameter values have a straightforward meaning; e.g., linear models tend to be more interpretable than nonlinear models. Second, models are appropriately sparse, including only relevant predictors, while not excluding any relevant predictors. Third, parameter estimates are understandable in light of existing background knowledge. In a translational neuroimaging context, this would mean that the brain regions implicated by the model estimates are neuroscientifically plausible according to existing knowledge or provide new insight into the neurobiological mechanism influencing the clinical outcome, and can potentially be used to establish biomarkers.

In Section 4, we apply fused sparse group lasso to a resting state fMRI dataset from the Autism Brain Imaging Data Exchange (ABIDE), a public repository of MRI datasets (Di Martino et al., 2014). Autism spectrum disorder (ASD) is a group of developmental disorders characterized by impaired social functioning and restrictive and repetitive behavior, and affects approximately 1% of children (Di Martino et al., 2014). Neuroimaging studies report abnormal functional connectivity between brain regions in ASD, although findings are mixed regarding the specific nature of the abnormalities (Di Martino et al., 2014). In our application, we show that incorporating prior information about voxel spatial location and functional connectivity using fused sparse group lasso increases accuracy when predicting a continuous measure of autistic social impairment from resting state fMRI data.

Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^n$ is a continuous outcome (e.g., score on a clinical depression rating scale), $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a predictor matrix (e.g., neuroimage voxel values), $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown vector of coefficients, $\epsilon \in \mathbb{R}^n$ is the error, $E(\boldsymbol{\epsilon}) = \mathbf{0}$, and $E(\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}_n$. This represents a high-dimensional setting where the number of subjects n is much less than the number of predictors p, which can be on the order of 100,000 voxels. To obtain a unique solution for $\boldsymbol{\beta}$, we can constrain the optimization problem using the penalized least squares estimator

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda J(\boldsymbol{\beta}), \tag{2}$$

where $\lambda \geqslant 0$ is a tuning parameter controlling the level of regularization. The penalty term $J(\boldsymbol{\beta})$ can impose both sparsity and structure, thereby constraining the solution space according to *a priori* information about relationships between elements of $\boldsymbol{\beta}$. Examples of unstructured and structured penalties are presented in Table 1.

Simulation studies and applications to real neuroimaging datasets (mostly using fMRI) have shown that penalties enforcing spatial smoothness frequently outperform unstructured penalties (Michel et al., 2011; Fiot et al., 2014). Not only do spatially-informed penalties yield more interpretable estimates insofar as they select contiguous groups of voxels in neuroscientifically plausible brain regions, but they often show better prediction performance. In addition to spatial regularization, group-structured regularization has shown promise in predictive neuroimaging models. For example, Shimizu et al. (2015) found that group lasso and sparse group lasso were superior to lasso and random forest and comparable to SVM in terms of classification accuracy, but unlike SVM, produced sparse and more interpretable models.

For neuroimaging applications, we aim to incorporate two types of structure into the penalty term $J(\boldsymbol{\beta})$ of the estimator in Equation (2): (1) local spatial information, to encourage smooth coefficient estimates across neighboring voxels; and (2) spatially distributed groups, such as those defined by functional or structural networks or anatomical regions, to allow voxels within the same group to be selected or shrunk to zero together. We achieve this by combining l_1, fusion, and group lasso penalties into a fused sparse group lasso penalty. We found one instance of this penalty in the literature, in a multi-task learning context where groups consist of repeated measures of the same task and smoothing is applied across time points within a group (Zhou et al., 2012). To our knowledge, the fused sparse group lasso penalty has never been studied via simulations or used in a predictive model with voxel-level neuroimaging data.

In the remainder of this paper, we present the fused sparse group lasso estimator in Section 2 and derive update steps to fit the fused sparse group lasso penalized least squares regression model using the alternating direction method of multipliers (ADMM) algorithm in Section 2.3. We report methods and results of a simulation study in Section 3 and apply our method to resting state fMRI data from the ABIDE repository in Section 4. We make concluding remarks in Section 5. We provide R and MATLAB functions for fitting the fused sparse group lasso estimator and additional supporting information online.

2. Fused Sparse Group Lasso

2.1 Model

Suppose we observe $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ from n independent subjects, indexed by $i = 1, \ldots, n$, where $y_i \in \mathbb{R}$ and $\mathbf{x}_i \in \mathbb{R}^p$. In the neuroimaging context considered here, y_i is a continuous scalar outcome for each subject such as age, depression scale score, or cognitive test score, and $\mathbf{x}_i$ is a vector of voxel values from a three dimensional brain image such that each element of $\mathbf{x}_i$ corresponds to one of p voxels. Assume that $\mathbf{y} = (y_1, \ldots, y_n)^T$ and the columns of the matrix $\mathbf{X} = (\mathbf{x}_1 | \ldots | \mathbf{x}_n)$ are centered, so we do not have an intercept term.

Furthermore, we standardize the columns of $\mathbf{X}$ to have unit standard deviation. We model the continuous outcome using standard linear regression as expressed in Equation (1).

2.2 Estimator

Since the number of voxels is typically orders of magnitude larger than the number of subjects, i.e., $p \gg n$, regularization is required to obtain a unique solution for $\boldsymbol{\beta}$. We propose estimating $\boldsymbol{\beta}$ by minimizing the sum of the loss function and three penalty terms:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} L(\boldsymbol{\beta}) + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\mathbf{D}\boldsymbol{\beta}\|_1 + \lambda_3 \Omega^{\mathcal{G}}(\boldsymbol{\beta}); \tag{3}$$

where $L(\boldsymbol{\beta})$ is the loss function (e.g., least squares); $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$ the l_1 norm of $\boldsymbol{\beta}$; $\mathbf{D}_{m \times p}$ is the three dimensional fusion matrix for fused lasso (see Web Appendix A for example), and $\|\mathbf{D}\boldsymbol{\beta}\|_1$ is the fusion penalty; $\Omega^{\mathcal{G}}(\boldsymbol{\beta}) = \sum_{g \in \mathcal{G}} \sqrt{p_g} \|\boldsymbol{\beta}_g\|_2$ is the $l_{2,1}$ group lasso penalty, which applies the l_2 norm, $\|\boldsymbol{\beta}_g\|_2 = \sqrt{\boldsymbol{\beta}_g^T \boldsymbol{\beta}_g}$, to the coefficients $\boldsymbol{\beta}_g$ for each group $g \in \mathcal{G}$, each of size p_g; and $\lambda_1, \lambda_2, \lambda_3 \geq 0$ are regularization tuning parameters.

