The Canadian Journal of Statistics
Vol. 47, No. 4, 2019, Pages 729–747
La revue canadienne de statistique

Doubly sparse regression incorporating graphical structure among predictors

Matthew STEPHENSON[1], R. Ayesha ALI[1*], and Gerarda A. DARLINGTON[1] for the Alzheimer's Disease Neuroimaging Initiative[†]

[1] *Department of Mathematics and Statistics, University of Guelph, 50 Stone Road East, Guelph, Ontario, Canada N1G 2W1*

Key words and phrases: Prediction; undirected graphs; variable selection.

MSC 2010: Primary 62J07; secondary 62M20; 62P10.

Abstract: Recent research has demonstrated that information learned from building a graphical model on the predictor set of a regularized linear regression model can be leveraged to improve prediction of a continuous outcome. In this article, we present a new model that encourages sparsity at both the level of the regression coefficients and the level of individual contributions in a decomposed representation. This model provides parameter estimates with a finite sample error bound and exhibits robustness to errors in the input graph structure. Through a simulation study and the analysis of two real data sets, we demonstrate that our model provides a predictive benefit when compared to previously proposed models. Furthermore, it is a highly flexible model that provides a unified framework for the fitting of many commonly used regularized regression models. *The Canadian Journal of Statistics* 47: 729–747; 2019 © 2019 Statistical Society of Canada

Résumé: La recherche récente montre que l'information apprise en construisant un modèle graphique sur l'ensemble des prédicteurs d'un modèle de régression régularisé peut être exploitée pour améliorer la prévision d'une réponse continue. Les auteurs présentent un nouveau modèle encourageant les solutions clairsemées à la fois au niveau des coefficients de régression et des contributions individuelles dans une représentation décomposée. Ils fournissent des estimateurs des paramètres et leurs bornes d'erreurs pour les échantillons finis et montrent leur robustesse aux erreurs dans la structure graphique fournie en entrée. À l'aide d'une étude de simulation et de l'analyse de deux jeux de données réelles, les auteurs démontrent que leur modèle comporte des bénéfices prédictifs en comparaison des modèles proposés précédemment. Ils expliquent également comment la grande flexibilité de leur modèle offre un cadre unifié pour l'ajustement de plusieurs modèles de régression régularisés utilisés couramment. *La revue canadienne de statistique* 47: 729–747; 2019 © 2019 Société statistique du Canada

Additional Supporting Information may be found in the online version of this article at the publisher's website.

[*] *Author to whom correspondence may be addressed.*
E-mail: aali@uoguelph.ca

[†] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

1. INTRODUCTION

A common objective in statistical learning is the prediction of an outcome $\mathbf{Y}$ from a set of input features $\mathbf{X}$. For example, how can volume measurements of various brain regions be used in the prediction of cognitive impairment? Linear regression has proven to be a simple yet effective technique that often outperforms complicated non-linear approaches, particularly in cases of low signal-to-noise ratios, small training sets, or sparse data (Hastie, Tibshirani & Friedman, 2001). However, as data sets continue to increase in size, the use of conventional statistical methods is challenged when the number of input features (explanatory variables), p, exceeds the sample size, n. Therefore, in addition to accurate prediction, a secondary (and often inherent) objective of many predictive models involves variable selection. In the context of a brain region volume example, we may be interested in not only the prediction of cognitive impairment, but also the identification of regions of the brain that are associated with cognitive function.

Regularized regression adds a penalty term $\mathcal{R}(\boldsymbol{\beta})$, weighted by a tuning parameter $\lambda > 0$, to an ordinary least squares model and has been proposed as a method to improve predictive accuracy. These models provide parameter estimates even when the matrix of predictors, $\mathbf{X}$, is not of full rank and are valid in the presence of multicollinearity or when $p > n$ (Hastie, Tibshirani & Friedman, 2001). Recently, Yu & Liu (2016) introduced structural sparsity among the regression coefficients by assuming a graphical structure of the predictors and then exploiting that structure during minimization.

Here, we introduce the *doubly sparse regression incorporating graphical* structure among predictors model (DSRIG). This regularized regression model provides parameter estimates with a finite sample error bound and is more robust to the presence of false-positive edges in the predictor graph. The graphical structure of the predictors is first modelled independently from the outcome and this predictor structure is then leveraged to improve prediction in the construction of $\mathcal{R}(\boldsymbol{\beta})$. DSRIG allows for sparsity not only at the level of contributions of the regression coefficients to the outcome but also at the level of individual contributions of the predictors to $\boldsymbol{\beta}$ based on a decomposed representation of the regression coefficients. Our new model has improved prediction, is highly flexible and provides a unified framework for fitting many available regularized linear regression models.

Section 2 motivates and introduces our new model, DSRIG, and outlines the estimation procedure. Section 3 theoretically evaluates parameter estimates and derives the finite sample error bound. Section 4 examines the empirical properties of the estimates through a comprehensive simulation study, while Section 5 presents the analysis of two real world data sets. Lastly, Section 6 discusses our results, provides conclusions and highlights areas for future research.

2. METHODS AND MOTIVATION

Let $\mathcal{J}$ represent a set of nodes defined by p predictors, labelled $1, \ldots, p$, each measured on n individuals such that the predictor matrix $\mathbf{X}$ can be arranged in an $n \times p$ matrix of observations. Assume that the p-dimensional observations on each individual are independently and identically distributed multivariate normal, $\mathbf{X}_k \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), k = 1, \ldots, n$, with the $p \times p$ precision matrix $\boldsymbol{\Omega}$ taking elements ω_{ij} for $i, j = 1, \ldots, p$; that is, $\boldsymbol{\Omega} = \left[\omega_{ij}\right]_{p \times p} = \boldsymbol{\Sigma}^{-1}$. Then, an undirected graph representing the joint distribution of the variables in $\mathbf{X}$ will have edges between any nodes (i, j), $i \neq j$, wherever $\omega_{ij} \neq 0$.

Let $\boldsymbol{\Sigma}_{xy} = (c_1, c_2, \ldots, c_p)^T$ be the cross covariance vector between predictors in $\mathbf{X}$ and an $n \times 1$ response vector $\mathbf{Y}$. Then, under the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \text{MVN}(0, \sigma^2 \mathbf{I}_n)$, the $p \times 1$ regression coefficient vector $\boldsymbol{\beta}$ may be decomposed as:

$$\boldsymbol{\beta} = \boldsymbol{\Omega}\boldsymbol{\Sigma}_{xy}. \tag{1}$$

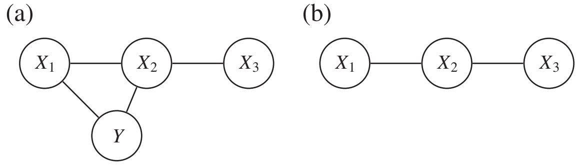

FIGURE 1: (a) Undirected graph representing the joint distribution over $\{Y, X_1, X_2, X_3\}$ and (b) predictor graph over $\{X_1, X_2, X_3\}$ obtained after marginalizing Y out of the joint distribution in (a).

For $i, j = 1, \ldots, p$, let $V_j^{(i)} = c_i \omega_{ji}$ represent the contribution of the ith predictor to the jth regression coefficient and let $\mathbf{V}^{(i)} = [V_1^{(i)}, \ldots, V_p^{(i)}]^T$ be a $p \times 1$ column vector representing the contributions of the ith predictor to $\boldsymbol{\beta}$. Then, Equation (1) can be re-expressed as $\boldsymbol{\beta} = \sum_{i=1}^{p} \mathbf{V}^{(i)}$. Further let the neighbourhood of node i, $\mathcal{N}_i$, be defined as the union of node i and the set of all its neighbours, that is, the set of all other nodes $j = 1, \ldots, p$, $j \neq i$, such that $\omega_{ij} \neq 0$. Consequently, the support of $\mathbf{V}^{(i)}$ will simply be given by the neighbourhood $\mathcal{N}_i$ since $\omega_{ji} = 0$ whenever there is no edge between nodes (i, j) in the predictor graph. In other words, node i only contributes to the estimation of regression coefficients associated with its neighbours and so learning the support of $\mathbf{V}^{(i)}$ is analogous to learning the structure of the predictor graph. Yu & Liu (2016) exploit this relation in the *sparse regression incorporating graphical* structure among predictors (SRIG) model.

