Stable Anatomy Detection in Multimodal Imaging Through Sparse Group Regularization: A Comparative Study of Iron Accumulation in the Aging Brain

Matthew Pietrosanu[1], Li Zhang[1], Peter Seres[2], Ahmed Elkady[3], Alan H. Wilman[2], Linglong Kong[1]* and Dana Cobzas[4,5]*

[1] Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB, Canada, [2] Department of Biomedical Engineering, University of Alberta, Edmonton, AB, Canada, [3] Department of Biomedical Engineering, McGill University, Montreal, QC, Canada, [4] Department of Computing Science, University of Alberta, Edmonton, AB, Canada, [5] Department of Computer Science, MacEwan University, Edmonton, AB, Canada

Multimodal neuroimaging provides a rich source of data for identifying brain regions associated with disease progression and aging. However, present studies still typically analyze modalities separately or aggregate voxel-wise measurements and analyses to the structural level, thus reducing statistical power. As a central example, previous works have used two quantitative MRI parameters—R2* and quantitative susceptibility (QS)—to study changes in iron associated with aging in healthy and multiple sclerosis subjects, but failed to simultaneously account for both. In this article, we propose a unified framework that combines information from multiple imaging modalities and regularizes estimates for increased interpretability, generalizability, and stability. Our work focuses on joint region detection problems where overlap between effect supports across modalities is encouraged but not strictly enforced. To achieve this, we combine L_1 (lasso), total variation (TV), and L_2 group lasso penalties. While the TV penalty encourages geometric regularization by controlling estimate variability and support boundary geometry, the group lasso penalty accounts for similarities in the support between imaging modalities. We address the computational difficulty in this regularization scheme with an alternating direction method of multipliers (ADMM) optimizer. In a neuroimaging application, we compare our method against independent sparse and joint sparse models using a dataset of R2* and QS maps derived from MRI scans of 113 healthy controls: our method produces clinically-interpretable regions where specific iron changes are associated with healthy aging. Together with results across multiple simulation studies, we conclude that our approach identifies regions that are more strongly associated with the variable of interest (e.g., age), more accurate, and more stable with respect to training data variability. This work makes progress toward a stable and interpretable multimodal imaging analysis framework for studying disease-related changes in brain structure and can be extended for classification and disease prediction tasks.

Keywords: ADMM, geometric regularization, group lasso, joint region, lasso, multiple sclerosis, sparse detection, total variation

1. INTRODUCTION

Following progressive developments in healthcare, the subpopulation of individuals over 60 years old has become the fastest growing worldwide (lr1, 2015). Thanks also in part to rapid advancements in neuroimaging and data storage technology, it is not surprising that much work in recent years has been devoted to studying structural changes in the brain during natural aging and disease progression (e.g., Alzheimer's and Parkinson's diseases Wyss-Coray, 2016 and multiple sclerosis Elkady et al., 2017) as well as differences between healthy and non-normative subjects, particularly in the elderly population.

The naturally high dimensionality of neuroimaging data and their inherent spatial correlation structures present significant challenges to the above goals. Implications of these challenges are apparent in the drawbacks (Davatzikos, 2004) of voxel-based analyses (Ashburner and Friston, 2000) and motivate many of the dimension reduction and regularization techniques commonly applied to neuroimaging data. These methods include pattern identification and discrimination (Fan et al., 2007), which rely on sample sizes not feasible in small neuroimaging studies, and sparse regularization for feature extraction or selection (Batmanghelich et al., 2011; Sabuncu and Van Leemput, 2012; He et al., 2016; Yu et al., 2019; Su et al., 2020). Despite successful applications of the latter to medical data (Krishnapuram et al., 2005; Zou and Hastie, 2005; Ryali et al., 2010), these approaches are unstable with respect to selected or extracted features. In the presence of spatial correlation structures, such as in neuroimaging data where each voxel carries similar information to its neighbors, sparsity comes at the severe detriment of interpretability. To address this, approaches to shape regularization have been developed, notably, total variation (TV) penalties which penalize spatial gradients to encourage a model to apply similar weights to neighboring voxels.

Nonetheless, incorporating multimodal imaging data into analytic techniques remains a difficult problem in general. Previous neuroimaging studies have demonstrated that the simultaneous analysis of multiple imaging modalities can improve statistical power (Elkady et al., 2017), although methods for efficiently doing so are limited. As a result, most studies either analyze different imaging modes independently (Betts et al., 2016) or aggregate voxel-wise observations to the structural level (Cherubini et al., 2009). This approach is inefficient from the perspective of both statistical power and estimate interpretability.

As a focal example and application for this article, we consider specific iron changes in deep gray matter (DGM), which have been histologically associated with healthy aging in the brain (Hallgren and Sourander, 1958). Gradient-echo magnetic resonance imaging (MRI) is used extensively *in vivo* due to its sensitivity to such changes (Peters, 2002). Transverse relaxation rate (R2*) and quantitative susceptibility (QS), both voxel-wise measures derived from gradient-echo MRI, are typically used for this purpose (Haacke et al., 2010; Acosta-Cabronero et al., 2016). Several studies have demonstrated a positive association between iron levels and age in healthy controls using quantitative MRI methods, although they consider only a single MRI measure or anatomical structure (Cherubini et al., 2009; Haacke et al., 2010;

Acosta-Cabronero et al., 2016). The combined use of R2* and QS has only recently been introduced, for the purpose of improving statistical power in delineating iron and non-iron changes in studies of multiple sclerosis (Elkady et al., 2017). This earlier study, however, only combines modalities in an initial feature selection stage rather than in a single model.

Related works have similarly investigated, with the previously-discussed methodological limitations, the complex associations between QS and R2* measures, iron, and myelin in brain structures. Through a biochemical and histological analysis of postmortem brains using Pearson correlation (e.g., between myelin and R2*) and linear mixed effect regression (e.g., of R2* on iron and myelin intensities) on region- or structurally-aggregated measures, Hametner et al. (2018) showed that R2* is only weakly sensitive to myelin in white matter regions that do not contain iron, but is strongly sensitive to iron in the basal ganglia and white matter regions that do contain iron. The authors further demonstrated that QS also has mixed sensitivities to iron and myelin in white matter. Taege et al. (2019) applied region of interest-based (i.e., structurally-aggregated) regression analyses using R2* and QS to investigate deep gray matter microstructure in the context of multiple sclerosis. Mangeat et al. (2015) used two-dimensional independent components analysis of magnetization transfer, cortical thickness, B_0 orientation, and R2* to extract information about the fraction of myelin in white matter. Bergsland et al. (2018) used QS and diffusion metrics to perform separate univariate statistical analyses of thalamic white matter tracts and used non-parametric correlation analysis to combine these measures and quantify their association with clinical disease metrics.

While all the previous works provide useful insight into the structure and composition of the brain, the use of region-based analyses relies of the typically incorrect assumption of intra-region homogeneity. Furthermore, these works generally do not jointly incorporate multiple imaging modes as model covariates. These are general limitations in the application of traditional statistical methods to high-dimensional neuroimaging data. In this work, we aim to address these limitations in the context of region-detection problems.

In this article, we propose a unified, multimodal framework for joint region detection associated with a numerical variable of interest such as age. We formulate this problem in terms of sparse regression with L_1 (lasso), TV, and L_2 group lasso penalties, with the last of these combining information across multiple modalities. These penalties have been widely used in the development of robust methods for medical data (He et al., 2016; Yu et al., 2019). Together, the penalties regularize spatial effect estimates for better interpretability, both within and between modalities. This work is concerned primarily with joint region detection, such as in our focal R2*-QS-age problem, where overlap between effect supports across modalities is encouraged but not strictly enforced. To address the computational difficulty in implementing these penalties together, we propose an optimization procedure using the alternating direction method of multipliers (ADMM) algorithm.

Our work addresses the need for multimodal neuroimaging techniques for joint region detection that can provide

interpretable estimates accounting for similarities in spatial effect supports. In our focal example, it is critical to account for both R2* and QS since, voxel-wise, both are positively associated with iron accumulation (Langkammer et al., 2012): the increased statistical power in multimodal MRI analysis has previously detected subtle pathological changes in neurological diseases such as multiple sclerosis (Elkady et al., 2017) and will similarly improve the detection of aging-related neurodegenerative effects. Generally, our approach emphasizes the interpretability and stability of support estimates in terms of spatial smoothness and variability (with respect to perturbations in training data), respectively. Our estimators are much less affected by spatial correlation structures that would otherwise result in severe undersegmentation in other sparse analyses. Furthermore, our work easily accommodates multiple anatomical regions simultaneously: this contrasts with traditional voxel-based analyses, which are generally known to have unsuitable performance when considering multiple regions due to their disregard for spatial correlation (Ashburner and Friston, 2000). In the limited literature addressing multimodal data (Michel et al., 2011; Gramfort et al., 2013), our formulation using sparse regularization is novel.

