Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–14

DOI: 10.1177/09622802211022404
journals.sagepub.com/home/smm
SAGE

A Bayesian group lasso classification for ADNI volumetrics data

Atreyee Majumder[1], Tapabrata Maiti[1] and Subha Datta[2]

Abstract

The primary objective of this paper is to develop a statistically valid classification procedure for analyzing brain image volumetrics data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) in elderly subjects with cognitive impairments. The Bayesian group lasso method thereby proposed for logistic regression efficiently selects an optimal model with the use of a spike and slab type prior. This method selects groups of attributes of a brain subregion encouraged by the group lasso penalty. We conduct simulation studies for high- and low-dimensional scenarios where our method is always able to select the true parameters that are truly predictive among a large number of parameters. The method is then applied on dichotomous response ADNI data which selects predictive atrophied brain regions and classifies Alzheimer's disease patients from healthy controls. Our analysis is able to give an accuracy rate of 80% for classifying Alzheimer's disease. The suggested method selects 29 brain subregions. The medical literature indicates that all these regions are associated with Alzheimer's patients. The Bayesian method of model selection further helps selecting only the subregions that are statistically significant, thus obtaining an optimal model.

Keywords

Alzheimer's Disease Neuroimaging Initiative data classification, Bayesian group lasso, group lasso, logistic regression, spike, slab

1 Introduction

Alzheimer's disease (AD) is the sixth leading cause of death in the United States. It is a form of dementia in which patients suffer loss of memory where they fail to identify people or objects, have difficulty with speech and, in later stages, are unable to perform daily life activities like getting up from the bed or brushing their teeth. Although it is mainly a disease of old age affecting people who are 65 or older, early onset of the disease can occur in 40 or 50 year olds in up to 5% of cases. AD is the most common case of dementia amounting to 60%–80% of all cases. It is a progressive disease where symptoms worsen over time. AD affected patients live an average of 4–20 years after the symptoms become noticeable. Medical scientists are yet to find a cure for Alzheimer's but it is possible to slow down the worsening of dementia and improve lifestyles of both the affected people and their caregivers. According to recent studies (Leifer[1]) early detection of AD is extremely helpful as it can be treated with novel drugs to delay AD progression. Extensive studies are being conducted to find a treatment for AD, delay its onset or curb its advancement. More information and facts about AD can be found at www.alz.org.

Dedicated research is done with neuroimaging techniques for early diagnosis of AD. Alzheimer's Disease Neuroimaging Initiative (ADNI) conducts multi-center case-control study of elderly people that was designed to find more sensitive and accurate methods to diagnose AD at earlier stages. ADNI studies use brain-imaging techniques, such as positron emission tomography (PET) and magnetic resonance imaging (MRI). ADNI

[1]Department of Statistics and Probability, Michigan State University, East Lansing, MI, USA
[2]Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ, USA

Corresponding author:
Atreyee Majumder, 240 Chubb Avenue, Lyndhurst, NJ 07071, USA.
Email: atreyee7@gmail.com

database has data from three phases (ADNI1, ADNIGO, and ADNI 2). MRI data points were obtained from 1737 subjects with baseline diagnosed as Normal, mild cognitive impairment (MCI), and AD. For all subjects at each visit, structural MRI scans were acquired from 1.5 T scanners for ADNI1 subjects and from 3 T scanners for ADNIGO and ADNI2 subjects. MRI protocols were performed across a variety of scanners such as GE, Siemens, or Philips to ensure comparability. MRI volumes were computed using FreeSurfer by UCSF/SF VA Medical Center. ADNI1's 1.5 T data were run with FreeSurfer version 4.3 and ADNIGO and ADNI2's 3 T data were run with FreeSurfer version 5.1.