2.1. Sparse Regression Incorporating Graphical

The SRIG model of Yu & Liu (2016) assumes that the predictor graph structure is known. As such, the support of $\mathbf{V}^{(i)}$ is also known and estimation of the SRIG coefficients proceeds by solving

$$\underset{\boldsymbol{\beta}, \mathbf{V}^{(i)}: i=1,\ldots,p}{\arg\min} \left\{ \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{i=1}^{p} \tau_i \|\mathbf{V}^{(i)}\|_2 \right\}, \tag{2}$$

where $\boldsymbol{\beta} = \sum_{i=1}^{p} \mathbf{V}^{(i)}$, $\lambda \geq 0$ is a tuning parameter and τ_i is a weight for the ith predictor. Often, we set τ_i as $\tau_i = \frac{\sqrt{d_i}}{|\hat{\beta}_i|^\gamma}$ for $n \geq p$ and $\tau_i = \frac{\sqrt{d_i}}{|\hat{c}_i|^\gamma}$ for $n < p$, where $\gamma > 0$ is a tuning parameter and $d_i = |\mathcal{N}_i|$ is the degree of node i, representing the size of its neighbourhood. Although the tuning parameter γ may be trained via cross-validation we assume $\gamma = 1$ from here forward for simplicity. Note that the SRIG penalty in Equation (2) has some similarity to the group-LASSO penalty of Yuan & Lin (2006) but acts only on the vectors of contributions $\left\{\mathbf{V}^{(i)}, i = 1, \ldots, p\right\}$. That is, the shrinkage and selection of the SRIG model is done at the level of the $\mathbf{V}^{(i)}$'s rather than directly on the regression coefficients in $\boldsymbol{\beta}$. However, there may be situations where encouraging sparsity within the $\mathbf{V}^{(i)}$'s is desirable.

Consider the augmented graph consisting of the union of the predictors in X and the outcome Y. Let $\mathcal{J}_0 = \{i : \beta_i \neq 0, i = 1, \ldots, p\}$ represent the set of nodes that are true (direct) predictors of $\mathbf{Y}$, that is, the neighbours of Y in the augmented graph (e.g., X_1 and X_2 of Figure 1a). Equivalently, the set of true (direct) predictors can be thought of as the set of nodes associated with non-zero regression coefficients in the underlying data-generating mechanism. Yu & Liu (2016) demonstrated that the finite sample bounds derived for SRIG prediction and estimation errors require that if any node $i \in \mathcal{J}_0$, then $\mathcal{N}_i \subseteq \mathcal{J}_0$ (Assumption A2 of Yu & Liu (2016)).

In other words, all neighbours of true predictors are also true predictors. This implies that any two nodes connected by a continuous path in the predictor graph are assumed to have associated regression coefficients that are zero or non-zero together. We too will make use of this assumption (see Lemma 3.2).

Suppose edge (i, j) is in the predictor graph such that node $i \in \mathcal{J}_0$ but node $j \notin \mathcal{J}_0$. For example, consider the edge between X_2 and X_3 of Figure 1b. SRIG would set $V_1^{(3)} = V_3^{(1)} = 0$ since $\omega_{13} = \omega_{31} = 0$. However, the vector of contributions $\mathbf{V}^{(3)}$ associated with X_3 may not be shrunk to zero, thereby increasing the bias in $\boldsymbol{\beta}$. Furthermore, SRIG does not provide a mechanism to shrink $V_3^{(2)}$ to 0 without (erroneously) shrinking the entire vector $\mathbf{V}^{(2)}$ to $\mathbf{0}$. Therefore, SRIG may produce a non-zero β_3, which would be equivalent to adding an edge between X_3 and Y in the augmented graph of Figure 1a. Through simulation, Yu & Liu (2016) found that the performance of SRIG decreased relative to the LASSO as the assumption that for any node $i \in \mathcal{J}_0$, then $\mathcal{N}_i \subseteq \mathcal{J}_0$ became increasingly challenged. By allowing small $V_j^{(i)}$ associated with an edge (i, j) for which $i \in \mathcal{J}_0$ but $j \notin \mathcal{J}_0$ to shrink to zero, the effects of violations to this assumption can be minimized.

From another perspective, in practice, the predictor graph structure is typically unknown and estimated from data. Accordingly, (i, j) may be a false positive edge in the predictor graph with a small $V_j^{(i)}$ relative to $V_k^{(i)}$ for some $k \in N_i, k \neq j$. Shrinking such $V_j^{(i)}$ to zero would make the estimation procedure more robust to the mis-specification of the predictor graph and help mitigate bias in the final estimate of $\boldsymbol{\beta}$. In short, a model that further encourages sparsity within the $\mathbf{V}^{(i)}$ by shrinking $V_j^{(i)}$'s that take on small values to zero while retaining other larger components can improve variable selection and predictive performance.

2.2. Doubly Sparse Regression Incorporating Graphical

In order to induce sparsity both between and within the $\mathbf{V}^{(i)}, i = 1, \ldots, p$, the new DSRIG model adds an ℓ_1 penalty to the SRIG objective function with regression coefficients found as the solution to

$$\underset{\boldsymbol{\beta}, \mathbf{V}^{(i)}: i=1,\ldots,p}{\arg\min} \left(\frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \left\{ \sum_{i=1}^{p} \left[\tau_i \|\mathbf{V}^{(i)}\|_2 + \xi \|\mathbf{V}^{(i)}\|_1 \right] \right\} \right), \tag{3}$$

where $\boldsymbol{\beta} = \sum_{i=1}^{p} \mathbf{V}^{(i)}$, for $\lambda \geq 0$ and $\xi \geq 0$. The tuning parameter ξ balances the contributions of the ℓ_1 and ℓ_2 components of the penalty term. The penalty proposed in Equation (3) is similar in nature to the sparse-group-LASSO (Simon et al., 2013) but performs shrinkage on the $V^{(i)}$'s rather than directly on $\boldsymbol{\beta}$. The ℓ_1 component induces sparsity within $\mathbf{V}^{(i)}$ by shrinking individual contributions $V_j^{(i)}$ to zero, while the ℓ_2 penalty functions as in SRIG and shrinks entire vectors of contributions, $\mathbf{V}^{(i)}$, to the zero vector.

As with SRIG, DSRIG first requires estimation of the predictor graph. To further mitigate any potential bias induced by graph mis-specification, we recommend that the predictor graph structure estimation be incorporated in a cross-validation scheme in which the graph structure is learned on only the training set. The optimization problem in Equation (3) can then be solved across a grid of (λ, ξ) using the training data with the optimal tuning parameters being chosen by an independent validation set. Since the learned predictor graphs from the training sets would exhibit variation, the estimated model parameters would implicitly reflect some of this uncertainty in the graph structure.

2.3. Estimation by Proximal Gradient Descent

The DSRIG regression parameters can be estimated using proximal gradient descent. First reformulate the DSRIG optimization problem in Equation (3) such that the predictor matrix $\mathbf{X}$

is represented in an expanded form (Obozinski, Jacob & Vert, 2011; Rao et al., 2013; Rao et al., 2014; Yu & Liu, 2016). Let $\mathbf{X}^{(j)}$ represent the jth column of the original predictor matrix $\mathbf{X}$. Further let $\tilde{\mathbf{X}} = \left[\mathbf{X}^{(j)}\right], j \in \mathcal{N}_i,\ i = 1, \dots, p$, be an augmented predictor matrix with columns consisting of replicates of the columns of the original predictor matrix, $\mathbf{X}^{(j)}$, for each neighbourhood in which predictor j belongs; therefore, $\tilde{\mathbf{X}}$ will be of dimension $n \times \sum d_i$. When $\mathcal{N}_i = \{i\}$ for $i = 1, \dots, p$ (no edges in the predictor graph), then $\tilde{\mathbf{X}} = \mathbf{X}$. If $\tilde{\mathbf{V}}$ is defined as the $\sum d_i \times 1$ column vector formed by concatenating all the non-zero elements of the $\mathbf{V}^{(i)}, i = 1, \dots, p$, then the expected value of $\mathbf{Y}$ is $\mathbb{E}\left[\mathbf{Y}|\mathbf{X}\right] = \mathbf{X}\boldsymbol{\beta} = \tilde{\mathbf{X}}\tilde{\mathbf{V}}$, and Equation (3) in its re-expressed expanded form is given by

$$\begin{aligned}&\underset{\boldsymbol{\beta},\mathbf{V}^{(i)}:i=1,\dots,p}{\arg\min}\left(\frac{1}{2n}\|\mathbf{Y}-\tilde{\mathbf{X}}\tilde{\mathbf{V}}\|_2^2 + \lambda\left\{\sum_{i=1}^{p}\left[\tau_i\|\tilde{\mathbf{V}}^{(i)}\|_2\right]+\xi\|\tilde{\mathbf{V}}\|_1\right\}\right)\\ &= \underset{\boldsymbol{\beta},\mathbf{V}^{(i)}:i=1,\dots,p}{\arg\min}\left(\mathcal{L}(\boldsymbol{\beta})+\lambda\,\mathcal{R}(\boldsymbol{\beta})\right),\end{aligned}\tag{4}$$

where $\boldsymbol{\beta} = \sum_{i=1}^{p} \mathbf{V}^{(i)}$, $\tilde{\mathbf{V}}^{(i)}$ represents the non-zero components in $\mathbf{V}^{(i)}$, $\mathcal{L}(\boldsymbol{\beta})$ is a smooth loss function, and $\mathcal{R}(\boldsymbol{\beta})$ is a non-smooth penalty function.