The three penalty terms incorporate prior information into the estimator, encouraging the solution to have both sparsity and a particular structure. The standard lasso l_1 penalty encourages overall sparsity. The fusion penalty penalizes the absolute differences between coefficients at neighboring voxels, thereby encouraging local smoothness. The group lasso penalty encourages a group-level structure; entire groups may be selected or shrunk to zero together. For example, if groups are defined by functional networks, the penalty allows voxels involved in a common network to be shrunk to zero if that network is not important for prediction. Given the overlapping structure of brain networks, overlapping groups are another possibility worth considering. With appropriate weighting and a latent variable approach (Obozinski et al., 2011), the estimator could also accommodate overlapping groups.

For ease of selecting values for the tuning parameters via cross-validation, it is convenient to reparameterize (3) as follows:

$$\begin{aligned}\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} L(\boldsymbol{\beta}) + \alpha\gamma\lambda \|\boldsymbol{\beta}\|_1 + (1-\gamma)\lambda \|\mathbf{D}\boldsymbol{\beta}\|_1 \\ + (1-\alpha)\gamma\lambda \Omega^{\mathcal{G}}(\boldsymbol{\beta}),\end{aligned} \tag{4}$$

such that $\lambda > 0$ controls the overall level of regularization, $\alpha \in [0,1]$ controls the balance between the two sparsity inducing penalties (lasso and group lasso), and $\gamma \in [0,1]$ controls the balance between the two sparsity inducing penalties and the fusion penalty. When $\alpha = 1$ and $\gamma = 1$, the estimator reduces to the standard lasso; when $\alpha = 0$ and $\gamma = 1$, the estimator reduces to the group lasso, and so on for other subsets of the three penalty terms.

2.3 Optimization Algorithm

While a coordinate descent algorithm is often used to fit lasso penalized models, and Yuan and Lin (2006) proposed a blockwise coordinate descent algorithm for group lasso,

coordinate descent does not work for the fused lasso penalty. As discussed in Friedman et al. (2007), there are two main reasons for this: (1) the fused lasso penalty is non-separable into a sum of functions of the elements of $\boldsymbol{\beta}$, and (2) the fused lasso penalty is not continuously differentiable, so coordinate descent can get stuck. The accelerated gradient method algorithm employed in Zhou et al. (2012) depends on the separability of the penalty term across groups of $\boldsymbol{\beta}$, which is possible because they applied the fusion penalty only within groups. We wanted to allow for fusion across groups as well. Thus, for the optimization problem expressed in Equation 3, we chose to implement an ADMM algorithm (Boyd et al., 2011).

For simplicity, we assume that the groups are non-overlapping and form a partition of $\boldsymbol{\beta}$, so that each coefficient belongs to exactly one group. For applying ADMM, we follow a strategy similar to that employed in Huo and Tseng (2017) and exploit the fact that $|\beta_j| = \sqrt{\beta_j^2}$ and $|\beta_j - \beta_{j-1}| = \sqrt{(\beta_j - \beta_{j-1})^2}$. Then we can reformulate the lasso and fusion l_1 penalty terms as sets of $l_{2,1}$ group penalties whose groups have only one member. If there are p coefficients, $\mathbf{D}$ has m rows, and there are G groups that form a partition of $\boldsymbol{\beta}$, then the total number of effective groups is $p + m + G = N$.

Using a least-squares loss function, we now write the objective function (3) as

$$\arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{N} \lambda_j w_j \|\mathbf{K}_j \boldsymbol{\beta}\|_2, \tag{5}$$

with

$$\{\lambda_j, \mathbf{K}_j\} = \begin{cases} \{\lambda_1, \mathbf{j}_j\} & \text{if } j \in \{1, \ldots, p\} \\ \{\lambda_2, \mathbf{d}_j\} & \text{if } j \in \{p+1, \ldots, p+m\} \\ \{\lambda_3, \mathbf{G}_j\} & \text{if } j \in \{p+m+1, \ldots, p+m+G\}, \end{cases}$$

where $\lambda_j \in \{\lambda_1, \lambda_2, \lambda_3\}$ are the regularization parameters for the lasso, fusion, and group lasso penalties, respectively; w_j are group weights (for group lasso typically $w_j = \sqrt{p_j}$ where p_j is the number of elements in group j); $\mathbf{j}_j \in \mathbb{R}^p$ corresponds to the jth row of the $p \times p$ identity matrix; $\mathbf{d}_j \in \mathbb{R}^p$ corresponds to the $(j-p)$th row of the fusion matrix $\mathbf{D}$ in the three dimensional fusion penalty; and $\mathbf{G}_j \in \mathbb{R}^{p_j \times p}$ is a sparse matrix where each row has a 1 at a column position corresponding to a member of group j.

For ADMM, we introduce the auxiliary variables $\boldsymbol{\theta}_j = \mathbf{K}_j\boldsymbol{\beta}$. The optimization problem becomes minimize $\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{N} \lambda_j w_j \|\boldsymbol{\theta}_j\|_2$, subject to $\boldsymbol{\theta}_j - \mathbf{K}_j\boldsymbol{\beta} = \mathbf{0}$ for $j \in \{1,2, \ldots, N\}$. Let $\mathbf{K} = (\mathbf{K}_1 | \ldots | \mathbf{K}_N)$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)^T$, and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N)^T$. The augmented Lagrangian is

$$\mathscr{L}_\rho(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\mu}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{N} \lambda_j w_j \|\boldsymbol{\theta}_j\|_2 + \sum_{j=1}^{N} [\boldsymbol{\mu}_j^T(\boldsymbol{\theta}_j - \mathbf{K}_j\boldsymbol{\beta}) + \frac{\rho}{2}\|\boldsymbol{\theta}_j - \mathbf{K}_j\boldsymbol{\beta}\|_2^2],$$

where $\rho > 0$ is the step-size parameter and $\boldsymbol{\mu}_j$ are the dual variables for ADMM. After