2. MATERIALS AND METHODS

2.1. Sparse Regression and Image Regularization

We first present more detail on the components of our proposed framework and define the two other models against which the proposed method will be compared. For notational simplicity, we ignore intercept terms in the following subsections and only consider two modalities, although our setup generalizes immediately.

Let X_1 and X_2 be $n \times p$ matrices corresponding to the two imaging modalities (e.g., R2* and QS) after columnwise standardization. These matrices have i-th row $\boldsymbol{x}_{1i}^\top$ and $\boldsymbol{x}_{2i}^\top$, respectively, corresponding to vectorized, observed image data for the i-th subject. In general, p can be smaller than the number of voxels in the image due to the application of a mask (that is constant across subjects). Let $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ denote corresponding voxel effects that are to be estimated, and $\boldsymbol{y}$ an observed length-n vector of a continuous variable corresponding to the n subjects. Loosely, we are interested in associations of X_1 and X_2 with $\boldsymbol{y}$, where $p \gg n$.

2.1.1. Sparse Regression

The traditional (least squares) linear regression problem in this setting takes the form $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2) = \arg\min_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2} \|\boldsymbol{y} - X_1\boldsymbol{\beta}_1 - X_2\boldsymbol{\beta}_2\|_2^2$. This estimator is unstable and has extremely high variability, attributable to overfitting in this high-dimensional setting. Furthermore, this estimator typically yields estimates with only non-zero entries, making it useless for region identification. To address both these problems, we introduce a sparsity assumption on $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$ by adding an L_1 (lasso) penalty to the previous problem, as an approximation of the L_0 penalty used in (NP-hard) best subset selection (Tibshirani, 1996; Huo and Ni, 2007). The lasso penalty is widely-established in

both statistical theory and application as a method for variable selection (Tibshirani, 1996; Tibshirani et al., 2005; Simon et al., 2013): loosely-speaking, when applied in 3D imaging contexts, this penalty identifies voxels of interest by shrinking entries of $\hat{\boldsymbol{\beta}}$ corresponding to "non-significant" voxels to zero.

The sparse regression problem becomes

$$\arg\min_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2} \|\boldsymbol{y} - X_1\boldsymbol{\beta}_1 - X_2\boldsymbol{\beta}_2\|_2^2 + \lambda_1\|\boldsymbol{\beta}_1\|_1 + \lambda_1\|\boldsymbol{\beta}_2\|_1, \quad (1)$$

where the hyperparameter λ_1 controls the balance between model fidelity and model sparsity. We use the same hyperparameter λ_1 for both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ since X_1 and X_2 are standardized.

While the sparse regression problem in (1) is typically encountered in models for predicting y, that is not our primary interest here. We are instead concerned with associations of X_1 and X_2 with $\boldsymbol{y}$: in a neuroimaging context, we wish to identify regions with MRI measures X that vary together with age $\boldsymbol{y}$. In particular, the unpenalized, least squares model maximizes the square sample correlation between y_i and $\boldsymbol{x}_{1i}^\top\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}^\top\boldsymbol{\beta}_2$, but requires further regularization to become practically usable.

2.1.2. TV Regularization

While the previous sparsity assumption solves initial problems in estimating $\boldsymbol{\beta}$, it does not suitably address estimate interpretability. In neuroimaging contexts, since measures from nearby voxels tend to be correlated, $\hat{\boldsymbol{\beta}}$ will select isolated voxels rather than contiguous, compact regions that are more amenable to interpretation (Kandel et al., 2013; Dubois et al., 2014; Eickenberg et al., 2015).

To encourage interpretability, we add an image-based penalty to (1). Both TV (Michel et al., 2011; Baldassarre et al., 2012; Gramfort et al., 2013; Dohmatob et al., 2014; Dubois et al., 2014; Eickenberg et al., 2015) and GraphNet (Ng et al., 2010; Grosenick et al., 2013; Kandel et al., 2013; Watanabe et al., 2014) penalties have been explored in sparse regression and classification contexts: we use the former as it has demonstrated superior performance with respect to variable selection (Michel et al., 2011; Gramfort et al., 2013; Eickenberg et al., 2015). The anisotropic TV penalty (Tibshirani et al., 2005; Watanabe et al., 2014) is an L_1 penalty on the gradient of an image, which we denote by

$$\|\nabla\boldsymbol{\beta}\|_1 = \|\nabla_1\boldsymbol{\beta}\|_1 + \|\nabla_2\boldsymbol{\beta}\|_1 + \|\nabla_3\boldsymbol{\beta}\|_1,$$

where ∇_i (for $i = 1, 2, 3$) denotes a gradient taken along the i-th orthogonal coordinate direction. The resulting penalized problem is

$$\arg\min_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2} \|\boldsymbol{y} - X_1\boldsymbol{\beta}_1 - X_2\boldsymbol{\beta}_2\|_2^2 + \lambda_1\|\boldsymbol{\beta}_1\|_1 + \lambda_1\|\boldsymbol{\beta}_2\|_1$$
$$+ \lambda_2\|\nabla\boldsymbol{\beta}_1\|_1 + \lambda_2\|\nabla\boldsymbol{\beta}_2\|_1, \quad (2)$$

where λ_2 is another hyperparameter.

2.1.3. Sparse Group Lasso Regularization

The TV penalty in (2) addresses the interpretability of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ estimates separately rather than jointly. The latter is a

natural consideration in neuroimaging when interested in the joint support, e.g., regions that are associated with age for both R2* and QS modalities, or if prior knowledge suggests that the effect supports should overlap. We introduce an additional sparse group lasso regularization term to this end.

Group lasso methods apply a penalty to pre-specified groups of variables and are widely used. To encourage overlap in the two estimate supports, we consider p groups of size 2, each composed of voxel-wise measures at the same location. The sparse group lasso penalty (Simon et al., 2013) in this case is given by

$$\lambda_3 \sum_{j=1}^{p} \sqrt{\beta_{1j}^2 + \beta_{2j}^2}.$$

In contrast, the standard group lasso (Tibshirani, 1996) applies an L_1 penalty to the estimates in each group rather than the L_2 penalty above, forcing all estimates in a group to be both either zero or non-zero. This is an undesireably hard constraint in our setting.

The resulting penalized problem and our proposed estimator is given by

$$\underset{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2}{\arg\min} \|\boldsymbol{y} - X_1 \boldsymbol{\beta}_1 - X_2 \boldsymbol{\beta}_2\|_2^2 + \lambda_1 \|\boldsymbol{\beta}_1\|_1 + \lambda_1 \|\boldsymbol{\beta}_2\|_1$$

$$+\lambda_2 \|\nabla \boldsymbol{\beta}_1\|_1 + \lambda_2 \|\nabla \boldsymbol{\beta}_2\|_1 + \lambda_3 \sum_{j=1}^{p} \sqrt{\beta_{1j}^2 + \beta_{2j}^2}. \quad (3)$$

2.2. Optimization

The optimization problem in (3) is difficult to solve due to the non-differentiability of the L_1 norm appearing in the lasso and TV penalties, to which standard gradient-based methods do not apply. Various numerical methods and have been explored for settings with these two penalties, including for logistic regression (Dohmatob et al., 2014).

In this work, we apply the ADMM algorithm (Boyd et al., 2011), an efficient convex optimization scheme amenable to parallel computing. Its computational benefits rely on splitting a given problem into two convex sub-problems. In the present case, denoting $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$, (3) can be written as

$$\underset{\boldsymbol{\beta}, \boldsymbol{\alpha}}{\arg\min} \, L(\boldsymbol{\beta}) + \|\Lambda^\top \boldsymbol{\alpha}\|_1 \quad (4)$$

$$\text{subject to } \boldsymbol{\beta} = A\boldsymbol{\alpha},$$

where $L(\boldsymbol{\beta}) = \|\boldsymbol{y} - X_1 \boldsymbol{\beta}_1 - X_2 \boldsymbol{\beta}_2\|_2^2 + \lambda_3 \sum_{i=1}^{p} \sqrt{\beta_{1j}^2 + \beta_{2j}^2}$ is a smooth function of $\boldsymbol{\beta}$, $A = [I_p \mid D^\top]^\top$, $\Lambda = [\lambda_1 I_p \mid \lambda_2 I_{3p}]^\top$, and D is the 3-dimensional differential operator (Rohr, 1997).