Note that $\tilde{\mathbf{V}}^{(i)}$ contains the contributions from the neighbourhood of i, $\mathcal{N}_i$. In fact, the predictor graph can be viewed as a set of neighbourhoods that may overlap. Consequently, the optimization in Equation (4) is analogous to the multi-task set-up in Rao et al. (2013) in which the neighbourhoods represent a set of overlapping groups and the optimization induces sparsity both among and within groups (neighbourhoods). Accordingly, we can utilize proximal point methods that alternate between taking a gradient step in the negative direction of $\mathcal{L}(\boldsymbol{\beta})$ followed by subsequent application of the proximal operator of $\mathcal{R}(\boldsymbol{\beta})$ to do the optimization (Rao et al., 2014).

The gradient of $\mathcal{L}(\boldsymbol{\beta})$ with respect to $\tilde{\mathbf{V}}$ is $\nabla\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{n}\tilde{\mathbf{X}}'\left(\tilde{\mathbf{X}}\tilde{\mathbf{V}} - \mathbf{Y}\right)$, yielding the parameter estimates for iteration $r+1$ as $\tilde{\mathbf{V}}^{\{r+1\}} = \text{Prox}\left[\tilde{\mathbf{V}}^{\{r\}} - t\nabla\mathcal{L}\left(\tilde{\mathbf{V}}^{\{r\}}\right)\right]$, for some step size t, where Prox$[\cdot]$ is the proximal operator of $\mathcal{R}(\boldsymbol{\beta})$. Let $\tilde{\mathbf{V}}_{\nabla}$ be the intermediary found after taking a gradient step, but before applying the proximal operator. This proximal operator can be divided into two steps: (i) a soft thresholding of the individual elements $\tilde{\mathbf{V}}_j, j = 1, \dots, \sum_{i=1}^{p} d_i$, to address the ℓ_1 component

$$\tilde{V}_j^* = \begin{cases}\text{sign}(\tilde{\mathbf{V}}_{\nabla j}) - \lambda\xi & \text{for } |\tilde{\mathbf{V}}_{\nabla j}| > \lambda\xi\\ 0 & \text{otherwise}\end{cases}; \tag{5}$$

and (ii) a group soft thresholding of each of the $\tilde{\mathbf{V}}^{(i)}, i = 1, \dots, p$, to address the ℓ_2 component

$$\left(\tilde{\mathbf{V}}^{(i)}\right)^{\{r+1\}} = \begin{cases}\frac{\tilde{\mathbf{V}}^{(i)*}}{\|\tilde{\mathbf{V}}^{(i)*}\|_2}\left(\|\tilde{\mathbf{V}}^{(i)*}\|_2 - \lambda\tau_i\right) & \text{for } \|\tilde{\mathbf{V}}^{(i)*}\|_2 > \lambda\tau_i\\ 0 & \text{otherwise}\end{cases}. \tag{6}$$

From the proximal operator steps in Equations (5) and (6), it is easy to see the hierarchical structure of our proximal operator where shrinkage and selection is first performed element-wise for the $\tilde{V}_j$ and then performed group-wise for the $\tilde{\mathbf{V}}^{(i)}$. In applications, we actually implemented the *fast iterative shrinkage-thresholding algorithm* (FISTA) with backtracking to choose the step size t (Beck & Teboulle, 2009) in order to accelerate convergence.

3. ESTIMATOR PROPERTIES

Since DSRIG is designed to accommodate the $p > n$ problem, standard consistency of the DSRIG estimator cannot be obtained. Instead, we derive an explicit finite sample error bound that holds with a high probability. Finding such a bound is not straightforward. The derivation of the consistency rate in Yu & Liu (2016) is not easily extended to the case where $\mathcal{R}(\boldsymbol{\beta})$ includes an ℓ_1-penalty term. While we may borrow the ideas from Rao et al. (2013), our bound makes different assumptions on the tuning parameters and is derived for the undirected predictor graph. In the derivation of a finite sample error bound, Rao et al. (2013) required equal group weights, $\tau_i, i = 1, \dots, p$; and an equal contribution of the ℓ_1 and ℓ_2 penalties to $\mathcal{R}(\boldsymbol{\beta})$ with $\tau_i = \xi = 1$ for $i = 1, \dots, p$. Here we derive a bound without requiring these assumptions.

Negahban et al. (2012) provided a framework for deriving finite sample error bounds for convex optimization problems like that in Equation (3) when $\lambda > 0$ and $\mathcal{R}(\boldsymbol{\beta})$ is a norm. Fortunately, several of the results needed here are corollaries of the results proven by Negahban et al. (2012); for self-containment purposes, we provide these results without proof, as needed. This framework hinges on two key properties of the objective function to be minimized: (i) the regularizer, or penalty function, $\mathcal{R}(\boldsymbol{\beta})$ is *decomposable*; and (ii) the loss function $\mathcal{L}(\boldsymbol{\beta})$ meets a *restricted strong convexity* condition.

Section 3.1 provides definitions of decomposability and restricted strong convexity and proves that the DSRIG optimization problem in Equation (3) exhibits these properties. Section 3.2 then derives the finite sample error bound. The assumptions required for the derivation of this bound can be found in the Appendix, while all proofs and intermediary results are contained in Section B of the Supplementary Information.

3.1. Properties of $\mathcal{R}(\boldsymbol{\beta})$ and $\mathcal{L}(\boldsymbol{\beta})$

Assume Assumption (A2) in the Appendix, that for any node $i \in \mathcal{J}_0$ we have $\mathcal{N}_i \subseteq \mathcal{J}_0$. Let $\mathcal{M}$ and $\overline{\mathcal{M}}$ be two subspaces such that $\mathcal{M} \subseteq \overline{\mathcal{M}} \in \mathbb{R}^p$ and let $\overline{\mathcal{M}}^{\perp}$ be the orthogonal complement of $\overline{\mathcal{M}}$. In what follows, we choose $\mathcal{M}$ to be the model subspace, which reflects the constraints imposed by DSRIG. Define the cardinality of the set of true predictors $\mathcal{J}_0$ as $|\mathcal{J}_0| = s$ such that $s \ll p$ and assume $\boldsymbol{\beta}$ is exactly sparse, that is, $\boldsymbol{\beta} \in \mathcal{M}$ where $\mathcal{M}$ can now be further defined as the s-dimensional model subspace spanned by the coordinates indexed by $\mathcal{J}_0$. Then $\overline{\mathcal{M}}^{\perp} = \mathcal{M}^{\perp}$ may be defined as the subspace spanned by the remaining $p - s$ coordinates indexed in $\mathcal{J}_0^c$. Similarly, define $A_0 = \{i : \mathbf{V}^{(i)} \neq \mathbf{0}, i = 1, \dots, p\}$, with cardinality $|A_0| = a$, where $a \ll p$, to be the set of active $\mathbf{V}^{(i)}$ in our decomposition of $\boldsymbol{\beta}$. If d^{max} is defined to be the maximum degree across all nodes, then the vectors $\mathbf{V}^{(i)}$ themselves are sparse with at most d^{max} non-zero elements, as stated in Assumption (A3) in the Appendix. Further note that $\mathcal{J}_0$ must lie within the union of the supports across all active $\mathbf{V}^{(i)}$, that is, $\mathcal{J}_0 = \bigcup_{i \in A_0} \text{supp}(\mathbf{V}^{(i)})$. These assumptions will be used in many of our subsequent results.

Definition 3.1 (Negahban et al., 2012). *Given the pair of subspaces $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$, a norm-based regularizer $\mathcal{R}(\cdot)$ is decomposable if $\mathcal{R}(\boldsymbol{\beta} + \boldsymbol{\beta}^*) = \mathcal{R}(\boldsymbol{\beta}) + \mathcal{R}(\boldsymbol{\beta}^*)$ for all $\boldsymbol{\beta} \in \mathcal{M}$ and $\boldsymbol{\beta}^* \in \overline{\mathcal{M}}^{\perp}$.*

Lemma 3.2. *Assume Assumptions (A1)–(A3) in the Appendix. Then $\mathcal{R}(\boldsymbol{\beta})$ is a norm and decomposable with respect to the subspace pair $\left(\mathcal{M}, \mathcal{M}^{\perp}\right)$.*

Similar to previous work on decomposed regression coefficients (Obozinski, Jacob & Vert, 2011; Rao et al., 2013; Rao et al., 2014; Yu & Liu, 2016), we require that our decomposition is *optimal* in the sense that there is no other decomposition for which the associated penalty $\mathcal{R}(\boldsymbol{\beta})$ is smaller. See Definition S.2 in Section A of the Supplementary Information. Optimality of the decomposition is needed to show that $\mathcal{R}(\boldsymbol{\beta})$ is a decomposable norm. Rao et al. (2013) noted that if $\mathcal{R}(\boldsymbol{\beta})$ is convex and coercive, which is also the case here, an optimal decomposition exists for any $\boldsymbol{\beta}$.