Historically, studies have shown that AD causes abnormal change to brain region volumes which causes shrinkage in the hippocampal volume or reduction in its thickness or enlargement of internal ventricles. Smith et al.[2] studied structural brain alterations before MCI. They had previously demonstrated that volume loss in bilateral anteromedial temporal lobe is present at baseline in longitudinally followed normal subjects who later developed MCI or AD. Arlt et al.[3] believed that fully automated MRI-based volumetric measurements may serve as a biomarker for the diagnosis in patients with MCI or dementia. They concluded that fully automated MRI-based volumetry allows detection of regional gray matter volume loss that correlates with neuropsychological performance in patients with amnestic MCI or mild AD. Our objective in this paper is to predict dementia in patients based on the volumetric measurements obtained from the MRI ADNI data. There is evidence of brain atrophy with increasing age but the atrophies differ significantly from normal aging to AD patients. We use the differences of brain region atrophies to distinguish subjects with or without AD. The volumetric data have brain parcellated subregions of the entire brain for the left and right hemispheres. Volume, area and thickness measurements of brain subregion is a simple way of detecting atrophied brain regions, thus the motivation of combined use of these measurements. It is believed that all these regions are not associated with dementia, but only a few (Haroutunian et al.,[4] Shivamurthy et al.[5]). Identification of a few brain regions from the large pool of regions makes appropriately a dimension reduction problem.

Numerous methods have been developed for the analysis of ADNI data to identify the brain subregions that are disease related. These methods usually single out a region of interest (ROI) and perform a univariate analysis based on the chosen ROI (Luo and Nichols,[6] Grimmer et al.[7]) Univariate analysis of ROIs neglects the effect of other significant ROIs. These methods aim at analyzing each hypothesized significant ROI and then looking at multiple hypothesis where careful adjustment of multiple comparisons have to be looked at. To include all the ROIs for analysis simultaneously, a regression framework seems plausible such that the model itself selects the most significant regions. However, due to high number of candidate ROIs, the standard regression analysis is not possible. The good thing is, there is a medical belief that only a few ROIs are informative for characterizing AD. Thus, a dimension reduction technique such as penalized regression can be developed. The ADNI MRI database has volume, area, and thickness measurements of various brain regions. Since these measurements are a direct manifestation of brain region atrophies, we should consider them as a single variable with multiple levels. Thus, the number of regression parameters (brain subregions with all levels) may exceed the number of patients being studied. There are a number of competitive penalized regression techniques that have been developed in recent years. Least absolute shrinkage and selection operator (LASSO[8]) is perhaps the most popular technique among all. However, the direct use of LASSO is not appropriate in the presence of multiple levels of a covariate in feature selection models. We employ, instead, a group lasso technique to build a model since it places group penalty on the parameters of a variable (feature) which makes easy selection of the whole set of volumetric measurements for an ROI. We treat different measurements of the same subregion as different levels of a covariate in this regression setup. Thus, it is easy to visualize structured correlation in the matrix of covariates (subregions) establishing the motivation of using a group lasso like method. The acuteness of AD makes its early detection imperative which is why classification of a subject into healthy individuals or AD patients is of immense importance. We have developed logistic regression in Bayesian setup for detection of AD for getting more reliable standard error estimates.

In this article, a Bayesian group lasso type technique has been developed with spike and slab prior following Xu and Ghosh[9] over other types of penalized regression because this approach presents many natural advantages. The biggest advantage is that the Bayesian approach provides reliable estimates of uncertainty which can be used for statistical inference beyond feature selection. A thorough literature review has shown that the Bayesian group lasso with logistic regression model is largely overlooked. This article develops this novel method motivated by the ADNI data. Bayesian group lasso with spike and slab prior deals with feature selection (dimension reduction) in a binary outcome scenario and produces reliable estimates for regression coefficients. Unlike commonly used Bayesian variable selection methods, we propose median thresholding to make insignificant coefficients are exactly zero. Another major contribution of this paper is that we look at the brain image volumetric data at a granular

level. We consider all available brain subregions mapped by FreeSurfer to include effects of all ROIs rather than looking at individual ROIs. An atrophied brain subregion is identified by selecting a group of volumetric measurements of the corresponding selected ROI. Zhang et al.[10] performed classification with MRI data based on 93 manually labeled ROIs. We use data of 116 automatically labeled ROIs (each having four different measurements) by FreeSurfer and treat this as a multivariate problem to perform a dimension reduction analysis. Zhang et al.[10] used a composite of three different modalities of biomarkers. Unlike Zhang et al.,[10] our method provides reliable parameter estimates which can be used to calculate the log of odds or relative risk of AD based on the selected subregions instead of solely classifying subjects. Group lasso encourages selection of all levels of a significant subregion, and spike and slab prior together with median ensure that a large number of subregions which have no impact on the disease are dropped from the model. So, the proposed method selects affected brain subregions automatically from a large pool of brain subregions. Furthermore, among the selected subregions only a few volumetric measurements serve as discriminative features in the model assessed by their statistical significance. So, we are able to narrow down the regions and their corresponding attribute that should be studied by scientists to stop progress of the disease or improve the quality of life of the affected individuals. Finally, we have provided theoretical foundation to our proposed methodology.