initialization of $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and $\boldsymbol{\mu}$, the update steps for ADMM consist of the following: $\boldsymbol{\beta}^{t+1} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathscr{L}_\rho(\boldsymbol{\beta}, \boldsymbol{\theta}^t, \boldsymbol{\mu}^t)$; $\boldsymbol{\theta}_j^{t+1} = \arg\min_{\boldsymbol{\theta}_j \in \mathbb{R}^{p_j}} \mathscr{L}_\rho(\boldsymbol{\beta}^{t+1}, \boldsymbol{\theta}, \boldsymbol{\mu}^t)$; $\boldsymbol{\mu}_j^{t+1} = \boldsymbol{\mu}^t + \rho(\boldsymbol{\theta}_j^{t+1} - \mathbf{K}_j\boldsymbol{\beta}^{t+1})$. For $\boldsymbol{\beta}^{t+1}$ and $\boldsymbol{\theta}_j^{t+1}$ updates, the corresponding $\mathscr{L}_\rho$ subgradient will equal zero at the optimal solution. Thus, the update for $\boldsymbol{\beta}$ is $\boldsymbol{\beta}^{t+1} = (\mathbf{X}^T\mathbf{X} + \rho\mathbf{K}^T\mathbf{K})^{-1}[\mathbf{X}^TY + \mathbf{K}^T(\boldsymbol{\mu}^t + \rho\boldsymbol{\theta}^t)]$, and the update for $\boldsymbol{\theta}_j$ is $\boldsymbol{\theta}_j^{t+1} = (1 - \lambda_j w_j / [\rho\|\boldsymbol{\eta}_j\|_2])_+\boldsymbol{\eta}_j$, Where $\boldsymbol{\eta}_j = \mathbf{K}_j\boldsymbol{\beta} - \boldsymbol{\mu}_j/\rho$ and $(\cdot)_+ = \max(0, \cdot)$. We derive this and discuss how we implement stopping criteria and adaptive step-size in Web Appendix B.

2.4 Adaptive Fused Sparse Group Lasso

Zou (2006) showed that the lasso only exhibits consistent variable selection (i.e., identifies the right subset of non-zero coefficients asymptotically) under a certain nontrivial condition, which includes an orthogonal design matrix $\mathbf{X}$ and $p = 2$ as special cases. The adaptive lasso, on the other hand, achieves consistent variable selection by differentially scaling the tuning parameter, λ, for each coefficient by the factor $|\hat{\beta}_j^*|^{-\gamma}$, where $\hat{\beta}_j^*$ is a consistent estimator for β_j such as the ordinary least squares estimator, and $\gamma > 0$ (Zou, 2006). Adaptive versions of fused lasso (Viallon et al., 2013) and group lasso (Wang and Leng, 2008) have also been developed. In our application to a real neuroimaging dataset in Section 4, we implement an adaptive version of fused sparse group lasso using ridge regression to obtain initial coefficient estimates, $\hat{\boldsymbol{\beta}}^{\text{ridge}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda^{\text{ridge}}\|\boldsymbol{\beta}\|_2^2$. The weights, w_j, introduced in Equation (5) are defined as

$$w_j = \begin{cases} \|\mathbf{j}_j\hat{\boldsymbol{\beta}}^{\text{ridge}}\|_1^{-1} & \text{if } j \in \{1, \ldots, p\} \\ \|\mathbf{d}_j\hat{\boldsymbol{\beta}}^{\text{ridge}}\|_1^{-1} & \text{if } j \in \{p+1, \ldots, p+m\} \\ \|\mathbf{G}_j\hat{\boldsymbol{\beta}}^{\text{ridge}}\|_2^{-1} & \text{if } j \in \{p+m+1, \ldots, p+m+G\}. \end{cases} \quad (6)$$

3. Simulation Study

3.1 Simulation Study Methods

Our simulation study aimed to show that, for a given modeling scenario, the optimal weighting of the three penalty terms in the fused sparse group lasso depends on the underlying structure of the true coefficients. We also sought to characterize the optimal penalty weights for a range of different coefficient structures. Accordingly, we evenly divided the pixels of two dimensional 20 × 20 images into 16 groups of 25 and considered three spatial arrangements of the groups (Figure 1): (A) members of a group were completely aggregated into 5 × 5 squares; (B) groups were partially aggregated, consisting of one 3 × 3 square, three 2 × 2 squares, and two 1 × 2 rectangles; (C) groups were completely distributed such that no pixels from the same group were touching sides. For each of these group structures, one group was selected to have non-zero coefficients, which were all set equal to 3. We also considered sparse versions of the coefficients, where 40% of

coefficients in the active group were set to zero. Additionally, we considered three more scenarios under the partially aggregated group structure: an extra sparse scenario, with 80% of active group coefficients set to zero; a misspecified group structure, where the set of true coefficients was divided among several groups; and a sparse version of the misspecified group structure. Thus, there were nine total scenarios of true coefficients (Figure 1).

For each of $n = 50$ training subjects and $n = 50$ test subjects, we generated a vector of 400 independent standard normal random variables to serve as predictors, where each corresponded to a pixel in the 20 × 20 image. The responses, $\mathbf{y}$, were then computed by the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where each element of $\boldsymbol{\epsilon}$ was independent normal with mean zero and variance 4. To select optimal tuning parameter values, we parameterized according to Equation (4). The fusion penalty was applied between coefficients of pixels that shared an edge, and the group penalty was applied to each of the 16 groups, as previously described. For each pair of $\alpha \in \{0, 0.2, 0.5, 0.8, 1\}$ and $\gamma \in \{0, 0.2, 0.5, 0.8, 1\}$, we performed 5-fold cross-validation over 50 values of $\lambda = 10^x$, where values of x formed a grid on the interval $[-3, 3]$, and selected the λ that resulted in the lowest cross-validation mean squared error. We fit the model to the entire sample of $n = 50$ training subjects at the given (α, γ) pair using this λ and calculated mean squared error of the estimated coefficients $(\text{MSE}(\hat{\boldsymbol{\beta}}) = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 / 400)$ and mean squared prediction error for the test data $(\text{MSE}(\hat{\mathbf{y}}_{\text{test}}) = \|\hat{\mathbf{y}}_{\text{test}} - \mathbf{y}_{\text{test}}\|_2^2 / 50)$. We repeated the entire procedure 100 times for each of the nine scenarios. We also generated another test sample of $n = 100$ for the purpose of decomposing the mean squared error into squared bias and variance at each (α, γ) combination across the 100 trained models. Analyses were done using R version 3.4.0 (R Core Team, 2017).