We now concentrate on properties of the loss function $\mathcal{L}(\boldsymbol{\beta})$ and specify the RSC condition needed to establish a finite sample bound. These conditions ensure that there is sufficient curvature in $\mathcal{L}(\boldsymbol{\beta})$ around its optimum to allow for parameter estimation. In particular, consider $\boldsymbol{\delta}\mathcal{L}(\boldsymbol{\Delta}, \boldsymbol{\beta})$, the error term in a first-order Taylor series expansion of $\mathcal{L}(\boldsymbol{\beta})$ in some direction $\boldsymbol{\Delta}$. Since $p > n$, it suffices to show that $\boldsymbol{\delta}\mathcal{L}(\boldsymbol{\Delta}, \boldsymbol{\beta})$ is lower bounded by $\kappa_{\mathcal{L}}\|\boldsymbol{\Delta}\|^2$, for some $\kappa_{\mathcal{L}} > 0$ for all $\boldsymbol{\Delta}$ in a restricted direction. The estimation error, $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, is the appropriate direction of interest here. Lemma 3.3 shows that $\hat{\boldsymbol{\Delta}}$ falls in a cone set for regularizers with dual norms that bound λ. This cone set is the appropriate space in which strong convexity is needed. Definition 3.4 states the restricted convexity condition.

Lemma 3.3. *Suppose $\mathcal{L}(\cdot)$ is a convex and differentiable loss function and consider any optimal solution $\hat{\boldsymbol{\beta}}$ to the optimization problem in Equation (3) with a strictly positive regularization parameter satisfying $\lambda \geq 2\mathcal{R}^*(\nabla\mathcal{L}(\boldsymbol{\beta}))$, where $\mathcal{R}^*(\cdot)$ is the dual norm of $\mathcal{R}(\cdot)$ and $\nabla\mathcal{L}(\boldsymbol{\beta})$ is the gradient of the loss function. Assume Assumptions (A1)–(A3) in the Appendix and let $\Pi_{\mathcal{M}}(\cdot)$ represent the projection onto the subspace $\mathcal{M}$. Then, the error, $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, belongs to the set*

$$\mathbb{C}(\mathcal{M}, \mathcal{M}^{\perp}, \boldsymbol{\beta}) := \{\boldsymbol{\Delta} \in \mathbb{R}^p | \mathcal{R}(\Pi_{\mathcal{M}^{\perp}}\boldsymbol{\Delta}) \leq 3\mathcal{R}[\Pi_{\mathcal{M}}(\boldsymbol{\Delta})]\}. \tag{7}$$

Definition 3.4. *The loss function $\mathcal{L}(\boldsymbol{\beta})$ satisfies an RSC condition with curvature parameter $\kappa_{\mathcal{L}}$ if it is convex, differentiable and*

$$\boldsymbol{\delta}\mathcal{L}(\boldsymbol{\Delta}, \boldsymbol{\beta}) := \mathcal{L}(\boldsymbol{\beta} + \boldsymbol{\Delta}) - \mathcal{L}(\boldsymbol{\beta}) - \langle \nabla\mathcal{L}(\boldsymbol{\beta}), \boldsymbol{\Delta} \rangle \geq \kappa_{\mathcal{L}}\|\boldsymbol{\Delta}\|_2^2,$$

for any $\boldsymbol{\Delta} \in \mathbb{C}(\mathcal{M}, \mathcal{M}^{\perp}, \boldsymbol{\beta})$, where $\langle \cdot, \cdot \rangle$ represents the inner product of two vectors and $\mathbb{C}(\cdot)$ is as defined in Equation (7).

Assumption (A5) in the Appendix simply states that the loss function meets this RSC condition. Note that Negahban et al. (2012) provided an RSC condition that involves a tolerance parameter, but since $\boldsymbol{\beta} \in \mathcal{M}$, this tolerance parameter is equal to zero and the corresponding term in their lower bound can be ignored.

3.2. Finite Sample Error Upper Bound

Negahban et al. (2012) provided a general error bound (see Theorem S.1 in Section B of the Supplementary Information) for regularized models whenever Lemma 3.2 and Definition 3.4 hold. In what follows, we tailor this bound to the specific optimization problem in Equation (3).

Our RSC condition is needed for a regularizer $\mathcal{R}(\cdot)$ that is not too large relative to the error norm. The subspace compatibility constant formalizes this notion by explicitly relating the error norm and the regularizer (see Definition S.3 in Section A of the Supplementary Information). Establishing a concrete error upper bound requires: (i) bounding the subspace compatibility constant $\psi(\cdot)$, as in Lemma 3.5; and (ii) bounding the dual norm $\mathcal{R}^*(\nabla\mathcal{L}(\boldsymbol{\beta}))$, as in Lemma 3.6, which in turn provides a tighter bound on λ.

Lemma 3.5. *The subspace compatibility constant associated with the optimization in Equation (3) is bounded by:*

$$\psi(\mathcal{M}) \leq \left(\tau^{max} + \xi\sqrt{d^{max}}\right)\sqrt{a}.$$

Lemma 3.6. *Assume Assumptions (A4) and (A6) in the Appendix. Then*

$$\mathcal{R}^*\left(\nabla\mathcal{L}(\boldsymbol{\beta})\right)^2 \leq \frac{\sigma^2\sigma^{max}\left(\log(p)+d^{max}\right)}{4\left(\tau^{min}\right)^2 n},$$

with probability greater than or equal to $1-c_1\exp(-c_2\sqrt{n})$ *for some* $c_1, c_2 > 0$.

Finally, we present our main result.

Theorem 3.7. *Assume Assumptions (A1)–(A6) in the Appendix for the optimization problem in Equation (3) and define* σ_i^* *to be the maximum singular value of* $X_{\mathcal{N}_i}^T X_{\mathcal{N}_i}$ *and* $\sigma^{*max} = \max_{i=1,\ldots,p}(\sigma_i^*)$. *Then for* $\lambda^2 \geq \frac{\sigma^2\sigma^{max}(\log(p)+d^{max})}{\left(\tau^{min}\right)^2 n}$, *any optimal solution,* $\hat{\boldsymbol{\beta}}$, *will satisfy*

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2^2 \leq \frac{9}{\left(\tau^{min}\right)^2}\frac{\sigma^2\sigma^{*max}\left(\tau^{max}+\xi\sqrt{d^{max}}\right)^2 a\left(\log(p)+d^{max}\right)}{n\kappa_{\mathcal{L}}}, \tag{8}$$

with probability greater than or equal to $1-c_1\exp(-c_2\sqrt{n})$ *for some* $c_1, c_2 > 0$.

The bound presented in Equation (8) can be found by substituting our bound on $\psi(\cdot)$ derived in Lemma 3.5 into Theorem S.1. If we assume that $\lambda \geq 2\mathcal{R}^*(\nabla\mathcal{L}(\boldsymbol{\beta}))$, then from Lemma 3.6, $\lambda^2 \geq \frac{\sigma^2\sigma^{max}(\log(p)+d^{max})}{\left(\tau^{min}\right)^2 n}$.

4. SIMULATION STUDY

Through a simulation study we investigate the empirical properties of the DSRIG estimates. Not only does DSRIG have a predictive benefit and decreased bias when compared to SRIG, but results demonstrate that DSRIG also exhibits robustness to both errors in graph estimation and certain violations of assumptions.