The rest of the paper is organized into seven sections. In Section 2 we have reviewed the literature of group lasso and Bayesian group lasso and then in Section 3 we have elaborated on Bayesian group lasso in logistic regression setup. Section 4 shows the posterior consistency of our estimator, i.e. the model selected by the proposed method converges to the true model for sufficiently large n. In Section 5 we have conducted a simulation study to test the performance of the proposed method. Section 6 contains the analysis on the ADNI dataset where we detail out our findings and our concluding remarks are in Section 7.

2 Group lasso

Variable selection is a technique of selecting an optimal model in predictive modeling. In many regression problems, we are interested in selecting feature variables that are important in predicting the response variable. The feature variables can be individual numeric variables, various levels of a categorical variable or a number of basis functions of the original measured variables. Recently proposed methods like the LASSO, SCAD (smoothly clipped absolute deviation[11]), etc. can efficiently perform variable selection by selecting individual feature variables. In case of an ANOVA (Analysis of Variance) type model where there are multiple levels of a feature variable or for an additive model where each component is a linear combination of a number of basis functions, selecting the important variable amounts to selecting all levels of the variable.

A very simple linear regression equation is of the form

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times p}\ \beta_{p\times 1} + \epsilon_{n\times 1} \tag{1}$$

Here, $\mathbf{X}$ is the design matrix whose columns are the feature variables, β is the vector of coefficients, is the error vector where each ϵ_i has a normal distribution with mean zero and variance σ^2 and $\mathbf{Y}$ is the vector of observations.

Each feature variable in equation (1) can be either categorical or continuous. ANOVA is a special case where all the input variables are categorical whereas an additive model is a special case of all continuous input variables. However, the input variables could be a mixture of both numeric and categorical variables in a regression problem given by equation (1).

When we want to work with factor variables with G factors (groups), we can modify our notations in equation (1) as follows

$$\boldsymbol{Y}_{n\times 1} = \sum_{g=1}^{G} \boldsymbol{X}_g \beta_g + \epsilon \tag{2}$$

where $\epsilon_{n\times 1} \sim \boldsymbol{N}_n\left(0, \sigma^2 \boldsymbol{I}_n\right)$, β_g is a coefficient vector of length m_g, and $\mathbf{X}_g$ is an $n \times m_g$ covariate (feature) matrix corresponding to the factor β_g, $g = 1,\ldots,G$. Let p be the total number of predictors, so $p = \sum_{g=1}^{G} m_g$. To eliminate the intercept from equation (2), we center response variables and each input variable so that the observed mean is zero.

The goal is to select important feature variables for accurate prediction. This amounts to selecting as well as estimating the parameter coefficients. Traditional approaches such as subset selection or stepwise procedures can be used for variable selection. Subset selection is impractical when there are a large number of predictors in the model since the number of candidate models grow exponentially as the number of predictors increase. Breiman[12] showed that subset selection methods are not satisfactory with respect to prediction accuracy and stability. Stepwise procedures, on the other hand, often lead to locally optimal solutions rather than globally optimal solutions. However, there are several optimization techniques that help to find exact (global) optimal solution in the field of optimal design of experiments and clinical trials (Hore et al.,[13] Hore et al.[14]). A variable neighborhood search algorithm coupled with a stochastic approach can be used for finding the optimal solution (Hore et al.[15]). These approaches are primarily used in clinical trial settings and have not been explored sufficiently in optimization problems related to other disciplines. The drawback of traditional methods indicates the need for development of sophisticated variable selection methods. Tibshirani[8] proposed the LASSO method. In this approach we minimize

$$\sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3}$$

where λ is the tuning parameter. This penalized approach forces many βs to take zero values. The LASSO is an attractive tool due to simultaneous estimation and variable selection. When the need for selecting a group of levels of a categorical variable or group of basis functions representing a numeric variable arises, these methods fail because they are designed to select individual feature variables and fail to select whole factors. Yuan and Lin[16] proposed group lasso as an alternative to LASSO in terms of factor selection and also exhibit superior model selection performance.