We hypothesized that, on the basis of $\text{MSE}(\hat{\boldsymbol{\beta}})$, $\text{MSE}(\hat{\mathbf{y}}_{\text{test}})$, or both, (1) the fusion penalty term would perform worse and the sparsity penalty terms would perform better (i.e., optimal γ value would increase) as the groups became more spatially distributed; (2) the group penalty term would perform worse and the l_1 lasso penalty term would perform better (i.e., optimal α value would increase) as the sparsity of true coefficients increased or with misspecification of group structure. We also sought to determine whether the lowest cross-validation error would correspond to the optimal values of (α, γ).

3.2 Simulation Study Results

Figure 2 shows the distributions of cross-validation error, $\text{MSE}(\hat{\boldsymbol{\beta}})$, and $\text{MSE}(\hat{\mathbf{y}}_{\text{test}})$ across the (α, γ) combinations for true coefficient scenario 5B. Optimal (α, γ) combinations for each scenario are presented in Web Appendix C, Web Table C1; detailed simulation results are reported in Web Tables C2-C10; box plots for the other scenarios are shown in Web Figures C1-C3; and bias-variance decompositions of $\text{MSE}(\hat{\mathbf{y}}_{\text{test}})$ are shown in Web Figure C4.

As expected, on the basis of both $\text{MSE}(\hat{\boldsymbol{\beta}})$ and $\text{MSE}(\hat{\mathbf{y}}_{\text{test}})$, as groups became more spatially distributed, the optimal value of γ increased from 0.2 for the completely aggregated to 1 for the completely distributed group structure, shifting weight from the fusion penalty term to

the sparsity penalty terms. This pattern was similar for the complete (1A, 2B, 3C) and sparse (4A, 5B, 6C) group scenarios. As the sparsity of true coefficients increased in the partially aggregated group scenarios (2B, 5B, 7B), the optimal value of α increased from 0 for the complete group to 0.8 for the extra sparse group scenario, shifting weight from the group penalty term to the l_1 lasso penalty term. When group structure was misspecified, the optimal α value was 1 for both scenarios (8B, 9B), putting zero weight on the group penalty term in favor of the l_1 lasso penalty term.

The results demonstrate that the combination of penalty terms in the fused sparse group lasso adapt to a wide range of spatial arrangements and sparsity levels of true coefficients. In all seven scenarios where group structure was correctly specified, the (α, γ) combination yielding the most frequent lowest cross-validation error corresponded to the most frequent lowest MSE($\hat{\boldsymbol{\beta}}$) and MSE($\hat{\mathbf{y}}_{\text{test}}$), indicating that selecting tuning parameters based on lowest cross-validation error tends to correspond to the optimal model. For the misspecified group scenarios, cross-validation error was lowest for either the first or second most frequent lowest MSE($\hat{\boldsymbol{\beta}}$) and MSE($\hat{\mathbf{y}}_{\text{test}}$) (see Web Appendix C, Web Tables C8 and C9).

4. Application to Neuroimaging Data

We applied fused sparse group lasso (FSGL) penalized regression to a resting state fMRI dataset of ASD (n = 111) and typically developing (TD, n = 108) male participants (mean (SD) age 17.4 (7.5) years, see Web Appendix D, Web Table D1 for descriptive summary) from the ABIDE repository (Di Martino et al. (2014), see Supporting Information for URL). In this set of participants, Cerliani et al. (2015) used independent components analysis to identify 19 resting state brain networks. The authors found that autistic traits as measured by Social Responsiveness Scale (SRS) scores were positively associated with functional connectivity between a subcortical network, comprising basal ganglia and thalamus, and two cortical networks: (1) dorsal and (2) ventral primary somatosensory and motor cortices. The association was only significant in the ASD group. Given that the resting state networks evaluated in Cerliani et al. (2015) represent relatively large brain regions, we used FSGL regression to more precisely define the cortical regions whose functional connectivity with a subcortical seed region best predicts SRS scores.

4.1 Application Methods

Preprocessed fMRI data was downloaded from the ABIDE I Preprocessed repository (Craddock et al., 2013). Data were preprocessed using the Connectome Computational System pipeline with no global signal regression and band pass filtering (0.01 – 0.1 Hz). The independent component resting state network data from Cerliani et al. (2015) was

[6]Supporting Information
Web Appendices, Tables, and Figures referenced in Sections 3 and 4, and data and code for the simulation study in Section 3 and application to ABIDE dataset in Section 4 (including an R package and Matlab functions to fit fused sparse group lasso) are available with this paper at the Biometrics website on Wiley Online Library. Data and code are also available at https://github.com/jcbeer/fsgl. ABIDE repository website is at http://fcon_1000.projects.nitrc.org/indi/abide/, and preprocessing pipeline information is available at http://preprocessed-connectomes-project.org/abide/Pipelines.html. Resting state network data was retrieved from https://github.com/sblnin/rsfnc.

downloaded and resampled to 3×3×3 mm^3 voxels to match the ABIDE data. (See Supporting Information for URLs.)

We partitioned the brain into 19 resting state networks by assigning each voxel to the maximal spatial independent component at that voxel out of the 19 components identified in Cerliani et al. (2015). We restricted our analyses to the three networks mentioned above: the basal ganglia/thalamus subcortical network and the two sensorimotor cortical networks. We defined a subcortical seed region by selecting the peak voxels in the subcortical network independent component spatial map. This yielded bilateral regions of the thalamus, with 12 voxels centered at MNI coordinates (11, −11, 11) and 11 voxels centered at (−11, −11, 12) (Figure 3A). For each participant, the first eigenvariate of the seed region time series was extracted, and its Pearson correlation was calculated with each voxel time series in the cortical regions of interest to form a seed-based connectivity map. Fisher's r-to-z transformation was applied to each voxel. After excluding voxels where any participants had missing data, this left $p = 5476$ voxels to serve as predictors for FSGL regression (Figure 3B).