4.1. Simulation Study Design

Our simulation study considered two predictor graph structures: (i) random (sparse); and (ii) scale-free. A scale-free structure comprises relatively few hub-nodes of high degree that connect the rest of the lesser connected nodes in the graph (Figure 2), as is commonly seen in biological network models (Jeong et al., 2000). For each predictor graph structure, we generated the precision matrix ($\boldsymbol{\Omega}$) for 30 parent graphs, each with $p = 100$ nodes as follows. Random parent graphs were generated following Yu & Liu (2016) whereby $\boldsymbol{\Omega} = \mathbf{B} + \delta\mathbf{I}_p$. The diagonal entries of $\mathbf{B}$ were initialized to zero while the off diagonal entries of $\mathbf{B}$ took on a value of 0.5 with probability 0.05 and 0 with probability 0.95; therefore the probability of an edge between any pair of nodes (i, j) was 0.05. The value of δ was chosen such that the condition number of $\boldsymbol{\Omega}$ was equal to p and $\boldsymbol{\Omega}$ was subsequently standardized to have unit diagonals. Scale-free parent graphs were generated by first generating a data set with 1,000,000 observations from a scale-free graph using the `huge.generator` function of the `huge` package (Zhao et al., 2015) for R (R Core Team, 2016). The empirical precision matrix, $\boldsymbol{\Omega}$, of these data was then taken to be the true parent precision matrix. From each parent graph, 100 independent $n \times p$ predictor sets, $\mathbf{X}$, were generated (Figure 2) using the `rmvnorm` function of the `mvtnorm` package (Genz et al., 2017; Genz & Bretz, 2009) for R (R Core Team, 2016).

Nodes were sorted according to their respective degrees. Then, for the random graph scenario, $c_i = 4$ for those nodes i that corresponded to the nodes with the four largest degrees. For the

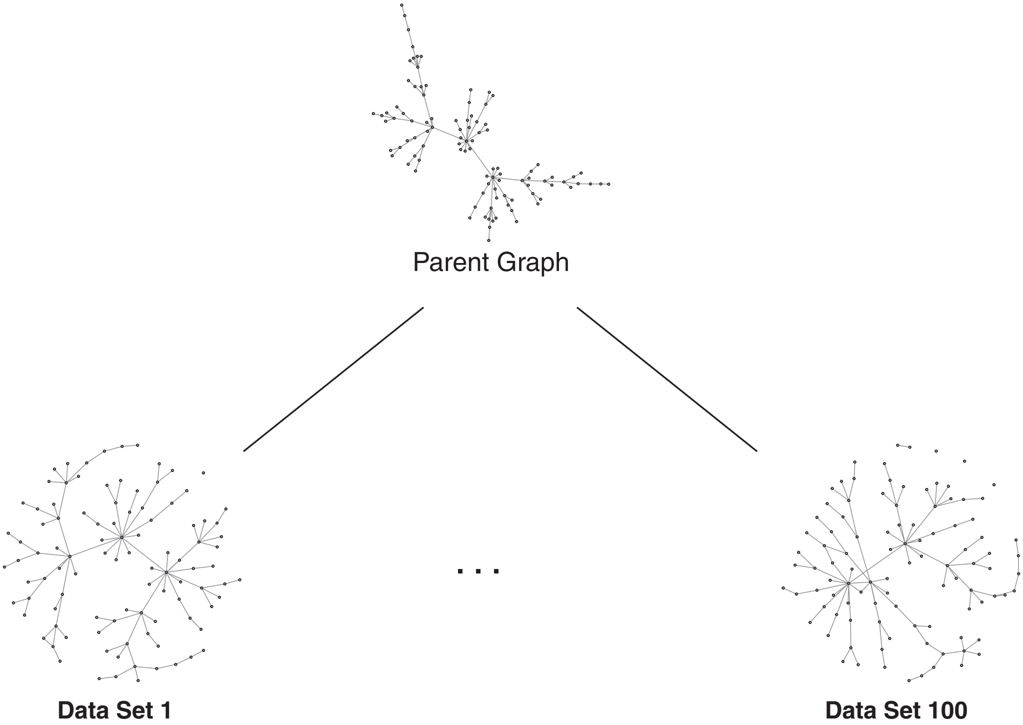

FIGURE 2: Estimated predictor graphs from independent data sets generated from the same scale-free parent graph.

scale-free graph, instead $c_i = 4$ for those nodes i that corresponded to nodes with the second through fifth largest degrees. The value of c_i for all other nodes was assigned to be 0. We considered two scenarios for generating the true regression parameter $\boldsymbol{\beta}$.

In Scenario 1, we wanted to ensure that, for the most part, neighbours of covariates correlated with $\mathbf{Y}$ were also correlated with $\mathbf{Y}$. We simply set $\boldsymbol{\beta} = \boldsymbol{\Omega}\boldsymbol{\Sigma}_{xy}$. While this calculation does not ensure Assumption (A2) will be met, the number of edges that violate this assumption will be minimized. To guarantee Assumption (A2) is violated, in Scenario 2 we took half of the original non-zero β's and switched them with β's that were originally calculated as zero. Covariate data were generated for $n = 480$ and $n = 560$ subjects, for a total of $30 \times 100 = 3,000$ independent data sets for each predictor graph structure and sample size combination. The response variable, $\mathbf{Y}$, was then calculated from the classical linear model, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \text{MVN}(\mathbf{0}, \sigma^2\mathbf{I}_n)$, with $\sigma = 5$.

Each data set was partitioned into a training set to learn the model, a validation set to choose the optimal value of the tuning parameter(s) and a test set to compare the predictive accuracy among the candidate models. For $n = 480$, the data were partitioned into train/validation/test set sample sizes of 40/40/400, while for $n = 560$, the sample sizes were 80/80/400. For each data set the LASSO (Tibshirani, 1996), SRIG (Yu & Liu, 2016) and DSRIG models were fit and compared. Furthermore, the SRIG and DSRIG models were fit under the best case scenario, which used the true graphical structure (SRIG-True, DSRIG-True) and the more realistic scenario in which the graphical structure was estimated from only the training data (SRIG-Est, DSRIG-Est).

For each data set, the predictor graph for the SRIG and DSRIG models was estimated using the `huge` and `huge.select` functions of the `huge` package (Zhao et al., 2015) in R (R Core

Team, 2016), which are based on the Meinshausen-Bühlmann (MB) (Meinshausen & Bühlmann, 2006) method with the *stability approach to regularization selection* (STARS) criterion (Liu, Roeder & Wasserman, 2010). The optimal value of the tuning parameter λ was chosen from a grid of 100 equally spaced values on a logarithmic scale. The maximum, λ_{max}, was chosen such that it was the minimum value for which $\hat{\boldsymbol{\beta}} = \mathbf{0}$ when $\xi = 0$, while the minimum was set at $\lambda_{min} = 0.01 \times \lambda_{max}$. Similarly, the optimal ξ was chosen out of a grid of 100 equally spaced values, but between 0 and ξ_{max}, where ξ_{max} was chosen to be just large enough such that the upper bound was never selected. Lastly, τ_i was calculated using the sample covariance with $\gamma = 1$.

For each model the final estimated regression parameter vector, $\hat{\boldsymbol{\beta}}$, was chosen to be that which minimized the prediction error of the validation set (i.e., minimized the Euclidean distance between the observed $\mathbf{Y}$ and predicted values $\hat{\mathbf{Y}}$). The ℓ_2-distance between the estimated and true regression parameter vectors ($\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$) was used to compare the bias between the candidate models. The relative prediction error (RPE), calculated as RPE $= \frac{1}{\sigma^2 N_{test}} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T \left(\mathbf{X}_{test}^T \mathbf{X}_{test}\right) \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ was used to compare predictive accuracy, where $\mathbf{X}_{test}$ is the predictor matrix for the test set and N_{test} is the size of the test set. For the case where the predictor graphs were estimated, we also recorded the proportion of times for which DSRIG was preferred, SRIG was preferred, or for which the models were equivalent (i.e., DSRIG parameter $\xi = 0$). The absolute median difference of both ℓ_2-distance and RPE between the two models was also recorded. The proximal gradient methods to estimate the SRIG and DSRIG models were implemented in MATLAB version 2016b (MATLAB, 2016) using code adapted from Cox (2014), while the LASSO was fit using the MATLAB function in the Statistics Toolbox. MATLAB code for the fitting of the DSRIG model, along with a sample data set and results, are contained in the Supplementary Information.

4.2. Simulation Study Results

Table 1 records the ℓ_2-distance between the estimated and true regression parameters as well as the RPE for each of the candidate models. As expected, both DSRIG and SRIG fit under the true graphical structure have decreased bias and RPE when compared to the corresponding models fit under an estimated predictor graph. However, DSRIG typically offers improved performance over SRIG both in terms of predictive accuracy and bias, particularly in the estimated predictor graph setting. DSRIG outperformed the LASSO in Scenario 1 while remaining competitive to the LASSO in Scenario 2. However, while SRIG outperformed the LASSO in Scenario 1, it suffered from decreased predictive accuracy and increased bias compared to both DSRIG and the LASSO in Scenario 2.