The ADMM algorithm solves (4) using the iterative updates

$$\boldsymbol{\beta}^{(t+1)} = \underset{\boldsymbol{\beta}}{\arg\min} L_\rho(\boldsymbol{\beta}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\eta}^{(t)})$$

$$\boldsymbol{\alpha}^{(t+1)} = \underset{\boldsymbol{\alpha}}{\arg\min} L_\rho(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\alpha}, \boldsymbol{\eta}^{(t)})$$

$$\boldsymbol{\eta}^{(t+1)} = \boldsymbol{\eta}^{(t)} + \rho(\boldsymbol{\alpha}^{(t+1)} - A\boldsymbol{\beta}^{(t+1)}),$$

where $\boldsymbol{\eta}$ is an auxiliary variable and $L_\rho(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = L(\boldsymbol{\beta}) + \|\Lambda^\top \boldsymbol{\alpha}\|_1 + \boldsymbol{\eta}^\top(\boldsymbol{\beta} - A\boldsymbol{\alpha}) + \frac{\rho}{2}\|\boldsymbol{\beta} - A\boldsymbol{\alpha}\|_2^2$ is the augmented Lagrangian with parameter ρ.

The $\boldsymbol{\beta}$-update step minimizes a smooth function (noting that $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ are held fixed) and can be performed using standard first-order methods. The $\boldsymbol{\alpha}$-update step, on the other hand, features a non-smooth objective function containing both L_2 and L_1 penalties on $\boldsymbol{\alpha}$. This update admits a closed-form solution using the soft thresholding operator $S_\lambda(x) = \text{sgn}(x)(|x| - \lambda)$, namely,

$$\alpha_i^{(t+1)} = S_{\Lambda_{ii}} \left((A\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\eta}^{(t)})_i \right).$$

For termination criterion, we follow the suggestion of Boyd et al. (2011) based on the problem's primal residuals $\boldsymbol{r}_{\text{primal}}^{(t)} = \boldsymbol{\alpha}^{(t)} - A\boldsymbol{\beta}^{(t)}$ and dual residuals $\boldsymbol{r}_{\text{dual}}^{(t)} = \rho A^\top(\boldsymbol{\alpha}^{(t)} - \boldsymbol{\alpha}^{(t-1)})$, namely,

$$\|\boldsymbol{r}_{\text{primal}}^{(t)}\|_2 \le \epsilon_{\text{primal}} \quad \text{and} \quad \|\boldsymbol{r}_{\text{dual}}^{(t)}\|_2 \le \epsilon_{\text{dual}},$$

for positive primal and dual tolerances ϵ_{primal} and ϵ_{primal}, respectively.

2.3. Models and Hyperparameter Tuning

In all subsequent analyses, we compare performance between three models. The first is an independent sparse (IS) model, composed of one model for each imaging modality, each fit with lasso and TV regularization. For this model, "predictions" for $\boldsymbol{y}$ are taken as averages across the independent models. Second, we consider a joint sparse (JS) model, given in (2), including all imaging modalities together with lasso and TV penalties. Third, we consider our proposed sparse group lasso (SGL) method in (3) including all imaging modalities and lasso, TV, and sparse group lasso regularization.

Generally, three hyperparameters λ_1, λ_2, and λ_3 require tuning. We reparameterize these in terms of λ, $r_{1,2}$, and r_3 as

$$(\lambda_1, \lambda_2, \lambda_3) = \lambda(r_{1,2}(1 - r_3), (1 - r_{1,2})(1 - r_3), r_3)$$
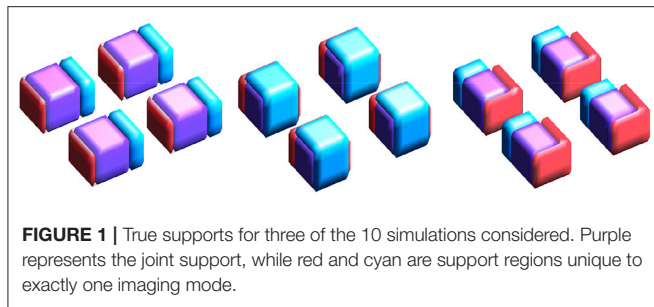
and use the Bayesian information criterion (BIC) (Fan and Tang, 2013), calculated on the full training dataset, as a performance criterion to choose optimal hyperparameter values. In all analyses, results using the generalized information criterion (GIC) (Fan and Tang, 2013) are similar.

In the following simulation studies, $(\lambda, r_{1,2}, r_3)$ are tuned on a grid over $[e^{-9}, 0] \times [0.1, 0.9] \times [0, 0.9]$. As further restrictions, the IS model enforces $r_3 = 0$ and the SGL model restricts $r_{1,2} = 0.05$ and $r_3 = 0.6$. Tuning in the subsequent neuroimaging analysis restricts the maximum value of λ to e^{-1}, but is otherwise identical.

2.4. Data, Analyses, and Evaluation
2.4.1. Synthetic Data and Simulation Study
Because ground truth for the support regions is not available in real-world imaging, we use synthetic data to more comprehensively investigate the performance of our proposed method. Our setup generalizes that in Zhang et al. (2018) by including a second imaging modality and inter-modality

FIGURE 1 | True supports for three of the 10 simulations considered. Purple represents the joint support, while red and cyan are support regions unique to exactly one imaging mode.

FIGURE 2 | Deep gray matter (DGM) structures considered in this study. The DGM mask applied to the data is shown in gray in all sub-panels. The caudate (blue), putamen (red), thalamus (green), and globus pallidus (purple) are highlighted in each sub-panel.

correlation. In all simulation studies, we use training datasets of size $n = 100$, which is comparable to the $n = 113$ used in our real-world MRI application.

Imaging data for the first mode is generated with size $32 \times 32 \times 8$. The true support is represented by four disjoint $8 \times 8 \times 4$ regions (which we call "blocks"). Imaging data for the second mode is generated similarly: however, to simulate (partial) overlap in the support between the two modes, we allow block size for the second mode to vary uniform randomly over $\{6, 7, 8, 9, 10\}^2 \times \{4, 5\}$. Example supports are provided in **Figure 1**.

Background noise is independently generated from a $\mathcal{N}(0, I)$ distribution. Voxel measures within a block are correlated only with voxels from the same block and from the corresponding block in the other mode. This inter-voxel correlation is controlled by a spatial coherence parameter ρ that decreases with L_0 distance, and is further multiplied by an inter-mode correlation parameter η for measures in different modes.

The true effect $\boldsymbol{\beta}$ is set to 0.1, 0.2, 0.3, or 0.4 in each block, for varying signal strength, and 0 elsewhere. The simulated response is generated as $\boldsymbol{y} = 0.01 + X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, I_n)$, following the notation in section 2.

We consider two small-sample ($n = 100, p = 32 \times 32 \times 8 \times 2 = 16384$) simulation settings: a "low correlation" ($\rho = 0.5, \eta = 0.2$) and a "high correlation" setting ($\rho = 0.8, \eta = 0.5$). In each case, we run two analyses. The first examines overall performance averaged over 10 simulations, each using independently-generated true supports. The second examines estimate stability over five independent datasets generated using the same true support (i.e., for 10 possible inter-simulation comparisons). From here on, we refer to these two analyses as simulation study #1 and simulation study #2, respectively.

Intuitively, simulation study #1 assesses average performance across a variety of ground truths. On the other hand, simulation study #2 assesses estimate stability across different training datasets for a single, fixed ground truth.