The group lasso penalty is a hybrid of the l_1 and l_2 penalties and encourages selection at a group level. The group lasso estimate, for linear regression, minimizes

$$\left\| Y - \sum_{g=1}^{G} X_g\beta_g \right\|_2^2 + \lambda \sum_{g=1}^{G} \|\beta_g\|_2 \tag{4}$$

where λ is the tuning parameter. Note that equation (3) is a special case of equation (4) when all groups have size 1, i.e. $m_1 = m_2 = \cdots = m_G = 1$.

2.1 Bayesian Group lasso. The limiting distribution of the group lasso estimator is complicated (Knight and Fu,[17] Chatterjee and Lahiri[18]). Thus, this estimator fails to give meaningful standard errors of the estimates which affect the statistical significance of the covariates in the chosen model. To deal with this drawback of frequentist lasso type estimators, Bayesian formulations have been developed. The Bayesian MAP estimators provide reliable standard errors for the estimates.

It is known that the lasso estimator for linear regression is equivalent to the posterior mode with independent Laplace priors on each regression coefficient. Park and Casella[19] developed a fully hierarchical Bayesian setup for the lasso using a scale mixture prior on the regression parameters. This mixture prior results in a Laplace marginal distribution for β. This idea has been further extended by Kyung et al.[20] to build similar fully Bayesian Hierarchical models for group lasso, fused lasso (Tibshirani et al.[21]) and the elastic net (Zou and Hastie[22]). They employ a multivariate m_g-dimensional Laplacian prior over each group of regression coefficients

$$\pi(\beta_g) \propto \exp\left\{ -\frac{\lambda}{\sigma} ||\beta_g||_2 \right\} \tag{5}$$

The classical group lasso is recovered as the MAP solution in log-space with $\frac{\lambda}{\sigma}$ having the role of a fixed Lagrangian multiplier. For a full Bayesian treatment, however, we place hyperpriors on λ and σ which lead to integrations that are analytically impossible to solve.

For finding closed form posterior distributions for all parameters, we extend the hierarchical scale mixture model approach of lasso to grouped predictors. Thus, we express the prior as a scale mixture of multivariate normals over β_g with Gamma hyperpriors over the variance hyperparameter. Specifically, with

$$\beta_g|\tau_g^2, \sigma^2 \overset{ind}{\sim} \boldsymbol{N}_{m_g}\left(0, \tau_g^2\sigma^2\boldsymbol{I}_{m_g}\right), \quad \tau_g^2 \overset{ind}{\sim} Gamma\left(\frac{m_g+1}{2}, \frac{\lambda^2}{2}\right) \tag{6}$$

the marginal distribution of β_g is of the form (5). This Bayesian formulation encourages shrinkage at the group level and provides comparable prediction performance with the group lasso. However, estimation of β_g by its posterior mean or median does not produce exact zero estimates; we need to bring in the concept of sparsity here. Thus, to introduce sparsity at group level, Xu and Ghosh[9] assumed a multivariate zero inflated mixture prior or a spike and slab prior for each β_g.

For variable selection, we want the estimates to produce exact zeroes such that they are dropped from the model. Zero inflated mixture priors are such that the slab part draws values from a known distribution and the spike part is degenerate distribution selecting zero. Xu and Ghosh[9] further showed that median thresholding is better than using posterior mean. The spike and slab prior keeps the scale mixture prior of normals and gamma intact thus providing full conditionals. This approach is thus computationally easy and gives exact zero estimates. Narisetty and He[23] used shrinking and diffusing priors for variable selection in a hierarchical Bayesian setup.