Table 2 records the proportion of runs for which each method (DSRIG, SRIG) was preferred, as well as the absolute median difference between the two models when the predictor graphs were estimated. DSRIG was preferred in a majority of runs and, when chosen, had a larger performance benefit (larger absolute median difference) than those runs for which SRIG was preferred. A greater discrepancy between the two models can be seen in Scenario 2, both in the proportion of runs preferred and in the absolute median difference. Figure 3 plots the difference (SRIG−DSRIG) for both the ℓ_2-distance and RPE across all runs for the random graph scenario under Scenario 2 with sample sizes of 40/40/400. Plots with similar characteristics were obtained for all other combinations of the study parameters. Points above the horizontal zero line correspond to the models for which DSRIG outperformed SRIG; points below the line correspond to the models for which DSRIG was outperformed by SRIG. Again, it can be seen that not only does DSRIG outperform SRIG more frequently, but DSRIG also provides a greater performance benefit.

TABLE 1: The ℓ_2-distance and relative prediction error of model fits under two different cross-validation splits for training/validation/test for random and scale-free graphs.

	Scenario 1				Scenario 2			
	40/40/400		80/80/400		40/40/400		80/80/400	
Method	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$	RPE	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$	RPE	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$	RPE	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$	RPE
Random graph								
LASSO	10.654	2.882	8.659	2.197	8.873	3.550	6.567	1.750
SRIG-True	9.237	2.446	5.969	1.236	9.068	3.707	6.904	1.994
DSRIG-True	9.087	2.365	6.013	1.225	8.324	3.091	6.567	1.777
SRIG-Est	9.669	2.600	6.663	1.484	9.148	3.783	7.086	2.090
DSRIG-Est	9.440	2.486	6.619	1.417	8.454	3.191	6.698	1.840
Scale-free graph								
LASSO	13.225	3.279	8.574	1.628	9.742	2.953	6.577	1.356
SRIG-True	11.864	3.009	6.393	1.080	10.396	3.428	6.855	1.504
DSRIG-True	11.734	2.847	6.463	1.074	9.679	2.953	6.629	1.383
SRIG-Est	12.955	3.348	7.706	1.488	10.500	3.476	6.955	1.535
DSRIG-Est	12.567	3.084	7.760	1.424	9.789	2.988	6.739	1.414

TABLE 2: Proportion of runs where each method had the lowest ℓ_2-distance or relative prediction error when the predictor graph is estimated from the training data. The median difference between SRIG and DSRIG for the ℓ_2-distance and RPE is shown in brackets.

	Scenario 1				Scenario 2			
	40/40/400		80/80/400		40/40/400		80/80/400	
Method	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$	RPE	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$	RPE	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$	RPE	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$	RPE
Random graph								
DSRIG	0.513	0.535	0.347	0.482	0.797	0.793	0.753	0.778
	(0.351)	(0.174)	(0.245)	(0.116)	(0.764)	(0.624)	(0.457)	(0.266)
Equivalent	0.302	0.300	0.361	0.361	0.145	0.145	0.163	0.163
SRIG	0.186	0.165	0.292	0.157	0.058	0.062	0.084	0.058
	(0.158)	(0.093)	(0.212)	(0.055)	(0.123)	(0.098)	(0.104)	(0.064)
Scale-free graph								
DSRIG	0.513	0.613	0.245	0.482	0.719	0.745	0.572	0.660
	(0.673)	(0.357)	(0.219)	(0.103)	(0.834)	(0.512)	(0.325)	(0.135)
Equivalent	0.263	0.262	0.376	0.376	0.185	0.185	0.244	0.244
SRIG	0.224	0.125	0.379	0.142	0.096	0.0700	0.184	0.096
	(0.274)	(0.095)	(0.235)	(0.034)	(0.210)	(0.076)	(0.118)	(0.030)

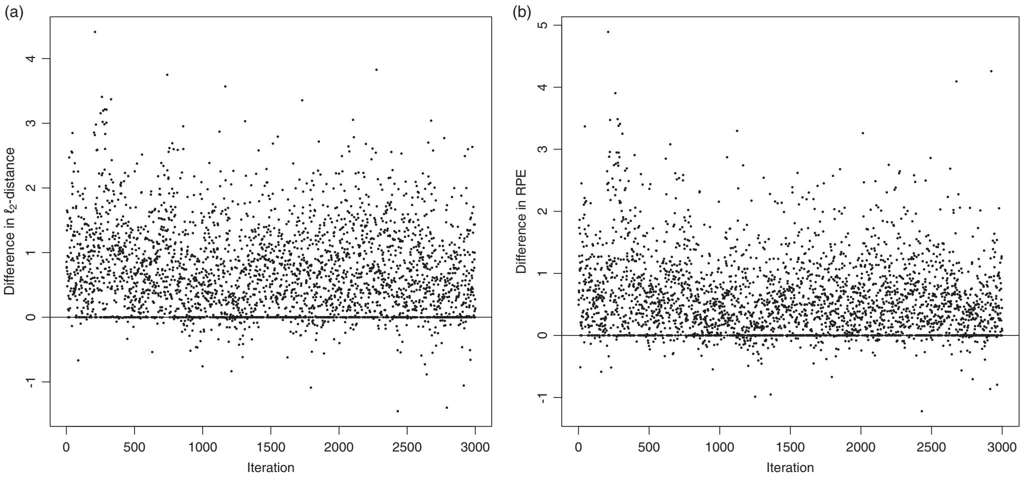

FIGURE 3: Difference (SRIG−DSRIG) for 100 simulations from each of 30 random parent graphs under Scenario 2 with sample sizes of 40/40/400 for (a) ℓ_2-distance and (b) RPE. Points above the zero line indicate when DSRIG performed better.

5. APPLICATION

We demonstrate DSRIG on two real data sets and compare performance with the LASSO and SRIG. The description and results for the second data set, which includes predictors that violate the multivariate normality assumption, can be found in Section C of the Supplementary Information. All data were initially scaled such that each column of the predictor matrix ($\mathbf{X}$) and the outcome vector ($\mathbf{Y}$) had a mean of 0 and standard deviation of 1. The data were then split into 10 roughly equal segments: 8 segments on which to train the model, 1 segment to validate the model and choose the optimal value of the tuning parameter(s) and 1 segment to test the model and compare prediction accuracy across various candidate models. We fit and compared the results for all 90 possible permutations of training/validation/test sets.

As in the simulation study, the predictor graph for the SRIG and DSRIG procedures was learned on the training data using the `huge` and `huge.select` functions of the `huge` package (Zhao et al., 2015) in R (R Core Team, 2016) using the MB method (Meinshausen & Bühlmann, 2006) and based on optimization of the STARS criterion (Liu, Roeder & Wasserman, 2010). The mean square prediction error for the model chosen by the validation set applied to the independent test set was calculated as $MSPE = \frac{1}{N_{Test}} \|\mathbf{Y} - \hat{\mathbf{Y}}\|_2^2$. The identities of covariates with associated non-zero regression coefficient estimates ($abs(\hat{\beta}_i) > 0.01$) were also tracked.

5.1. Alzheimer's Disease Data

Data used in this analysis were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). In this analysis, patients collected under all three ADNI protocols (ADNI1, ADNI2 and ADNIGO) with cross-sectional MRI were included. The objective of the analysis was two-fold: (i) to predict cognitive function from measured volumes of brain regions; and (ii) to identify which regions of the brain best contribute to this prediction. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

Cortical reconstruction and volumetric segmentation was performed with the FreeSurfer 5.1 image analysis suite, which is documented and freely available for download online (http://surfer.nmr.mgh.harvard.edu/). The processing of the MRI images was conducted by Hartig et al. (2012). The technical details of these procedures are described in prior publications (Dale, Fischl & Sereno, 1999; Dale & Sereno, 1993; Fischl, Sereno & Dale, 1999a,b; Fischl & Dale, 2000, 2001; Fischl et al., 2002, 2004a,b; Han et al., 2006; Jovicich et al., 2006; Ségonne et al., 2004). Briefly, this processing includes motion correction and averaging (Reuter, Rosas & Fischl, 2010) of multiple volumetric T1 weighted images (when more than one is available), removal of non-brain tissue using a hybrid watershed/surface deformation procedure (Ségonne et al., 2004), automated Talairach transformation, segmentation of the subcortical white matter and deep gray matter volumetric structures (including hippocampus, amygdala, caudate, putamen, ventricles) (Fischl et al., 2002, 2004a) intensity normalization (Sled, Zijdenbos & Evans, 1998), tessellation of the gray matter white matter boundary, automated topology correction (Fischl, Liu & Dale, 2001; Ségonne, Pacheco & Fischl, 2007) and surface deformation following intensity gradients to optimally place the gray/white and gray/cerebrospinal fluid borders at the location where the greatest shift in intensity defines the transition to the other tissue class (Dale & Sereno, 1993; Dale, Fischl & Sereno, 1999; Fischl & Dale, 2000). The longitudinal processing capabilities of the Freesurfer software were used in the processing of sequential scans for a single patient (Reuter et al., 2012).