Participant data were divided into training ($n = 175$) and test ($n = 44$) sets. Test set data was put aside until all model fitting was completed. The following steps were carried out with the training set data:

1. For the group penalty term, voxels in the cortical regions of interest were partitioned into 50 groups using agglomerative hierarchical clustering on the voxel time series. First, for each training participant, Pearson correlations between time series were calculated for all possible pairs of voxels in the cortical regions of interest. Correlation matrices were averaged across participants, and a distance matrix was formed by applying the elementwise transformation $d = \sqrt{2 * (1 - r)}$. Finally, hierarchical clustering using Ward's method was performed based on the distance matrix, and the resulting tree was cut to form 50 groups which ranged in size from 43 to 268 voxels (Figure 3C).
2. A linear regression model adjusted raw SRS scores for age, full-scale IQ, site of acquisition, eye status at scan (open or closed), and mean framewise displacement. The residuals were used as the outcome for FSGL regression. (See Web Appendix D, Web Table D2)
3. After centering the SRS outcome and standardizing columns of the predictor matrix, 5-fold cross-validation was used to determine the λ value yielding the minimum cross-validation error at selected values of (α, γ) for the FSGL regression. We chose to compare (α, γ) equal to (1.0, 1.0) (standard lasso), (0.2, 1.0) (sparse group lasso), (0.2, 0.8) (fused sparse group lasso), and (0.0, 0.8) (fused group lasso). Cross-validation folds were stratified to ensure that they had similar distributions of the adjusted SRS outcome.
4. To estimate the coefficients, the model was fit to the entire training set at the optimal λ for selected values of (α, γ).

5. For adaptive FSGL regression, first ridge regression estimates were obtained (using R package glmnet, Friedman et al. (2010)) and adaptive weights were formed according to Equation 6, and then steps (3) and (4) were completed.

For the test set data, adjusted SRS scores were predicted for each participant by taking the dot product of the estimated FSGL regression parameters with the participant's predictor variables, which were first standardized according to the training set column means and standard deviations. Raw test set SRS scores were adjusted using the linear regression model parameters estimated with the training set data. Prediction accuracy was assessed via mean squared error and Pearson correlation of predicted with actual adjusted SRS scores. Since prior studies have used resting state fMRI data to classify ASD and TD subjects into diagnostic groups rather than predict SRS, in order to compare our results we used receiver operating characteristic (ROC) curve analysis to find the best classification threshold for predicted SRS scores from the best-performing models and calculated the corresponding classification accuracies.

Analyses were done using a combination of AFNI version 17.1.03 (Cox, 1996), MATLAB version 9.1.0 (R2016b) (MATLAB, 2016), and R version 3.4.0 (R Core Team, 2017).

4.2 Application Results

Results for non-adaptive and adaptive fused sparse group lasso as well as for ridge and elastic net penalties are summarized in Table 2. The best test set prediction was achieved by the adaptive fused sparse group lasso with (α, γ) equal to (0.2, 0.8), which gave a mean squared error of 1165.2 and Pearson correlation $r = 0.437$ ($p = 0.003$) (Figure 4C; see Web Appendix D, Web Figure D1 for other adaptive penalties). (For comparison, Cerliani et al. (2015) reported correlations of $r = 0.21$ and 0.25 for the respective cortical networks in ASD participants.) Cross-validation error was in general much lower for adaptive penalties than for non-adaptive penalties (Figure 4A). The superior performance of adaptive, ridge, and elastic net penalties over the non-adaptive penalties in this particular application is likely due at least in part to high multicollinearity of the predictors and the influence of many small, weak effects of predictors on the outcome rather than a few strong effects. Estimated coefficient brain maps for two sets of α and γ values are shown in Figure 4B (see Web Appendix D, Web Figure D1 for other adaptive penalties). Higher SRS scores correspond to greater autistic social impairment. Thus the coefficient maps reflect multivariate thalamic seed connectivity patterns predictive of greater social impairment. Penalties including the fusion term, i.e., with $\gamma = 0.8$, resulted in larger clusters of contiguous regions, rather than the more scattered coefficient maps resulting when $\gamma = 1.0$.

A couple of questions arise regarding the clinical significance of the findings. First, does the result provide insight into the neurobiology of ASD? While the sparse, structured penalty succeeded in narrowing down the predictors to a smaller subset of the most predictive voxels, it is not immediately clear why these particular scattered regions of sensorimotor cortex are most informative. We invite interested readers to further explore the coefficient brain maps available at the URL noted in the Supporting Information Section, below. Second, does the result represent a good diagnostic biomarker? We consider this question in the following context. ASD has been associated with abnormalities in connectivity between

multiple brain regions (Di Martino et al., 2014). Accordingly, studies using resting state fMRI data to define ASD biomarkers have often summarized voxel-level data using regions of interest and considered connectivity between multiple regions, rather than use a focused, voxel-level approach as we did. Previous studies using such methods on the ABIDE dataset have achieved diagnostic classification (ASD versus TD) accuracies in the range of 60% to 71% (Abraham et al., 2017; Nielsen et al., 2013; Plitt et al., 2015; Kassraian-Fard et al., 2016), well below the classification accuracy that can be achieved using behavioral measures such as the SRS, which can attain accuracies of up to 95% (Plitt et al., 2015). The difficulty of identifying fMRI biomarkers for ASD may be due to the noisiness of fMRI data or the neurobiological heterogeneity of the disorder (Plitt et al., 2015). What is remarkable about our result is that, when we dichotomize our predicted outcomes for the purpose of diagnostic classification, we can achieve similar accuracies (classification accuracy of 29 to 30 out of 44 test subjects, or 66% to 68%, for the adaptive penalties), even though we only considered connectivity between a single thalamic seed region and sensorimotor cortex. This may reflect the richness imparted by voxel-level as opposed to region of interest functional connectivity data. It seems possible that adding other features to the model, e.g., using not only a thalamic seed region, but also including voxel-level connectivity data from other seed regions that have shown abnormal connectivity in ASD, such as the default mode network and regions implicated in social cognition, would likely improve prediction performance further.

5. Conclusions

The fused sparse group lasso penalty offers a flexible way to incorporate pertinent structure into a predictive model, which can lead to more interpretable coefficient estimates and better predictive performance on test data. The fusion penalty term constrains coefficients that we expect to have similar estimated values, and we can use it to enforce local spatial smoothness in an image. The group penalty term groups together coefficients that we do not necessarily expect to have similar values, but we expect to be selected simultaneously, such as voxels residing in the same functional brain networks. The l_1 penalty term allows sparse groups, and may also be useful when groups are misspecified. Cross-validation over a range of weights for the three penalty terms allows a data-driven way of incorporating information about coefficient structure into a prediction model.