Zero inflated mixture priors, in recent years, have been extensively utilized in Bayesian variable selection setups. George and McCulloch[24] used zero inflated normal mixture priors in the hierarchical formulation for variable selection in a linear regression model. Chen and Dunson[25] used a spike and slab type prior for the random effects variances in a linear setup allowing probabilistic selection of random effects. Also see Zhao and Sarkar,[26] Lykou and Ntzoufras,[27] and Zhang et al.[28] Heavy tailed distributions, such as double exponential, are often used as the slab part. The slab part can be further segmented to a scale mixture of normal and gamma distributions as is done by Xu and Ghosh.[9]

The following hierarchical Bayesian formulation with spike and slab prior for linear regression (2) comparable to a group lasso type estimator is proposed by Xu and Ghosh[9]

$$\boldsymbol{Y}|\boldsymbol{X}, \beta, \sigma^2 \sim \boldsymbol{N}_n\left(\boldsymbol{X}\beta, \sigma^2\boldsymbol{I}_n\right),$$

$$\beta_g|\sigma^2, \tau_g^2 \ , \ \pi_0 \overset{ind}{\sim} (1- \ \pi_0)N_{m_g}\left(0, \tau_g^2 \ \sigma^2 I_{m_g}\right) + \ \pi_0\delta_0\left(\beta_g\right), \ g = 1, \ \ldots, \ G$$

$$\tau_g^2 \overset{ind}{\sim} Gamma\left(\frac{m_g+1}{2}, \frac{\lambda^2}{2}\right), \ g = 1, \ldots, G,$$

$$\sigma^2 \sim Inverse \ Gamma(\alpha, \gamma), \ \sigma^2 > 0,$$

$$\pi_0 \sim Beta(a, b),$$

$$\lambda^{(k)} = \sqrt{\frac{p+G}{\sum_{g=1}^{G} \boldsymbol{E}_{\lambda^{(k-1)}}\left[\tau_g^2|\boldsymbol{Y}\right]}}$$

where $\delta_0(\beta_g)$ denotes a point mass as $\mathbf{0} \in \mathrm{R}^{m_g}, \beta_g = (\beta_{gm_1}, \ldots, \beta_{gm_g})^T$. The posterior expectation of τ_g^2 will be replaced by the sample average of τ_g^2 generated in the Gibbs sampler based on $\lambda^{(k-1)}$. The value of λ should be carefully tuned. A large value of λ will overshrink the estimates while a small value will lead to overfitting. Xu and Ghosh[9] suggested a conjugate Gamma prior can be placed on λ^2. Using an empirical Bayes approach, λ is estimated from data using marginal maximum likelihood. Since marginal maximum likelihood of λ does not have a closed form, a Monte Carlo EM algorithm (Park and Casella[19] and Casella[29]) can be used to estimate λ. The kth EM update for λ is given in the above setup.

3 Bayesian group lasso with logistic regression

So far we have talked about group lasso in a linear regression setup, i.e. when the response variable has a Gaussian error. In many practical problems, we come across response values that cannot be fitted into a linear model. For example, when the outcome is a binary categorical variable, count data or multi-level categorical variable, then the Gaussian error assumption does not hold. In such cases we have to use generalized linear models (GLM) with various link functions. Since, the occurrence of binary outcome is very common in the real world, we will focus on GLM with a logit link.

Since, the outcome is binary we cannot model this data with equation (1) having normal errors. Meier et al.,[30] developed the logistic group lasso in a frequentist setup. Before delving into its Bayesian counterpart, let us summarize the frequentist group lasso in logistic regression setup.

Assume that we have independent and identically distributed observations $(x_i,y_i),i=1,\ldots,n$, of a p-dimensional vector $x_i \in \mathrm{R}^p$ of G predictors and a binary response variable $y_i \in \{0,1\}$, where each group has m_g levels. We can write $x_i = \left(x_{i1}^T, \ldots, \ x_{iG}^T\right)^T$. Linear logistic regression models the conditional probability $p_\beta(x_i) = \mathrm{P}_\beta(\mathbf{Y}=1|x_i)$ by

$$\log\left\{\frac{p_\beta(x_i)}{1-p_\beta(x_i)}\right\} = \eta_\beta(x_i)$$

also known as the logit link with the link function

$$\eta_\beta(x_i) = \sum_{g=1}^{G} x_{ig}^T \beta_g$$

The logistic group lasso estimator, β_{GL}, is given by the minimizer of the convex function

$$S_\lambda(\beta) = -l(\beta) + \lambda \sum_{g=1}^{G} \|\beta\|2$$

where $l(.)$ is the log-likelihood function i.e.

$$l(\beta) = \sum_{i=1}^{n} \left(y_i \eta_\beta(x_i) - \log\left[1+\exp\left\{\eta_\beta(x_i)\right\}\right]\right)$$

The tuning parameter $\lambda \geq 0$ controls the amount of penalization.