The mini-mental state exam (MMSE) is a tool used by clinicians in the evaluation of cognitive functioning of psychiatric patients (Folstein, Folstein & McHugh, 1975). This exam, scored on a 30-point scale, consists of 11 questions and attempts to isolate and assess cognitive functioning from other mental functions such as mood. Higher MMSE scores correspond to higher functioning cognitive abilities. Scores greater than 27 are typically associated with normal cognitive function, whereas scores of 19−24 (mild), 10−18 (moderate), or ≤ 9 (severe) correspond to varying degrees of cognitive impairment (Kukull et al., 1994; Mungas, 1991). Yu & Liu (2016) applied the SRIG model to predict the MMSE scores of 103 subjects from the ADNI database, based on the volume measurements of 93 brain regions.

We considered 135 volume measurements (mm^3) obtained from the Freesurfer segmented data and restricted analysis to data available at the month 6 visit post-baseline. Of the available 696 subjects, 177 had measurements that did not pass the study's overall quality control and a further 197 subjects did not have complete data. We included the remaining 322 subjects in our analysis, along with their $p = 135$ brain region measurements and MMSE scores (range: 15−30). While these data do not represent a high-dimensional scenario where $p > n$, many of the predictors are highly correlated and the predictor matrix $\mathbf{X}$ is not of full rank. Therefore, ordinary least squares would not be expected to perform well and the problem is well-suited to using regularized regression. Figure 4 contains the predictor graph from a single data segmentation. Note that a continuous path can be traced between all nodes in the graph, which suggests a potential violation of Assumption (A2) if sparsity among the regions that predict cognitive impairment is assumed.

5.2. Results of Data Analysis

Table 3 summarizes model performance in terms of mean MSPE and model complexity (number of non-zero coefficients found in the final model) across all 90 data permutations for the ADNI data set. DSRIG had the smallest mean MSPE followed by SRIG and then the LASSO. LASSO typically chose the most sparse model with the least variability in the number of non-zero predictors, while SRIG resulted in the least sparse regression models and was the most variable in the number of non-zero coefficients selected. Our new DSRIG model fell between the LASSO and SRIG both in terms of sparsity and in the variability of the number of non-zero coefficients.

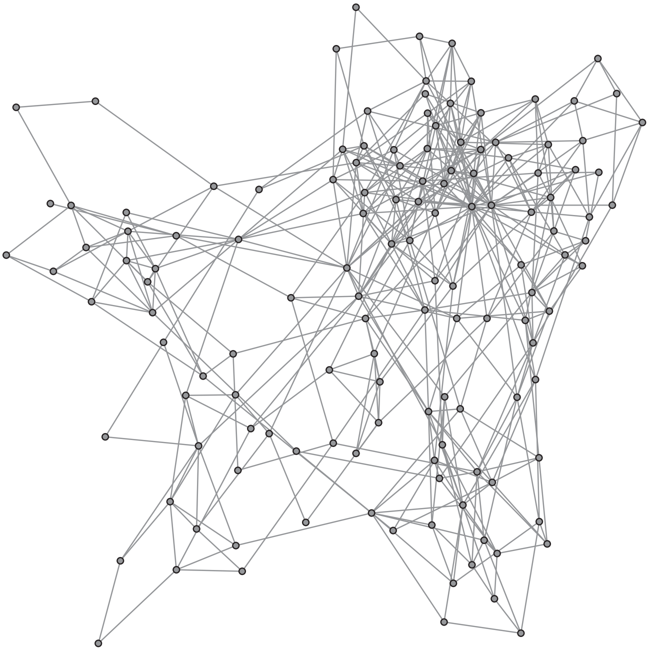

FIGURE 4: Estimated predictor graph for a single training segmentation of the ADNI data.

TABLE 3: Average mean square error in prediction and mean and standard deviation for the number of non-zero regression coefficients for 90 permutations of the ADNI and blood brain barrier data.

	LASSO	SRIG	DSRIG
ADNI			
Mean MSPE	0.722	0.717	0.707
Non-zero coefficients			
Mean number	30.967	40.122	38.600
Standard deviation	20.269	25.947	23.643

Figure 5 shows the differences, SRIG−DSRIG, between the MSPEs across each of the 90 permutations for the ADNI data. Values above the horizontal line correspond to the 27 models for which DSRIG outperformed SRIG; values below the line correspond to the 12 models for which DSRIG was outperformed by SRIG. We can see that not only did DSRIG perform better more often than SRIG, but also had a larger performance improvement (absolute mean difference 0.037 vs. 0.010).

Table 4 records the identity of predictors with non-zero regression coefficient estimates in at least 80 of the final models for the ADNI data. Interestingly, across the LASSO models, there was only one brain region, the right inferior lateral ventricle, that consistently was selected. There was a large overlap between the predictors commonly chosen by both DSRIG and SRIG with both models choosing three right hippocampal subfields, the right temporal pole and third ventricle. This is similar to the analysis performed by Yu & Liu (2016) which also found the right temporal pole and hippocampus to be associated with MMSE score. In our analyses, SRIG

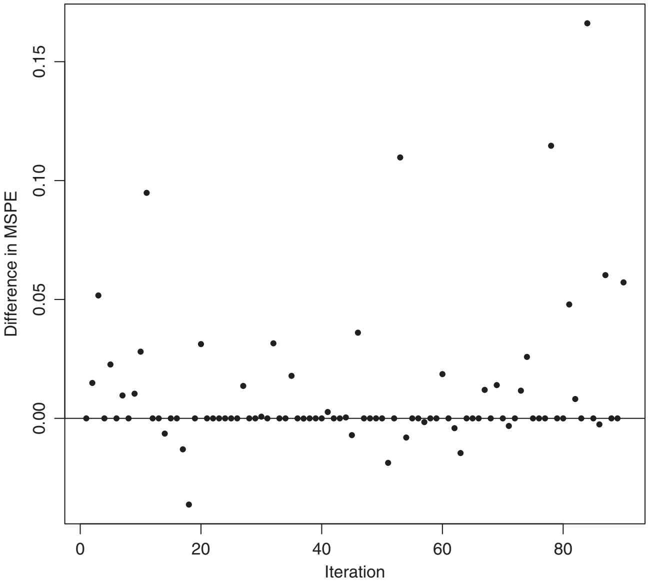

FIGURE 5: Difference in MSPE (SRIG−DSRIG) for 90 permutations of ADNI data. Points above the zero line indicate when DSRIG performed better.

TABLE 4: Predictors found to be non-zero in at least 80 models for the ADNI data.

Predictor[a]	LASSO	SRIG	DSRIG
R Temporal Pole		✓	✓
Third Ventricle		✓	✓
L Hipp		✓	
L Inf Lateral Ventricle		✓	
R Hipp		✓	
R Inf Lateral Ventricle	✓	✓	✓
R Hipp Subfield 1		✓	✓
2		✓	✓
3		✓	✓
4		✓	

[a] Hipp, hippocampus; Inf, inferior; L, left; R, right.

often identified a fourth right hippocampal subfield, as well as the right hippocampus, left hippocampus and left inferior lateral ventricle in at least 80 of the 90 final models fit.

6. DISCUSSION AND FUTURE WORK

We have introduced the DSRIG model, which performs shrinkage and selection on components of a decomposed representation of the regression coefficients. We have also derived a proximal gradient descent algorithm for parameter estimation and identified a finite sample error bound.

Like the SRIG model of Yu & Liu (2016), the predictor graph structure was exploited to improve the performance of regularized regression. Our model improves upon SRIG by encouraging sparsity both within and among the $\mathbf{V}^{(i)}$. This additional level of sparsity makes DSRIG more robust to predictor graph mis-specification and to violations of Assumption (A2), particularly when the predictor graph is unknown and estimated from the data.

The results presented do not enforce the restriction $\max_{i=1,\dots,p}(\tau_i) \leq \xi$ (Assumption (A4)) because we found that prediction error was improved without it. For the ADNI data, we gained a 0.3% improvement through its removal. Recall that Assumptions (A1)–(A6) were required to find a finite sample error bound. Assumption (A2) is used to prove the decomposability of $\mathcal{R}(\boldsymbol{\beta})$ while Assumption (A4) is used to bound the dual norm of $\mathcal{R}(\boldsymbol{\beta})$. When these assumptions are not met, it does not mean DSRIG is invalid, or that a finite error bound does not exist, but rather that we are not able to derive one at this time.

DSRIG is a highly flexible model that provides a unified framework for fitting several regularized regression models. When $\xi = 0$ our DSRIG model is equivalent to SRIG. Whenever all nodes in the graph are singletons and $\tau_i = \tau$ for $i = 1, \dots, p$, we get the LASSO. When the predictor graph consists of a series of complete disconnected subgraphs (where each subgraph represents a group) and $\xi = 0$, the method is equivalent to the group-LASSO. Lastly, when the predictor graph is complete and $\xi = 0$, DSRIG is equivalent to ridge regression.