In this paper we have presented an ADMM optimization algorithm to fit fused sparse group lasso. A simulation study featuring a range of coefficient structures demonstrated instances where a combination of the three penalty terms together outperforms any smaller subset, and showed that cross-validation is a reliable way to select optimal tuning parameter weights. On real fMRI data, we found that incorporating adaptive weights derived from initial ridge regression coefficient estimates greatly improved performance over non-adaptive fused sparse group lasso as well as ridge and elastic net penalties. The adaptive fused sparse group lasso produced the best test set prediction, and the addition of fusion and group penalty terms resulted in less dispersed, more clustered coefficient maps. Fused sparse group lasso, a generalization of lasso, group lasso, and fused lasso, has potential application not only to prediction problems in neuroimaging, but also to other contexts where coefficients are expected to be both smooth and group-structured.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was partially supported by the National Institute of General Medical Sciences grant R01GM113243 and National Institute on Aging grants 2P01AG025204-11A and R01AG052446. This research was supported in part by the University of Pittsburgh Center for Research Computing through the resources provided. We specifically acknowledge the assistance of Kim Wong. The authors thank George Tseng, Helmet Karim, Dana Tudorascu, and Leonardo Cerliani for helpful comments. We also thank the Autism Brain Imaging Data Exchange and the authors of Cerliani et al. (2015) for making their data available.

References

ABIDE Preprocessed: Functional Preprocessing (2016). http://preprocessed–connectomes–project.org/abide/Pipelines.html Accessed 22 June 2018.

Abraham A, Milham MP, Di Martino A, Craddock RC, Samaras D, Thirion B, et al. (2017). Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example. NeuroImage 147, 736–745. [PubMed: 27865923]

Autism Brain Imaging Data Exchange (2012). http://fcon_1000.projects.nitrc.org/indi/abide/. Accessed 22 June 2018.

Boyd S, Parikh N, Chu E, Peleato B, and Eckstein J (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine Learning 3, 1–122.

Cerliani L, Mennes M, Thomas RM, Di Martino A, Thioux M, and Keysers C (2015). Increased functional connectivity between subcortical and cortical resting-state networks in autism spectrum disorder. JAMA Psychiatry 72, 767–777. [PubMed: 26061743]

Cerliani L, Mennes M, Thomas RM, Di Martino A, Thioux M, and Keysers C (2016). Analysis of resting state functional connectivity network developed for ABIDE. https://github.com/sblnin/rsfnc. Accessed 22 June 2015.

Cox RW (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. Computers and Biomedical Research 29, 162–173. [PubMed: 8812068]

Craddock R, Benhajali Y, Chu C, Chouinard F, Evans A, Jakab A, et al. (2013). The Neuro Bureau Preprocessing Initiative: Open sharing of preprocessed neuroimaging data and derivatives. Frontiers in Neuroinformatics.

Di Martino A, Yan C-G, Li Q, Denio E, Castellanos FX, Alaerts K, et al. (2014). The Autism Brain Imaging Data Exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular Psychiatry 19, 659–667. [PubMed: 23774715]

Fiot J-B, Raguet H, Risser L, Cohen LD, Fripp J, Vialard F-X, et al. (2014). Longitudinal deformation models, spatial regularizations and learning strategies to quantify Alzheimer's disease progression. NeuroImage: Clinical 4, 718–729. [PubMed: 24936423]

Friedman J, Hastie T, Hfling H, and Tibshirani R (2007). Pathwise coordinate optimization. The Annals of Applied Statistics 1, 302–332.

Friedman J, Hastie T, and Tibshirani R (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33, 1–22. [PubMed: 20808728]

Grosenick L, Klingenberg B, Katovich K, Knutson B, and Taylor JE (2013). Interpretable whole-brain prediction analysis with GraphNet. Neuroimage 72, 304–321. [PubMed: 23298747]

Hoerl AE and Kennard RW (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67.

Huo Z and Tseng G (2017). Integrative sparse k-means with overlapping group lasso in genomic applications for disease subtype discovery. The Annals of Applied Statistics 11, 1011–1039. [PubMed: 28959370]

Kapur S, Phillips AG, and Insel TR (2012). Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? Molecular Psychiatry 17, 1174–1179. [PubMed: 22869033]

Kassraian-Fard P, Matthis C, Balsters JH, Maathuis MH, and Wenderoth N (2016). Promises, pitfalls, and basic guidelines for applying machine learning classifiers to psychiatric imaging data, with autism as an example. Frontiers in Psychiatry 7, 177. [PubMed: 27990125]

MATLAB (2016). version 9.1.0 (R2016b). The MathWorks Inc., Natick, Massachusetts.

Michel V, Gramfort A, Varoquaux G, Eger E, and Thirion B (2011). Total variation regularization for fMRI-based prediction of behavior. IEEE Transactions on Medical Imaging 30, 1328–1340. [PubMed: 21317080]

Nielsen JA, Zielinski BA, Fletcher PT, Alexander AL, Lange N, Bigler ED, et al. (2013). Multisite functional connectivity MRI classification of autism: ABIDE results. Frontiers in Human Neuroscience 7, 599. [PubMed: 24093016]

Obozinski G, Jacob L, and Vert J-P (2011). Group lasso with overlaps: The latent group lasso approach. arXiv preprint, https://arxiv.org/abs/1110.0413.

Plitt M, Barnes KA, and Martin A (2015). Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards. NeuroImage: Clinical 7, 359–366. [PubMed: 25685703]

R Core Team (2017). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

Rudin LI, Osher S, and Fatemi E (1992). Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena 60, 259–268.

Shimizu Y, Yoshimoto J, Toki S, Takamura M, Yoshimura S, Okamoto Y, et al. (2015). Toward probabilistic diagnosis and understanding of depression based on functional MRI data analysis with logistic group lasso. PloS One 10, e0123524. [PubMed: 25932629]

Simon N, Friedman J, Hastie T, and Tibshirani R (2013). A sparse-group lasso. Journal of Computational and Graphical Statistics 22, 231–245.

Tibshirani R (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) pages 267–288.

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

Viallon V, Lambert-Lacroix S, Höfling H, and Picard F (2013). Adaptive generalized fused-lasso: Asymptotic properties and applications. HAL preprint, https://hal.archives-ouvertes.fr/hal-00813281/.