Motivated by Xu and Ghosh[9]'s work, we construct a Bayesian formulation for the logistic regression case. Here, our likelihood is Bernoulli probability mass function with a logit link. We abide by using a multivariate zero inflated mixture prior with point mass at zero and the continuous part as double exponential distribution. Since a double exponential prior on β_g can be expressed as a scale mixture of normal and Gamma priors (as in (6)), we use priors very similar to the linear setup

$$\begin{aligned}
&y_i|x_i,\beta \ \sim \ \text{Bernoulli}\left(\frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}\right), i=1,\ldots,n,\\
&\beta_g|\tau_g^2 \ , \ \pi_0 \overset{ind}{\sim}(1-\ \pi_0)\mathrm{N}_{m_g}\left(0,\tau_g^2\ \mathrm{I}_{m_g}\right)+\ \pi_0\delta_0\left(\beta_g\right), \quad g=1,\ \ \ldots,\ G,\\
&\tau_g^2 \overset{ind}{\sim} \text{Gamma}\left(\frac{m_g+1}{2},\frac{\lambda^2}{2}\right), \quad g=1,\ldots,G,\\
&\pi_0 \ \sim \ \text{Beta}(a,b)
\end{aligned}$$

The full posterior conditional distributions are as follows

$$\begin{aligned}
p\left(\beta,\tau^2,\pi_0|\boldsymbol{Y},\boldsymbol{X}\right) &\propto \prod_{i=1}^{n}\left[\left(\frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}\right)^{y_i}\left(\frac{1}{1+e^{x_i^T\beta}}\right)^{1-y_i}\right] \propto \prod_{g=1}^{G}\left[(1-\pi_0)\left(2\pi\tau_g^2\right)^{-\frac{m_g}{2}} e^{-\frac{\beta_g^T\beta_g}{2\tau_g^2}} I_{[\beta_g\neq 0]} + \pi_0\delta_0\left(\beta_g\right)\right]\\
&\propto \prod_{g=1}^{G}(\lambda^2)^{\frac{m_g+1}{2}}\left(\tau_g^2\right)^{\frac{m_g+1}{2}-1} e^{-\frac{\lambda^2\tau_g^2}{2}} \propto \pi_0^{\mathrm{a}-1}(1-\pi_0)^{b-1}
\end{aligned} \tag{7}$$

We can simulate an efficient block Gibbs sampler to simulate from the posterior distribution above. Details of the block Gibbs sampler is part of the supplemental material.

4 Posterior consistency

Xu and Ghosh[9] showed that the posterior median is an adaptive thresholding estimator for a linear regression setup. Theorem 1 in their paper gives a proof of this idea. We will extend this idea for logistic regression model numerically.

To prepare the ground for posterior consistency, we will rewrite our model using different notations just so it is aligned with the model setup of Jiang.[31] Our proof of consistency is in line with Jiang's[31] paper. Similar notations will ease understanding of the proof.

Let $D^n = \{y; X_1, \ldots, X_{Pn}: y \in \{0,1\}, X_i \in \mathrm{R}^n, i = 1, \ldots, P_n\}$ denote a dataset of n observations each consisting of P_n predictors where P_n can increase with increasing n. We want to model this data using logistic regression. Let ξ_n denote a chosen (subset) model, and $|\xi_n|$ denote the model size of ξ_n. Note that, here ξ_n is the sum of all dummy variables (factor levels) of the groups that are chosen in the subset model. Let us call G^*, the number of selected groups then $G^* \leq \xi_n$. A major difference of this and Jiang's[31] setup is the multivariate β_g, $g = 1, \ldots, G$. An interesting thing to note is that, if we express our setup in terms of the dummy variables, then this layout is similar to what Jiang[31] proposed. Thus, when the chosen model is ξ_n, we are really considering our chosen group size to be G^* and the model size to be $\sum_{g=1}^{G^*} m_g = |\xi_n|$. To make the proof here in line with Jiang's[31] paper, we express the chosen model in terms of the dummy variables rather than the groups. Clearly, this is an extension of Jiang's[31] model since we consider a grouped structure for βs. Conditional on ξ_n, the regression coefficients