2.4.2. Neuroimaging Data, Pre-processing, and Analyses

The neuroimaging data used in this work is from an in-house study of multiple sclerosis and controls (Walsh et al., 2014; Elkady et al., 2017), similar to that used in Zhang et al. (2018) but composed of only $n = 113$ control scans (mean age 40.25 with SD 10.91, minimum 21.8, and maximum 65.1), of which 40 were obtained from male subjects. Our primary interest is in
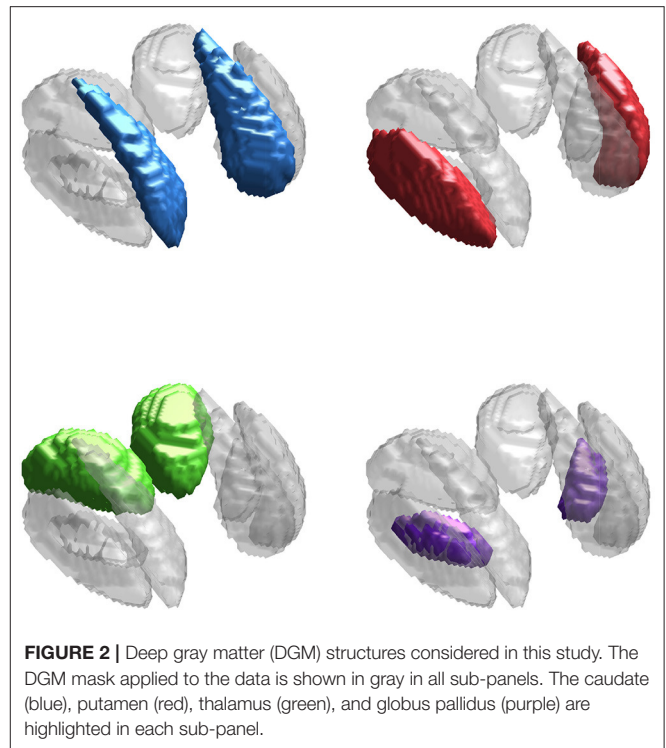
identifying regions of the brain in which variability in iron levels is associated with age. We focus on four subcortical deep gray matter (DGM) structures: the caudate, putamen, thalamus, and globus pallidus, shown in **Figure 2** (Zhang et al., 2017, 2018). We use R2* and QS modalities, as both are known to be highly sensitive to changes in non-heme iron (Wang and Liu, 2015).

Imaging was conducted using a 4.7 T system (Varian Inova, Palo Alto, CA), with two imaging sequences obtained per session. R2* and QS maps were calculated using three-dimensional multi-echo gradient echo acquisitions (acquisition time = 9.4 min, repetition time 44 ms, number of echoes = 10, time to first echo = 2.93 ms, echo spacing = 4.1 ms, monopolar readout, flip angle = $10°$, number of contiguous slices = 80, field of view = $160 \times 256 \times 160$ mm^3, voxel size $1 \times 1 \times 2$ mm^3). R2* maps were computed using a mono-exponential temporal fit of image magnitude, while QS maps were calculated using an image phase inversion (Bilgic et al., 2012) following phase unwrapping using FSL PRELUDE and background field removal using regularization-enabled sophisticated harmonic removal of phases (RESHARP) (Sun and Wilman, 2013). An anatomical MRI sequence was acquired by 3D T1w volumetric imaging using magnetization-prepared rapid gradient-echo (MPRAGE) (acquisition time = 4.8 min, flip angle = $108°$, TE/TR = 4.5/8.5 ms, inversion time to start of readout = 300 ms, sequential phase encoding, number of contiguous slices = 84, voxel size = $0.9 \times 0.9 \times 2$ mm^3).

Prior to analysis, the MRI data was pre-processed and aligned with an in-house unbiased template using ANTs (ANTS, 2011), built from T1w, R2*, and QS data from 10 healthy controls. Pre-processing involved intra-subject alignment of the R2* and QS

TABLE 1 | Performance results for simulation study #1, averaged over 10 independent simulations.

Method	Setting	MAE		Test R^2	Dice	
		Train	Test		Full	Joint
IS	LC	0.187 (0.140)	29.286 (7.895)	0.732 (0.144)	0.427 (0.236)	0.552 (0.152)
	HC	0.147 (0.102)	27.840 (7.334)	0.762 (0.139)	0.532 (0.151)	0.596 (0.128)
JS	LC	**0.127 (0.056)**	29.060 (6.078)	0.734 (0.124)	0.380 (0.148)	0.537 (0.112)
	HC	**0.127 (0.056)**	29.066 (6.072)	0.734 (0.123)	0.380 (0.149)	0.537 (0.113)
SGL	LC	0.128 (0.055)	**8.310 (4.682)**	**0.998 (0.001)**	**0.641 (0.107)**	**0.824 (0.032)**
	HC	0.129 (0.055)	**8.309 (4.682)**	**0.998 (0.001)**	**0.641 (0.107)**	**0.823 (0.031)**

"LC" and "HC" denote the low correlation ($\rho = 0.5$, $\eta = 0.5$) and high correlation ($\rho = 0.8$, $\eta = 0.5$) settings, respectively. "IS," "JS," and "SGL" refer to the independent sparse, joint sparse, and proposed sparse group lasso models, respectively. All results are presented in the form "mean (SD)." Entries with the best mean results for the LC and HC settings in each column are in bold.

maps with the T1w map. Non-linear registration in the template space was accomplished by applying SyN (ANTS, 2011) to the multimodal MRI data. Observation row vectors $\boldsymbol{x}_i^\top$ were obtained by taking voxels within a DGM mask manually traced on the atlas, shown in **Figure 5**. Data matrix columns were standardized before analysis.

2.5. Evaluation Methodology

For simulation study #1, we report prediction mean absolute error (MAE) on the training set ($n = 100$) and both MAE and R^2 on the independent testing set ($n = 500$). Since the true support is known for synthetic data, we also report Dice scores between the estimated and true supports (called "full Dice") as well as between the estimated and true joint supports (called "joint Dice"). Here, "joint" refers to the region of overlap in the supports of the two imaging modes.

For simulation study #2, we additionally report Dice scores between estimated supports obtained from different training datasets. This performance measure quantifies the stability of the estimated support with respect to training set variability.

In the neuroimaging analysis, we consider a 23-fold cross-validation approach due to the small sample size. We report training and testing set MAE, testing set R^2, and pairwise Dice scores between the supports estimated using each training dataset.

3. RESULTS

3.1. Simulation Studies

Results for simulation study #1 are provided in **Table 1**. **Figure 3** gives example visualizations of the estimated and true supports for one simulation. These results are similar between the high and low correlation settings. Full and joint Dice scores are significantly higher for the proposed SGL model, with 1.4–4.9 times lower variability across simulations, suggesting more accurate and stable support estimation when using the proposed SGL method. **Figure 3** supports these conclusions and illustrates how the SGL estimator more clearly and accurately identifies the joint support: in contrast, IS and JS estimates suggest higher false-positive voxel selection rates. These points are visually evident
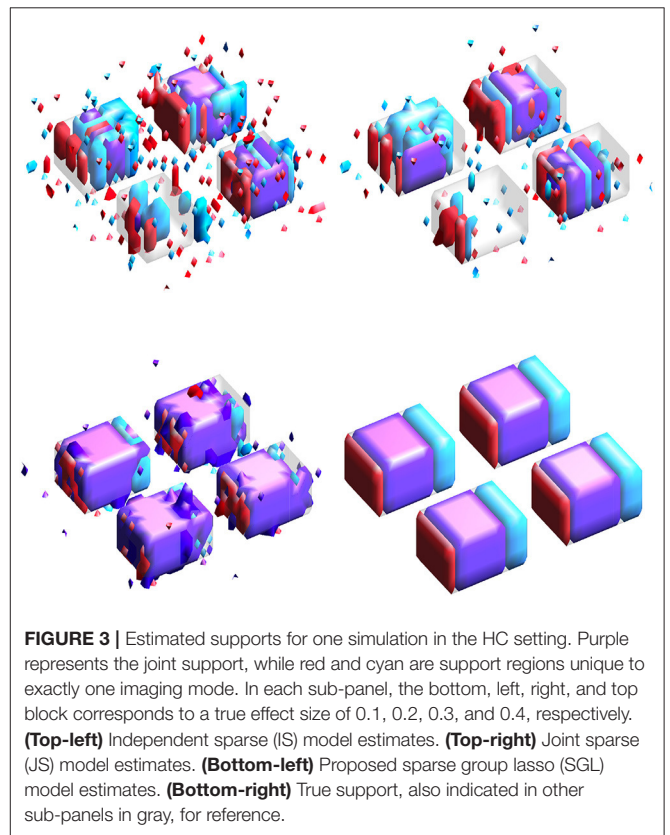

FIGURE 3 | Estimated supports for one simulation in the HC setting. Purple represents the joint support, while red and cyan are support regions unique to exactly one imaging mode. In each sub-panel, the bottom, left, right, and top block corresponds to a true effect size of 0.1, 0.2, 0.3, and 0.4, respectively. **(Top-left)** Independent sparse (IS) model estimates. **(Top-right)** Joint sparse (JS) model estimates. **(Bottom-left)** Proposed sparse group lasso (SGL) model estimates. **(Bottom-right)** True support, also indicated in other sub-panels in gray, for reference.