We implemented DSRIG using an expanded form of the predictor matrix. For dense graphs with many edges, this approach can become computationally intensive. For the ADNI data, using a 2013 Mac Pro with a 6 Core 3.5 GHz Intel Xeon processor running on a single core, the average computation time for a single data split was 6.1 s for the LASSO, 2.1 s for SRIG and 52.8 s for DSRIG. For all three models, we considered 100 possible tuning parameter values for λ. However, DSRIG has an additional tuning parameter ξ which resulted in 100×100 possible combinations of (λ, ξ). Accordingly, part of the additional computation time was due to the additional tuning parameter rather than the expanded predictor set representation. We are working on a more efficient estimation algorithm that better scales to large data sets.

APPENDIX: Assumptions

The following assumptions are needed in Section 3 to establish that the DSRIG regularizer defined in Equation (3) is decomposable and that the loss function meets a restricted strong convexity condition.

(A1) The decomposition of our regression coefficients into the set of vectors $\mathbf{V}^{(i)}, i = 1, \dots, p$, is an optimal decomposition.

(A2) For any node $i \in \mathcal{J}_0$, then $\mathcal{N}_i \subseteq \mathcal{J}_0$.

(A3) The true regression parameter vector $\boldsymbol{\beta}$ is exactly sparse with s non-zero components that can be decomposed into a set of a active vectors $\mathbf{V}^{(i)}$ with at most $d^{max} = \max_{i=1,\dots,p}(d_i)$ non-zero elements.

(A4) $\tau^{max} = \max_{i=1,\dots,p}(\tau_i)$, is upper bounded by ξ and $\tau^{min} = \min_{i=1,\dots,p}(\tau_i)$, is lower bounded by 1.

(A5) The loss function $\mathcal{L}(\boldsymbol{\beta})$ satisfies the restricted strong convexity (RSC) conditions with curvature parameter κ_L (see Definition 3.4).

(A6) The design matrix $\mathbf{X}$ is fixed; the observation errors $\epsilon_k, k = 1, \dots, n$, are additive, independent of $\mathbf{X}$ and $\epsilon_k \overset{i.i.d.}{\sim} \text{Normal}(0, \sigma)$.

Based on these assumptions, we can obtain a finite sample error bound for the ℓ_2 error between any optimal solution, $\hat{\boldsymbol{\beta}}$, of Equation (3) and the true parameter vector $\boldsymbol{\beta}$. Assumption (A4) explicitly states $\tau^{min} \geq 1$; however, $\mathbf{X}$ and $\mathbf{Y}$ are typically scaled to have columnwise standard deviations of 1 and therefore, $|c_i| \leq 1$, $d_i \geq 1$ and for $\tau_i = \frac{\sqrt{d_i}}{|\hat{c}_i|}$, we have $\tau^{min} \geq 1$. Although Assumption (A4) is needed to obtain the theoretical results presented, we found that prediction improved when the restriction $\max(\tau_i) \leq \xi, i = 1, \dots, p$, is not enforced. All analysis results reported here are with this restriction removed.

ACKNOWLEDGEMENTS

This work was funded by the Natural Sciences and Engineering Research Council of Canada. Additionally, data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education and the study is coordinated by the Alzheimer?s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

BIBLIOGRAPHY

Beck, A. & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2, 183–202.

Cox, C. (2014). WholeBrain MVPA, *GitHub Repository*. https://github.com/\crcox/WholeBrain_MVPA.

Dale, A., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *NeuroImage*, 9, 179–194.

Dale, A. M. & Sereno, M. I. (1993). Improved localizadon of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: A linear approach. *Journal of Cognitive Neuroscience*, 5, 162–176.

Fischl, B. & Dale, A. M. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences of the United States of America*, 97, 11050–11055.

Fischl, B., Liu, A., & Dale, A. M. (2001). Automated manifold surgery: Constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE Medical Imaging*, 20, 70–80.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., & Dale, A. M. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33, 341–355.

Fischl, B., Salat, D. H., van der Kouwe, A. J., Makris, N., Ségonne, F., Quinn, B. T., & Dale, A. M. (2004a). Sequence-independent segmentation of magnetic resonance images. *NeuroImage*, 23, S69–S84.

Fischl, B., Sereno, M. I., & Dale, A. (1999a). Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9, 195–207.

Fischl, B., Sereno, M. I., Tootell, R. B., & Dale, A. M. (1999b). High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping*, 8, 272–284.

Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D. H., Busa, E., Seidman, L. J., Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B., & Dale, A. M. (2004b). Automatically parcellating the human cerebral cortex. *Cerebral Cortex*, 14, 11–22.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12, 189–198.

Genz, A. & Bretz, F. (2009). Computation of multivariate normal and *t* PRobabilities. In *Lecture Notes in Statistics*, Springer-Verlag, Heidelberg.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2017). *mvtnorm: Multivariate normal and t distributions*. R package version 1.0-6.

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J., Albert, M., Killiany, R., Maguire, P., Rosas, D., Makris, N., Dale, A., Dickerson, B., & Fischl, B. (2006). Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *NeuroImage*, 32, 180–194.

Hartig, M., Truran-Sacrey, D., Raptentsetsang, S., Schuff, N., & Weiner, M. (2012). *UCSF FreeSurfer Methods*. Alzheimer's Disease Neuroimaging Initiative, San Francisco, CA.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of statistical learning. In *Springer Series in Statistics*, Springer New York Inc., New York, NY.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., & Barabási, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, 407, 651–654.

Jovicich, J., Czanner, S., Greve, D., Haley, E., van der Kouwe, A., Gollub, R., Kennedy, D., Schmitt, F., Brown, G., MacFall, J., Fischl, B., & Dale, A. (2006). Reliability in multi-site structural MRI studies: Effects of gradient non-linearity correction on phantom and human data. *NeuroImage*, 30, 436–443.

Kukull, W. A., Larson, E. B., Teri, L., Bowen, J., McCormick, W., & Pfanschmidt, M. (1994). The mini-mental state examination score and the clinical diagnosis of dementia. *Journal of Clinical Epidemiology*, 47, 1061–1067.

Liu, H., Roeder, K., & Wasserman, L. (2010). Stability approach to regularization selection (StARS) for high dimensional graphical models. *Advances in Neural Information Processing Systems*, Vol. 23, 1432–1440.

MATLAB. (2016). *MATLAB and Statistics Toolbox Version: R2016b*. The MathWorks Inc., Natick, MA.

Meinshausen, N. & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34, 1436–1462.

Mungas, D. (1991). In-office mental status testing: A practical guide. *Geriatrics*, 46, 54–67.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., & Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27, 538–557.

Obozinski, G., Jacob, L., & Vert, J. P. (2011). *Group lasso with overlaps: The latent group lasso approach*, arXiv preprint arXiv:1110.0413.

R Core Team. (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, N., Cox, C., Nowak, R., & Rogers, T. T. (2013). Sparse overlapping sets lasso for multitask learning and its application to fMRI analysis. In *Advances in Neural Information Processing Systems*, Vol. 26, 2202–2210.

Rao, N., Nowak, R., Cox, C., & Rogers, T. (2014). *Classification with sparse overlapping groups*, arXiv preprint arXiv:1402.4512.

Reuter, M., Rosas, H. D., & Fischl, B. (2010). Highly accurate inverse consistent registration: A robust approach. *NeuroImage*, 53, 1181–1196.

Reuter, M., Schmansky, N. J., Rosas, H. D., & Fischl, B. (2012). Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, 61, 1402–1418.

Ségonne, F., Dale, A., Busa, E., Glessner, M., Salat, D., Hahn, H., & Fischl, B. (2004). A hybrid approach to the skull stripping problem in MRI. *NeuroImage*, 22, 1060–1075.

Ségonne, F., Pacheco, J., & Fischl, B. (2007). Geometrically accurate topology-correction of cortical surfaces using nonseparating loops. *IEEE Transactions on Medical Imaging*, 26, 518–529.

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

Sled, J., Zijdenbos, A., & Evans, A. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging*, 17, 87–97.

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

Yu, G. & Liu, Y. (2016). Sparse regression incorporating graphical structure among predictors. *Journal of the American Statistical Association*, 111, 707–720.

Yuan, M. & 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.

Zhao, T., Li, X., Liu, H., Roeder, K., Lafferty, J., & Wasserman, L. (2015). *huge: High-dimensional undirected graph estimation*, R package version 1.2.7.

Received 26 September 2017
Accepted 13 May 2019