$$\beta_{\xi_n} | \tau_{\xi_n} \sim N(0, V_{\xi_n})$$

where $\mathrm{V}_{\xi n}$ is a $|\xi_n| \times |\xi_{\mathrm{n}}|$ covariance matrix and a function of τ_{ξ_n}. Here, $\beta_{\xi_n} = \left(\beta_1^{*T}, \ldots, \beta_{G^*}^{*T}\right)$ and $\tau_{\xi_n} = (\tau_1^*, \ldots, \tau_{G^*}^*)$ denote the vector of true regression coefficients and true variance parameters, respectively, such that $\sum_{g=1}^{G^*} m_g = |\xi_n|$. Let $\left\{X_1^*, \ldots, X_{\xi_n}^*\right\} \subset \{X_1, \ldots, X_{P_n}\}$ denote the predictors chosen in model ξ_n. Note that

$$V_{\xi_n} = \begin{pmatrix} \tau_1^{*2} I_{m_1} & 0 & \cdots & 0 \\ 0 & \tau_2^{*2} I_{m_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tau_{G^*}^{*2} I_{m_{G^*}} \end{pmatrix}$$

and $\tau_g^2 \overset{ind}{\sim} \mathrm{Gamma}\left(\frac{m_g+1}{2}, \frac{\lambda^2}{2}\right), g = 1, \ldots G$. Let model ξ_n have the prior $\prod(\xi_n | \pi_0) = \pi_0^{P_n - |\xi_n|}(1 - \pi_0)^{|\xi_n|}$, and $\pi_0 \sim \mathrm{Beta}(a,b)$
where a and b are pre-specified hyperparameters. Thus

$$\prod(\xi_n) = \frac{Beta(a + P_n - |\xi_n|, |\xi_n| + b)}{Beta(a,b)}, \text{ and } \Pi\left(\beta_{\xi_n}\right) = \int \Pi\left(\beta_{\xi_n}, \tau^2 \xi_n\right) d\tau^2 \xi_{\mathrm{n}} = \prod_{g=1}^{G^*} \int_0^\infty \Pi\left(\beta_{\xi_n} | \tau^2 \xi_n\right) \Pi\left(\tau^2 \xi_n\right) d\tau_g^2$$

$$= \prod_{g=1}^{G^*} \int_0^\infty \frac{e^{-\frac{\beta_g^T \beta_g}{2\tau_g^2}}}{\left(2\pi\tau_g^2\right)^{\frac{m_g}{2}}} \left(\lambda^2\right)^{\frac{m_g+1}{2}} \left(\tau_g^2\right)^{\frac{m_g+1}{2}-1} e^{-\frac{\lambda^2 \tau_g^2}{2}} d\tau_g^2$$

Substituting $\alpha_g^2 = \frac{1}{\tau_g^2}$ we have

$$\Pi\left(\beta_{\xi_n}\right) = \prod_{g=1}^{G^*} \left(\lambda^2\right)^{\frac{m_g}{2}} \frac{e^{-\lambda\sqrt{\beta_g^T \beta_g}}}{(2\pi)^{\frac{m_g-1}{2}}} \int_0^\infty \left(\frac{\lambda^2}{2\pi\left(\alpha_g^2\right)^3}\right)^{\frac{1}{2}} e^{-\frac{\lambda^2 \beta_g^T \beta_g}{2\lambda^2 \alpha_g^2}\left(\alpha_g^2 - \frac{\lambda}{\sqrt{\beta_g^T \beta_g}}\right)^2} d\alpha_g^2$$