though the differences in the size of the joint region (shown in purple) detected by the SGL and the IS/JS estimates, the noise present in the IS (and JS, to a lesser extent) estimate unique to one imaging modality (indicated in red and cyan), and the notable deterioration of IS/JS estimate quality when the true effect size is small. While not the primary goal, the proposed SGL model retains excellent association with the response, shown in testing set MAE and R^2: as hypothesized, this suggests that the proposed SGL model generalizes better to new data and that the
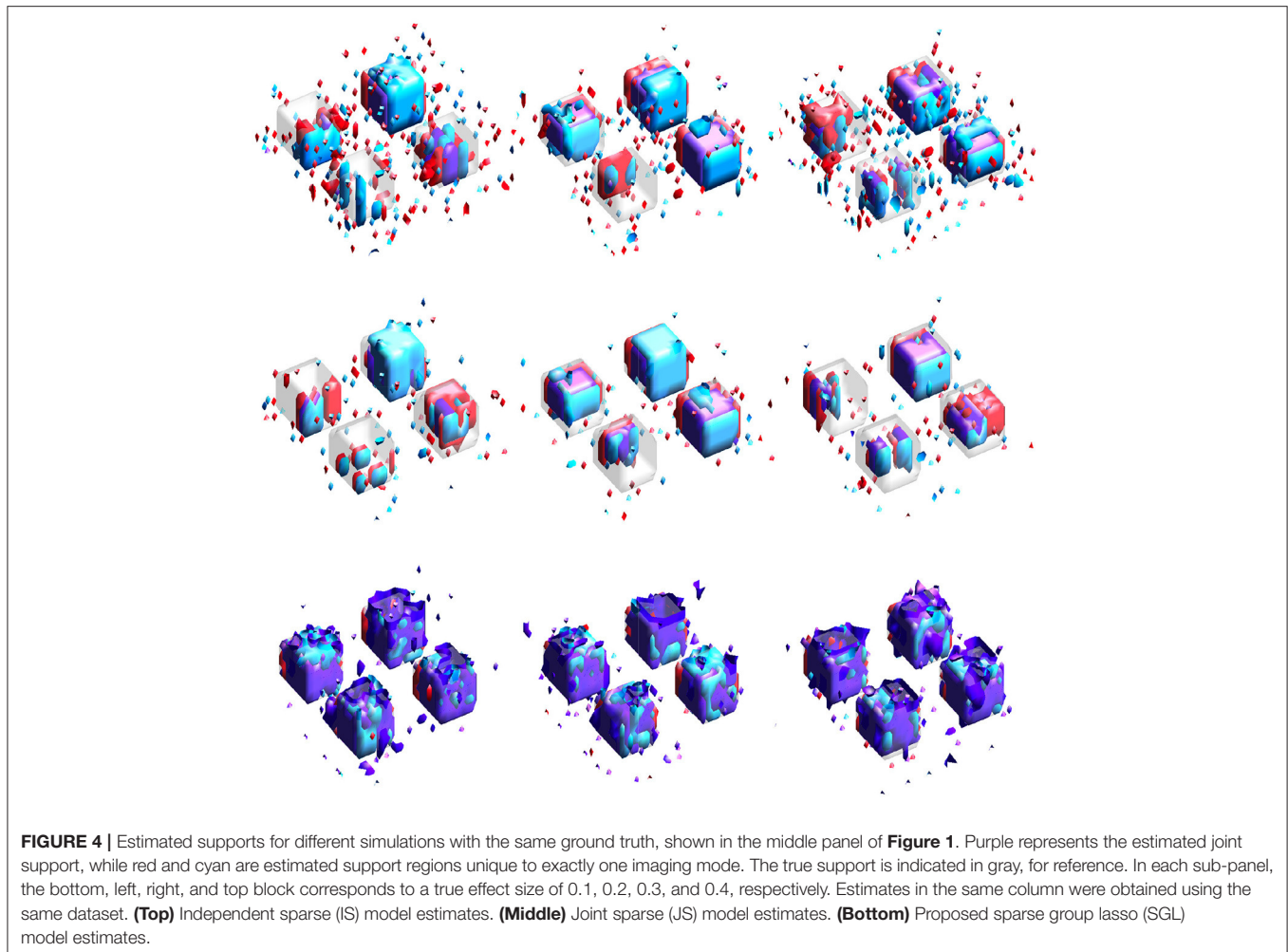
FIGURE 4 | Estimated supports for different simulations with the same ground truth, shown in the middle panel of **Figure 1**. Purple represents the estimated joint support, while red and cyan are estimated support regions unique to exactly one imaging mode. The true support is indicated in gray, for reference. In each sub-panel, the bottom, left, right, and top block corresponds to a true effect size of 0.1, 0.2, 0.3, and 0.4, respectively. Estimates in the same column were obtained using the same dataset. **(Top)** Independent sparse (IS) model estimates. **(Middle)** Joint sparse (JS) model estimates. **(Bottom)** Proposed sparse group lasso (SGL) model estimates.

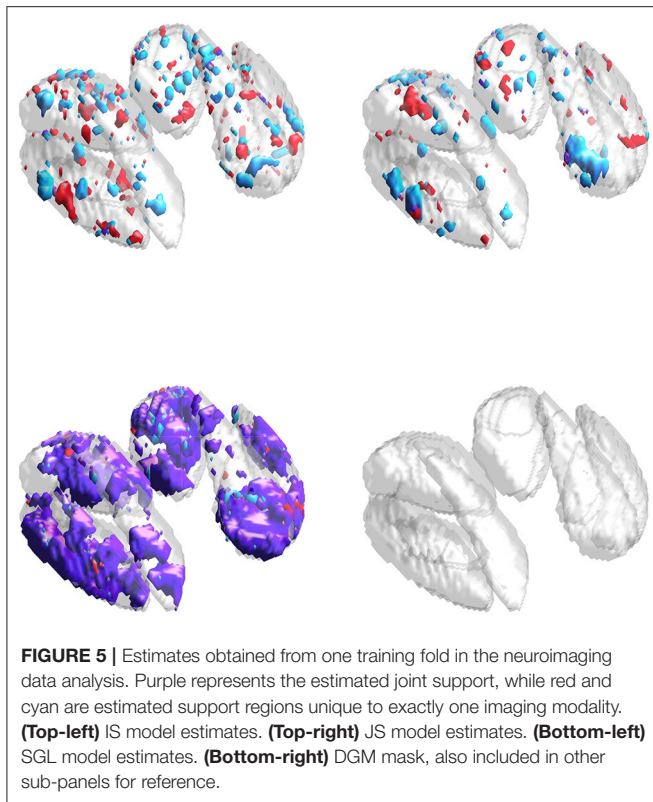**TABLE 2 |** Performance results for simulation study #2.

Method	Setting	MAE		Test R^2	Dice (truth)		Dice (pairwise)	
		Train	**Test**		**Full**	**Joint**	**Full**	**Joint**
IS	LC	0.207 (0.129)	25.940 (6.789)	0.811 (0.056)	0.641 (0.069)	0.659 (0.097)	0.579 (0.072)	0.688 (0.080)
	HC	0.207 (0.129)	25.940 (6.789)	0.811 (0.056)	0.641 (0.069)	0.659 (0.097)	0.579 (0.072)	0.688 (.080)
JS	LC	**0.112 (0.026)**	26.426 (6.902)	0.791 (0.067)	0.609 (0.110)	0.568 (0.190)	0.556 (0.076)	0.541 (0.153)
	HC	**0.112 (0.026)**	26.426 (6.902)	0.791 (0.067)	0.609 (0.110)	0.568 (0.190)	0.556 (0.076)	0.541 (0.153)
SGL	LC	0.144 (0.040)	**2.234 (0.402)**	**0.999 (0.001)**	**0.832 (0.012)**	**0.700 (0.024)**	**0.835 (0.007)**	**0.760 (0.011)**
	HC	0.141 (0.042)	**2.241 (0.404)**	**0.999 (0.001)**	**0.831 (0.011)**	**0.700 (0.021)**	**0.835 (0.006)**	**0.760 (0.009)**

"LC" and "HC" denote the low correlation ($\rho = 0.5, \eta = 0.5$) and high correlation ($\rho = 0.8, \eta = 0.5$) settings, respectively. "IS," "JS," and "SGL" refer to the independent sparse, joint sparse, and proposed sparse group lasso models, respectively. All results are presented in the form "mean (SD)." Entries with the best mean results for the LC and HC settings in each column are in bold.

estimated relationship (in $\hat{\boldsymbol{\beta}}$) between voxel-wise measures and the response is less-affected by over-fitting to the training dataset, relative to the IS and JS models.