Wang H and Leng C (2008). A note on adaptive group lasso. Computational Statistics & Data Analysis 52, 5277–5286.

Woo C-W, Chang LJ, Lindquist MA, and Wager TD (2017). Building better biomarkers: Brain models in translational neuroimaging. Nature Neuroscience 20, 365–377. [PubMed: 28230847]

Yuan M and Lin Y (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68, 49–67.

Zhou J, Liu J, Narayan VA, and Ye J (2012). Modeling disease progression via fused sparse group lasso. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1095–1103. Association for Computing Machinery.

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

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

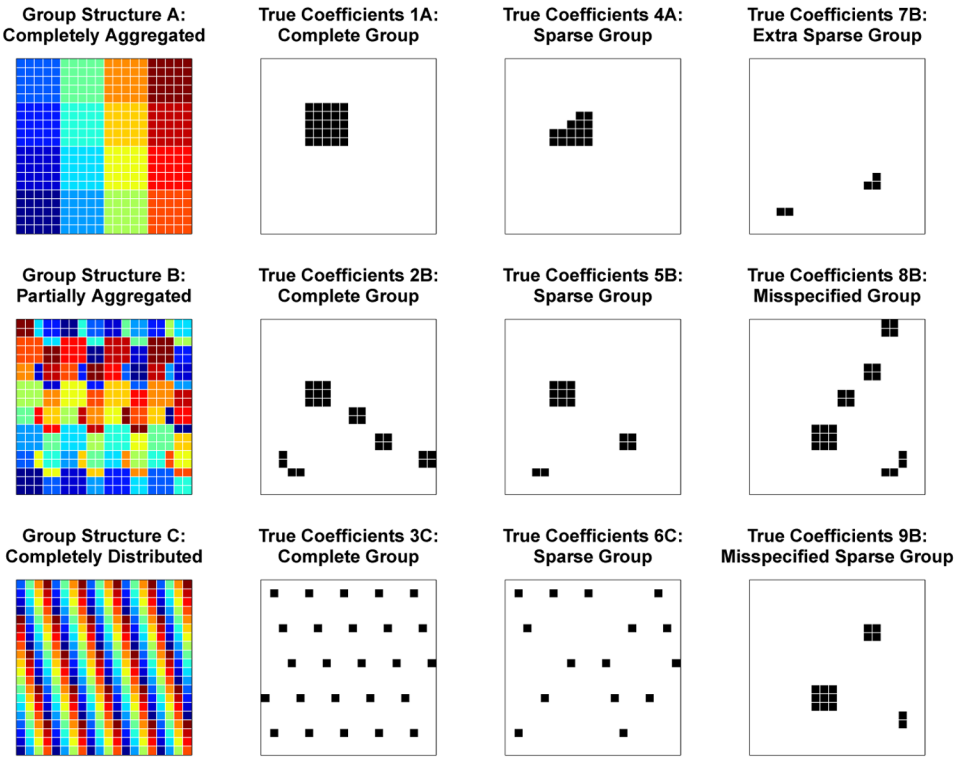

Figure 1.
Simulation study group structures (top row) and true coefficients. This figure appears in color in the electronic version of the manuscript.

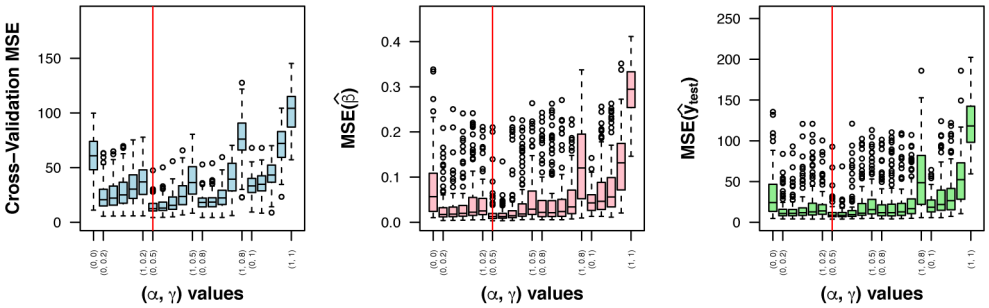

Figure 2.

Simulation study results for true coefficients 5B. Values of $\gamma \in \{0, 0.2, 0.5, 0.8, 1\}$ increase from left to right on the *x*-axis, corresponding to complete fusion penalty on the left ($\gamma = 0$) and complete sparsity penalties on the right ($\gamma = 1$). Intervals of increasing $a \in \{0, 0.2, 0.5, 0.8, 1\}$ values correspond to complete group penalty on the left ($a = 0$) and complete l_1 lasso penalty on the right ($a = 1$). Vertical lines indicate (a, γ) combination yielding most frequent lowest error over 100 simulations. This figure appears in color in the electronic version of the manuscript. MSE: mean squared error.

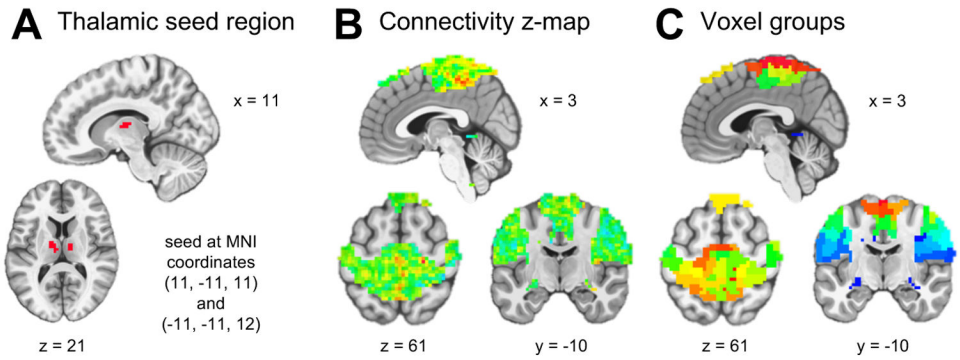

Figure 3.

(A) The thalamic seed region consisted of 23 3 × 3 × 3 mm^3 voxels. (B) Connectivity z-maps of 5476 voxels for each participant served as predictors for fused sparse group lasso regression. (C) Voxels were partitioned into 50 groups using agglomerative hierarchical clustering on the training set resting state fMRI voxel time series. This figure appears in color in the electronic version of the manuscript.