Results for simulation study #2 are provided in **Table 2**. **Figure 4** gives example visualizations of the estimated supports (obtained from different training datasets) and the true support for one simulation. These results agree with the conclusions of simulation study #1: the proposed SGL method more accurately and stably selects the true support (either full or joint) while retaining high association with the response. Despite a more

FIGURE 5 | Estimates obtained from one training fold in the neuroimaging data analysis. Purple represents the estimated joint support, while red and cyan are estimated support regions unique to exactly one imaging modality. **(Top-left)** IS model estimates. **(Top-right)** JS model estimates. **(Bottom-left)** SGL model estimates. **(Bottom-right)** DGM mask, also included in other sub-panels for reference.

complex true support, **Figure 4** demonstrates the stability of SGL estimates across independent training sets. This contrasts with IS and JS estimates, which again show higher false-positive voxel selection rates (visually, with red, cyan, or purple outside of the gray true support) and struggle to identify the joint support (visually, with estimates in purple), particularly when the true signal is weak.

3.2. Neuroimaging Data Analysis

Results for the neuroimaging analysis are given in **Table 3**. Examples of the estimated supports are provided for one training fold in **Figure 5** and for multiple folds in **Figure 6**. While no ground truth is available, these results support the simulation studies' conclusions. **Figure 5** illustrates the expected behavior of typical sparse methods in real-world settings: both the IS and JS estimates show many small, non-contiguous regions of selected voxels (although the JS estimates are somewhat smoother). More importantly, the IS and JS estimates suggest little to no joint support (in purple), which is clearly not consistent with findings from previous neuroimaging studies. In contrast, the SGL estimates show a substantial joint support attributable to the proposed sparse group penalty. Furthermore, this region is reasonably smooth, amenable to clinical interpretation, and stable across the (nearly identical) training folds (shown in **Figure 6**), unlike the IS and JS estimates. This conclusion regarding the stability of SGL estimates is supported in **Table 3** by the higher Dice scores (both full and joint) between estimates obtained from different training folds. Numerically, all estimated

models retain a comparable and moderate association between predicted and true age (shown by validation set MAE and R^2 in **Table 3**).

4. DISCUSSION

4.1. Summary and Key Findings

Our simulation studies and analysis of a real-world neuroimaging dataset consistently supported our proposed SGL approach over both IS and JS models. In the simulation studies, SGL estimates of the full support gave notably higher (estimate-truth) Dice scores with lower variation across simulations. This result was strengthened when considering joint support detection and can be attributed to our inclusion of a sparse group lasso penalty. **Figures 3**, **4** illustrate the high false-positive rate and instability of IS and JS estimates and the difficulty they have in obtaining the joint support, especially for weaker signals. SGL estimates do not suffer in these ways and still recover the joint support even if the signal is weaker. A second set of simulation studies over independent datasets with the same ground truth similarly confirmed that our proposed method provides more stable estimates.

The neuroimaging analysis highlighted a major disadvantage of typical sparse methods (e.g., IS and JS), namely, their tendency to select a small number of voxels in non-contiguous regions, shown clearly in **Figure 5**. Furthermore, neither IS nor JS estimates suggested the presence of a joint support region, inconsistent with previous neuroimaging studies (Peters, 2002; Langkammer et al., 2012). These estimates again showed high instability, even with minimal variation in the training dataset during 23-fold cross validation (with total sample size $n = 113$). In contrast, our SGL method estimates a substantial joint support that was reasonably smooth, compact, and stable across training folds.

In both the simulation and neuroimaging analyses, all methods demonstrate a difference between the training and testing (or validation) set MAE, as expected. We attribute variation in this difference between the IS/JS and SGL models to the effect of overfitting when using the former methods, which is emphasized in the simulation studies (with about 82–90% of truly non-predictive voxels, much greater than the 32–50% estimated by the SGL model in the neuroimaging analysis). We suspect this variation is lessened in the neuroimaging analysis due to the naturally more complex data for which our simulation is a simple imitation.
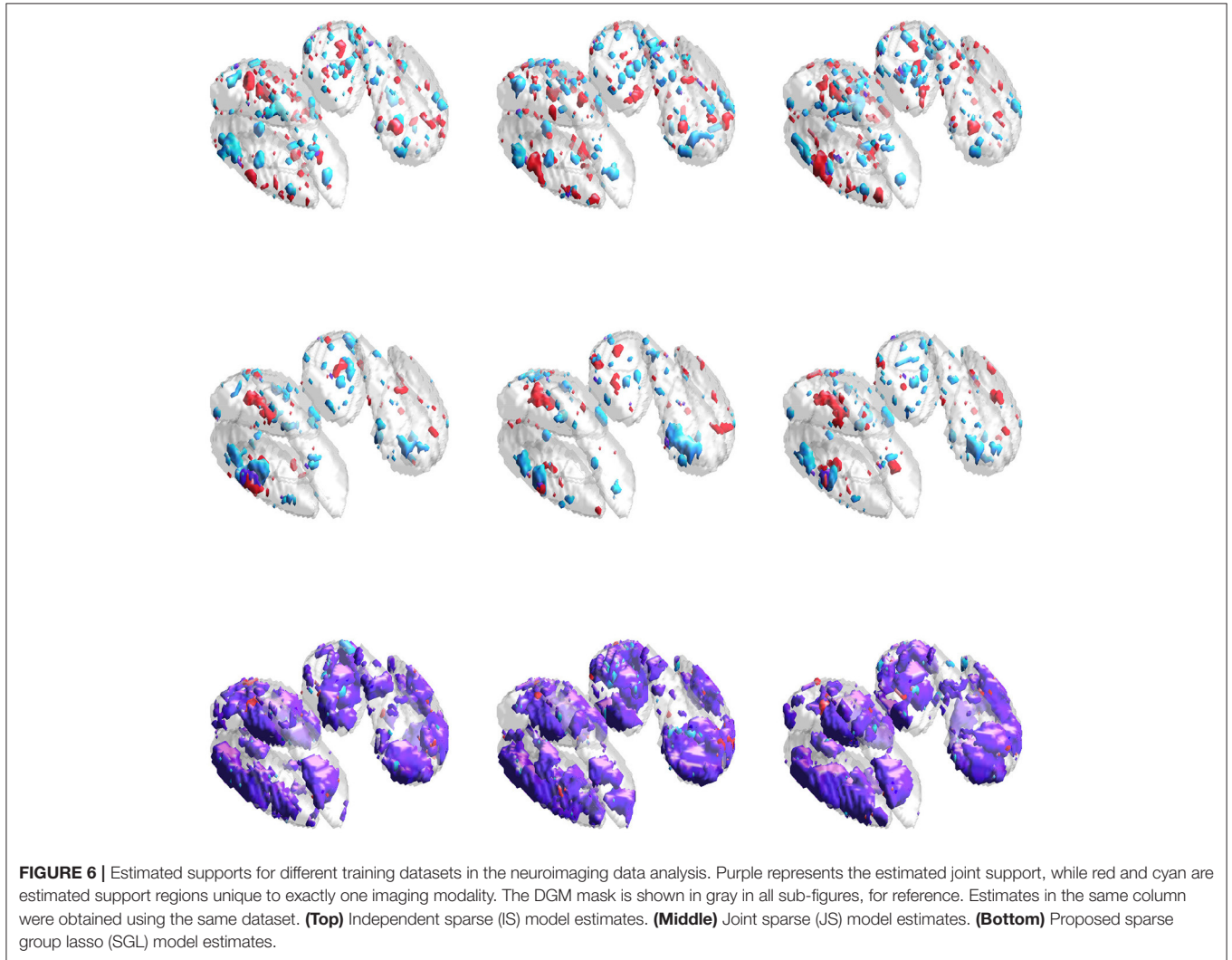
4.2. Related Work and Impact

Existing approaches to multimodal neuroimaging data analysis commonly use techniques requiring a large sample size (Fan et al., 2007), analyze modes separately (Cherubini et al., 2009), aggregate voxel-wise measures to the structural level (Betts et al., 2016), or combine modalities in only an initial feature selection stage (Elkady et al., 2017). These approaches are either impractical or come at the cost of statistical power.

In this article, we proposed a unified, multimodal framework for joint region detection associated with a numerical variable of interest. Our work addresses the above limitations in existing

TABLE 3 | Performance results for the neuroimaging data analysis for 23-fold cross-validation.

Method	MAE		Validation R^2	Dice (pairwise)	
	Train	**Validation**		**Full**	**Joint**
IS	**0.193 (0.051)**	**4.548 (1.563)**	**0.845 (0.171)**	0.529 (0.062)	0.395 (0.114)
JS	0.607 (0.188)	5.515 (1.640)	0.784 (0.183)	0.519 (0.070)	0.308 (0.131)
SGL	0.648 (0.117)	5.059 (1.967)	0.824 (0.150)	**0.678 (0.050)**	**0.666 (0.049)**

"IS," "JS," and "SGL" refer to the independent sparse, joint sparse, and proposed sparse group lasso models, respectively. All results are presented in the form "mean (SD)". Entries with the best mean results in each column are in bold.

FIGURE 6 | Estimated supports for different training datasets in the neuroimaging data analysis. Purple represents the estimated joint support, while red and cyan are estimated support regions unique to exactly one imaging modality. The DGM mask is shown in gray in all sub-figures, for reference. Estimates in the same column were obtained using the same dataset. **(Top)** Independent sparse (IS) model estimates. **(Middle)** Joint sparse (JS) model estimates. **(Bottom)** Proposed sparse group lasso (SGL) model estimates.

neuroimaging analytic approaches as well as disadvantages of methods that employ sparse regularization. We are primarily interested in settings where overlap in the effect supports between modes is encouraged (rather than enforced) due to prior knowledge or a need for estimate interpretability. Our proposed method combines lasso, TV, and sparse group lasso penalties to this end.

Our primary motivating example concerns the association of R2* and QS MRI maps with age: in particular, previous

works have shown that both modalities need to be considered simultaneously to delineate the complex association between iron and myelin levels (Hametner et al., 2018). An understanding of iron accumulation in healthy aging can give insights into the association between brain iron levels and neurological disease such as multiple sclerosis (Elkady et al., 2019). In this article, we considered the problem of identifying deep gray matter regions where variation in R2* and QS measures is jointly associated with aging in healthy controls. Although the main focus of this study

was the detection of specific iron changes in deep gray matter, our work may be extended to delineate iron and myelin changes in myelin-rich white matter tissue.

The neuroimaging literature demonstrates a need for methods that can simultaneously accommodate multiple modalities and yield interpretable results in joint region detection problems, such as in our central R2*-QS-age example. Methods employing sparse regularization suffer due to spatial correlation in imaging data and tend to select a non-compact set of disjoint voxels. Image-based (e.g., TV) penalties help address this, but do not consider relationships between modalities. In practice, this may be problematic when it is expected that relevant brain regions are similar across modalities.

Our introduction of a sparse group lasso penalty is novel in this neuroimaging context and addresses the above gap in the current literature. We have shown that our SGL model is capable of obtaining estimates that are compact, interpretable, and stable, both within and between imaging modalities. Our work here is a first step in more general approaches to the multimodal analysis of age- or disease-related alterations in brain structure and function.

4.3. Limitations and Future Developments

In this article, we only considered a continuous response due to its wide applicability and our primary motivation to study age-related variation in DGM iron levels. However, our approach is readily applied to other tasks by modification of the objective function in (3). For example, Zhang et al. (2018) considers a discriminative anatomy detection problem based on logistic regression using a single MRI modality. This could be developed to include multiple modalities and a sparse group penalty for joint discriminative anatomy detection. To focus on our approach to multiple imaging modalities in this work, we also did not consider subject-level scalar predictors z_i (e.g., demographic or clinical variables). By changing the $y - X_1\beta_1 - X_2\beta_2$ portion of the objective function in (3) to $y - Z\gamma - X_1\beta_1 - X_2\beta_2$, where Z has row vectors $z_i^\top$, scalar covariates can be included with only straightforward changes to the ADMM algorithm. In both cases, our work here is a general framework that supports other standard tools employed in regression analysis.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://github.com/pietrosa/qsr2-sim.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Health Research Ethics Board, University of Alberta. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MP performed the analyses presented in this article, designed the simulation study, and was responsible for the text, tables, and figures in this work. LZ authored the optimization code used for model-fitting, provided some initial analytic neuroimaging results (not presented), and authored a previous version of this manuscript. PS was heavily involved in subject recruitment, acquired all neuroimaging data used in this study, and provided continued support with neuroimaging data handling and all pre-processing tasks. AE contributed to writing the introduction of the manuscript. AHW designed the image acquisition protocol, funded data collection, and led initial project discussions. LK provided methodological development, modified the general ADMM algorithm for this problem, and provided senior project supervision. DC provided methodological support and senior supervision, motivated this work, and contributed to previous versions of the manuscript. All authors contributed to manuscript editing.

FUNDING

ACKNOWLEDGMENTS

REFERENCES

(2015). *World Health Organization: World Report on Ageing and Health.*

Acosta-Cabronero, J., Betts, M. J., Cardenas-Blanco, A., Yang, S., and Nestor, P. J. (2016). In vivo MRI mapping of brain iron deposition across the adult lifespan. *J. Neurosci.* 36, 364–374. doi: 10.1523/JNEUROSCI.1907-15.2016

ANTS (2011). Available online at: http://www.picsl.upenn.edu/ants/

Ashburner, J., and Friston, K. J. (2000). Voxel-based morphometry – the methods. *NeuroImage* 11, 805–821. doi: 10.1006/nimg.2000.0582

Baldassarre, L., Mourao-Miranda, J., and Pontil, M. (2012). "Structured sparsity models for brain decoding from fMRI data," in *International Workshop on Pattern Recognition in Neuroimaging* (London, UK), 5–9. doi: 10.1109/PRNI.2012.31

Batmanghelich, N. K., Taskar, B., and Davatzikos, C. (2011). Generative-discriminative basis learning for medical imaging. *IEEE Trans. Med. Imaging* 31, 51–69. doi: 10.1109/TMI.2011.2162961

Bergsland, N., Schweser, F., Dwyer, M. G., Weinstock-Guttman, B., Benedict, R. H. B., and Zivadinov, R. (2018). Thalamic white matter in multiple sclerosis: a combined diffusion-tensor imaging and quantitative susceptibility mapping study. *Hum. Brain Mapp.* 39, 4007–4017. doi: 10.1002/hbm.24227

Betts, M. J., Acosta-Cabronero, J., Cardenas-Blanco, A., Nestor, P. J., and Düzel, E. (2016). High-resolution characterisation of the aging brain using simultaneous quantitative susceptibility mapping (QSM) and R2* measurements at 7T. *NeuroImage* 138, 43–68. doi: 10.1016/j.neuroimage.2016.05.024

Bilgic, B., Pfefferbaum, A., Rohlfing, T., Sullivan, E. V., and Adalsteinsson, E. (2012). MRI estimates of brain iron concentration in normal aging using quantitative susceptibility mapping. *NeuroImage* 59, 2625–2635. doi: 10.1016/j.neuroimage.2011.08.077

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* 3, 1–122. doi: 10.1561/2200000016

Cherubini, A., Peran, P., Caltagirone, C., Sabatini, U., and Spalletta, G. (2009). Aging of subcortical nuclei: microstructural, mineralization and atrophy modifications measured in vivo using MRI. *NeuroImage* 48, 29–36. doi: 10.1016/j.neuroimage.2009.06.035

Davatzikos, C. (2004). Why voxel-based morphometric analysis should be used with great caution when characterizing group differences. *NeuroImage* 23, 17–20. doi: 10.1016/j.neuroimage.2004.05.010

Dohmatob, E., Gramfort, A., Thirion, B., and Varoquaux, G. (2014). "Benchmarking solvers for TV-l_1 least-squares and logistic regression in brain imaging," in *International Workshop on Pattern Recognition in Neuroimaging* (Tübingen), 1–4. doi: 10.1109/PRNI.2014.6858516

Dubois, M., Hadj-Selem, F., Lofstedt, T., Perrot, M., Fischer, C., Frouin, V., et al. (2014). "Predictive support recovery with TV-elastic net penalty and logistic regression: an application to structural MRI," in *International Workshop on Pattern Recognition in Neuroimaging* (Tübingen), 1–4. doi: 10.1109/PRNI.2014.6858517

Eickenberg, M., Dohmatob, E., Thirion, B., and Varoquaux, G. (2015). "Grouping total variation and sparsity: statistical learning with segmenting penalties," in *MICCAI* (Munich), 685–693. doi: 10.1007/978-3-319-24553-9_84