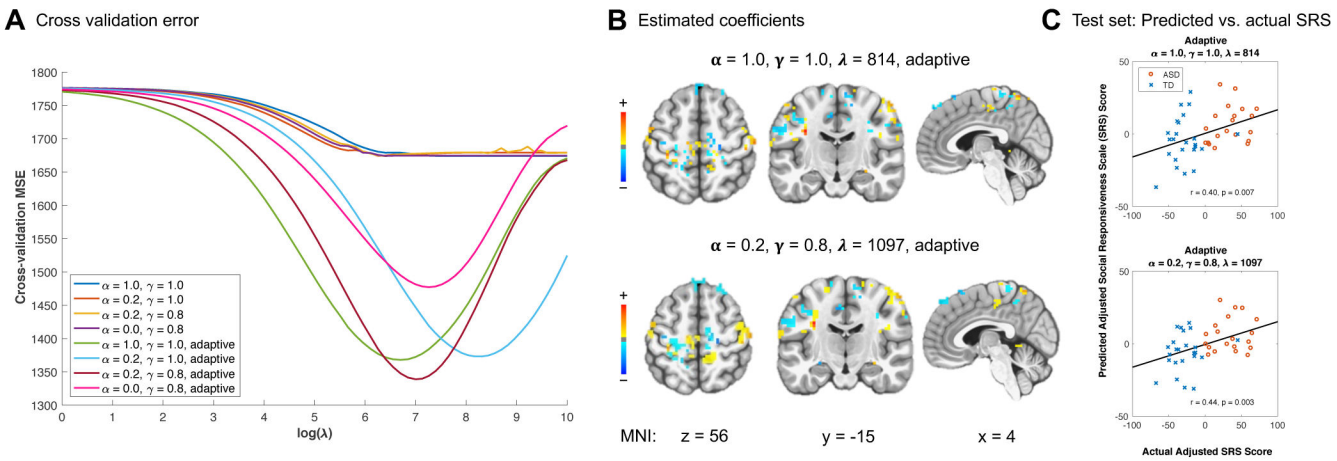

Figure 4.

(A) Five-fold cross-validation was carried out over a range of λ values for several sets of α and γ values. (B) Correlation of predicted and actual adjusted SRS scores for selected α and γ values. Points are distinguished by autism spectrum disorder (ASD, red circle) and typically developing (TD, blue cross) diagnosis groups. (C) Estimated coefficients at the optimal λ for selected α and γ values. Higher coefficient values contribute to higher predicted Social Responsiveness Scale (SRS) scores, which indicate greater autistic social impairment. MNI: Montreal Neurological Institute coordinates. This figure appears in color in the electronic version of the manuscript.

Table 1

Examples of unstructured and structured penalty terms

Unstructured penalties	$J(\boldsymbol{\beta})$
Lasso (Tibshirani, 1996)	$\|\boldsymbol{\beta}\|_1$
Ridge (Hoerl and Kennard, 1970)	$\|\boldsymbol{\beta}\|_2^2$
Elastic net (Zou and Hastie, 2005)	$\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2^2;\ \alpha \in [0,1]$
Structured penalties	$J(\boldsymbol{\beta})$
Isotropic total variation (Rudin et al., 1992)	$\|\mathbf{D}\boldsymbol{\beta}\|_{2,1}$; matrix $\mathbf{D}$ encodes spatial structure
Fused lasso * (Tibshirani et al., 2005)	$\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\mathbf{D}\boldsymbol{\beta}\|_1;\ \alpha \in [0,1]$
Graph net ** (Grosenick et al., 2013)	$\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\mathbf{D}\boldsymbol{\beta}\|_2^2;\ \alpha \in [0,1]$
Group lasso (Yuan and Lin, 2006)	$\sum_{g \in \mathcal{G}} \sqrt{p_g}\|\boldsymbol{\beta}_g\|_2$; groups $\mathcal{G}$ form a partition of $\boldsymbol{\beta}$
Sparse group lasso (Simon et al., 2013)	$\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\sum_{g \in \mathcal{G}} \sqrt{p_g}\|\boldsymbol{\beta}_g\|_2;\ \alpha \in [0,1]$

*Also known as anisotropic total variation–l_1

**Also known as sparse graph Laplacian

Table 2

Comparison of estimators applied to ABIDE dataset

					Training set (n = 175)				Test set (n = 44)		
Method	α	γ	**Estimator**	* **Optimal** λ	**CVMSE**	**MSE**	r	p	**MSE**	r	p
Glmnet	0.0		Ridge	2627	1646.7	954.3	0.879	< 0.001	1325.8	0.285	0.060
	0.01		Elastic Net	289	1661.2	935.7	0.883	< 0.001	1305.4	0.320	0.034
	1.0	1.0	Lasso	1848	1674.2	1689.1	0.159	0.036	1426.7	0.035	0.821
FSGL	0.2	1.0	Sparse Group Lasso	521	1674.4	1572.0	0.383	< 0.001	1427.8	0.069	0.654
	0.2	0.8	Fused Sparse Group Lasso	604	1673.9	1633.2	0.254	< 0.001	1434.3	0.038	0.805
	0.0	0.8	Fused Group Lasso	604	1674.3	1641.2	0.232	0.002	1435.4	0.032	0.838
Adaptive FSGL	1.0	1.0	Adaptive Lasso	814	1368.1	120.1	0.986	< 0.001	1193.1	0.406	0.006
	0.2	1.0	Adaptive Sparse Group Lasso	4041	1373.2	129.1	0.985	< 0.001	1203.1	0.397	0.008
	0.2	0.8	Adaptive Fused Sparse Group Lasso	1097	1338.9	168.6	0.977	< 0.001	1165.2	0.437	0.003
	0.0	0.8	Adaptive Fused Group Lasso	1424	1477.2	144.2	0.981	< 0.001	1211.6	0.394	0.008

*Note: λ for glmnet R package is scaled by factor n^{-1}.

Mean total sum of squares for training set = 1697.5; Mean total sum of squares for test set = 1428.0. ABIDE: Autism Brain Imaging Data Exchange; FSGL: fused sparse group lasso; CVMSE: cross-validation mean squared error; MSE: mean squared error; r: Pearson correlation.