Elkady, A., Cobzas, D., Sun, H., Blevins, G., and Wilman, A. H. (2017). Progressive iron accumulation across multiple sclerosis phenotypes revealed by sparse classification of deep gray matter. *J. Magn. Reson. Imaging* 46, 1464–1473. doi: 10.1002/jmri.25682

Elkady, A. M., Cobzas, D., Sun, H., Seres, P., Blevins, G., and Wilman, A. H. (2019). Five year iron changes in relapsing-remitting multiple sclerosis deep gray matter compared to healthy controls. *Multiple Scleros. Relat. Disord.* 33, 107–115. doi: 10.1016/j.msard.2019.05.028

Fan, Y., Shen, D., Gur, R. C., Gur, R. E., and Davatzikos, C. (2007). COMPARE: Classification of morphological patterns using adaptive regional elements. *IEEE Trans. Med. Imaging* 26, 93–105. doi: 10.1109/TMI.2006.886812

Fan, Y., and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *J. R. Stat. Soc. Ser. B* 75, 531–552. doi: 10.1111/rssb.12001

Gramfort, A., Thirion, B., and Varoquaux, G. (2013). "Identifying predictive regions from fMRI with TV-l_1 prior," in *International Workshop on Pattern Recognition in Neuroimaging* (Washington, DC), 17–20. doi: 10.1109/PRNI.2013.14

Grosenick, L., Klingenberg, B., Katovich, K., Brian, K., and Jonathan, E. T. (2013). Interpretable whole-brain prediction analysis with GraphNet. *NeuroImage* 72, 304–321. doi: 10.1016/j.neuroimage.2012.12.062

Haacke, E. M., Miao, Y., Liu, M., Habib, C. A., Katkuri, Y., Liu, T., et al. (2010). Correlation of putative iron content as represented by changes in R2* and phase with age in deep gray matter of healthy adults. *J. Magn. Reson. Imaging* 32, 561–576. doi: 10.1002/jmri.22293

Hallgren, B., and Sourander, P. (1958). The effect of age on the non-haemin iron in the human brain. *J. Neurochem.* 3, 41–51. doi: 10.1111/j.1471-4159.1958.tb12607.x

Hametner, S., Endmayr, V., Deistung, A., Palmrich, P., Prihoda, M., Haimburger, E., et al. (2018). The influence of brain iron and myelin on magnetic susceptibility and effective transverse relaxation: a biochemical and histological validation study. *NeuroImage* 179, 117–133. doi: 10.1016/j.neuroimage.2018.06.007

He, Q., Kong, L., Wang, Y., Wang, S., Chan, T. A., and Holland, E. (2016). Regularized quantile regression under heterogeneous sparsity with application to quantitative genetic traits. *Comput. Stat. Data Anal.* 95, 222–239. doi: 10.1016/j.csda.2015.10.007

Huo, X., and Ni, X. (2007). When do stepwise algorithms meet subset selection criteria? *Ann. Stat.* 35, 870–887. doi: 10.1214/009053606000001334

Kandel, B., Avants, B., Gee, J., and Wolk, D. (2013). "Predicting cognitive data from medical images using sparse linear regression," in *Information Processing in Medical Imaging* (Asilomar, CA), 86–97. doi: 10.1007/978-3-642-38868-2_8

Krishnapuram, B., Carin, L., Figueiredo, M. A., and Hartemink, A. J. (2005). Sparse multinomial logistic regression: fast algorithms and generalization bounds. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 957–968. doi: 10.1109/TPAMI.2005.127

Langkammer, C., Krebs, N., Goessler, W., Scheurer, E., Yen, K., Fazekas, F., et al. (2012). Susceptibility induced gray-white matter MRI contrast in the human brain. *NeuroImage* 59, 1413–1419. doi: 10.1016/j.neuroimage.2011.08.045

Mangeat, G., Govindarajan, S., Mainero, C., and Cohen-Adad, J. (2015). Multivariate combination of magnetization transfer, t2* and b0 orientation to study the myelo-architecture of the *in vivo* human cortex. *NeuroImage* 119, 89–102. doi: 10.1016/j.neuroimage.2015.06.033

Michel, V., Gramfort, A., and Varoquaux, G. (2011). Total variation regularization for fMRI-based prediction of behaviour. *IEEE Trans. Med. Imaging* 30, 1328–1340. doi: 10.1109/TMI.2011.2113378

Ng, B., Vahdat, A., Hamarneh, G., and Abugharbieh, R. (2010). "Generalized sparse classifiers for decoding cognitive states in fMRI," in *Machine Learning in Medical Imaging 2010* (Beijing), 108–115. doi: 10.1007/978-3-642-15948-0_14

Peters, A. (2002). The effects of normal aging on myelin and nerve fibers: a review. *J. Neurocytol.* 31, 581–593. doi: 10.1023/A:1025731309829

Rohr, K. (1997). On 3D differential operators for detecting point landmarks. *Image Vis. Comput.* 15, 219–233. doi: 10.1016/S0262-8856(96)01127-4

Ryali, S., Supekar, K., Abrams, D. A., and Menon, V. (2010). Sparse logistic regression for whole-brain classification of fMRI data. *NeuroImage* 51, 752–764. doi: 10.1016/j.neuroimage.2010.02.040

Sabuncu, M. R., and Van Leemput, K. (2012). The relevance voxel machine (RVoxM): a self-tuning Bayesian model for informative image-based prediction. *IEEE Trans. Med. Imaging* 31, 2290–2306. doi: 10.1109/TMI.2012.2216543

Simon, N., Friedman, G., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *J. Comput. Graph. Stat.* 22, 231–245. doi: 10.1080/10618600.2012.681250

Su, T., Wang, Y., Liu, Y., Branton, W. G., Asahchop, E., Power, C., et al. (2020). Sparse multicategory generalized distance weighted discrimination in ultra-high dimensions. *Entropy* 22:1257. doi: 10.3390/e22111257

Sun, H., and Wilman, A. H. (2013). Background field removal using spherical mean value filtering and Tikhonov regularization. *Magn. Reson. Med.* 71, 1151–1157. doi: 10.1002/mrm.24765

Taege, Y., Hagemeier, J., Bergsland, N., Dwyer, M. G., Weinstock-Guttman, B., Zivadinov, R., et al. (2019). Assessment of mesoscopic properties of deep gray matter iron through a model-based simultaneous analysis of magnetic susceptibility and r2* – a pilot study in patients with multiple sclerosis and normal controls. *NeuroImage* 186, 308–320. doi: 10.1016/j.neuroimage.2018.11.011

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B* 67, 91–108. doi: 10.1111/j.1467-9868.2005.00490.x

Walsh, A. J., Blevins, G., Lebel, R. M., Seres, P., Emery, D. J., and Wilman, A. H. (2014). Longitudinal MR imaging of iron in multiple sclerosis: an imaging marker of disease. *Radiology* 270, 186–196. doi: 10.1148/radiol.13130474

Wang, Y., and Liu, T. (2015). Quantitative susceptibility mapping (QSM): decoding MRI data for a tissue magnetic biomarker. *Magn. Reson. Med.* 73, 82–101. doi: 10.1002/mrm.25358

Watanabe, T., Kessler, D., Scott, C., Angstadt, M., and Sripada, C. (2014). Disease prediction based on functional connectomes using a scalable and spatially-informed support vector machine. *NeuroImage* 96, 183–202. doi: 10.1016/j.neuroimage.2014.03.067

Wyss-Coray, T. (2016). Ageing, neurodegeneration and brain rejuvenation. *Nature* 539, 180–186. doi: 10.1038/nature20411

Yu, D., Zhang, L., Mizera, I., Jiang, B., and Kong, L. (2019). Sparse wavelet estimation in quantile regression with multiple functional predictors. *Comput. Stat. Data Anal.* 136, 12–29. doi: 10.1016/j.csda.2018.12.002

Zhang, L., Cobzas, D., Wilman, A. H., and Kong, L. (2017). "An unbiased penalty for sparse classification with application to neuroimaging data," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, eds M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne (Cham: Springer International Publishing), 55–63. doi: 10.1007/978-3-319-66179-7_7

Zhang, L., Cobzas, D., Wilman, A. H., and Kong, L. (2018). Significant anatomy detection through sparse classification: a comparative study. *IEEE Trans. Med. Imaging* 37, 128–137. doi: 10.1109/TMI.2017.2735239

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x