The term in the integral is an Inverse Gaussian density i.e. ${\alpha_g}^2 \sim$ Inverse Gaussian$\left(\frac{\lambda}{\sqrt{\beta_g^T \beta_g}}, \lambda^2\right)$, thus the integrals integrate to one for all $g = 1,\ldots,G^*$. Therefore

$$\Pi(\beta_{\xi_n}) = \prod_{g=1}^{G^*} (\lambda^2)^{\frac{m_g}{2}} \frac{e^{-\lambda\sqrt{\beta_g^T \beta_g}}}{(2\pi)^{\frac{m_g-1}{2}}} \tag{8}$$

Let ξ_n be the model obtained from the median thresholding posterior probability and ξ_n^* be the true model. We want to show that the model ξ_n converges to the true model ${\xi_n}^*$ as the sample size n becomes sufficiently large. Define f^* as the true density under model ξ_n^* and f as the density proposed under model ξ_n. Hellinger distance between f and f^* is defined as.

$$d(f,f^*) = \sqrt{\iint \left(\sqrt{f} - \sqrt{f^*}\right)^2 \upsilon_y(d_y)\upsilon_x(d_x)}$$

To investigate posterior convergence, we formulate the following theorem based on Theorem 4 in Jiang's[31] paper. We consider logistic regression in this paper with a density of the form

$$p^*(y|x) = \exp\{a(h^*)y + b(h^*) + c(y)\} \equiv f(y, h^*)$$

where $h^* = x^T\beta^*$ is the linear parameter, $a(h)$ and $b(h)$ are continuously differentiable, and $a(h)$ has non-zero derivative. The mean function

$$\mu^* = E(y|x) = -\frac{b'(h^*)}{a'(h^*)} \equiv \psi(x^T\beta^*) = \frac{e^{h^*}}{1+e^{h^*}}$$

Thus ψ is the inverse of the logistic link function. Assume that $\lim_{n\to\infty} \sum_{g=1}^{G} \sqrt{\beta_g^{*T}\beta_g^*} < \infty$. For simplicity, let ξ be the corresponding subset model for which $|\beta| > 0$ and let r_n be the prior expectation of model size $|\xi|$. Define

$$\Delta(r_n) = \inf_{\xi:|\xi|=r_n} \sum_{j:j\notin\xi} |\beta_j^*| < \infty, \qquad B(r_n) = \sup_{\xi:|\xi|=r_n} ch_1\left(V_\xi^{-1}\right) \qquad \bar{B}(r_n) = \sup_{\xi:|\xi|=r_n} ch_1(V_\xi), \text{ and}$$

let, $\widetilde{B}_n = \sup_{\xi:|\xi|\le K_n} ch_1(V_\xi)$ where K_n is the maximal model size. Let $D(R) = 1 + R \times \sup_{|h|\le R} |a'(h)|.\sup_{|h|\le R} |\psi(h)|$ for any $R > 0$. Here, $ch_1(V_\xi)$ and $ch_1(V_\xi^{-1})$ are the largest eigenvalues of V_ξ and ${V_\xi}^{-1}$, respectively.

Let $\epsilon_n \in (0,1]$ for each n and $n\epsilon_n \succ 1$ and assume the following conditions hold

1. $K_n \log\left(\frac{1}{\epsilon_n^2}\right) \prec n\epsilon_n^2$
2. $K_n \log(P_n) \prec n\epsilon_n^2$
3. $K_n \log\left(D\left(K_n \frac{\bar{B}_n n\epsilon_n^2}{\lambda_n}\right)\right) \prec n\epsilon_n^2$
4. $\mathrm{r_n} \prec P_n$
5. $r_n \log \bar{B}_n(r_n) \prec n\epsilon_n^2$ and $\Delta(r_n) \prec n\epsilon_n^2$
6. $\log\left(\frac{r_n}{P_n}\right) \le -\frac{4n\epsilon_n^2}{P_n}$
7. m_g is such that $\sum_{g=1}^{G^*} m_g \prec P_n, \forall\ g = 1,\ldots,G$

We will replace λ by λ_n since λ and ${\tau_g}^2$, $g = 1,\ldots,G^*$ are dependent on n. Also, λ_n is inversely proportional to the sum of all ${\tau_g}^2$s.

Theorem 1. Assume the prior setting on (8) is used and the Assumption (A1) holds. Let $P\{.\}$ denote the probability measure for the data D^n. Assume, $G < P_n$, $1 \leq \lambda_n \leq B(r_n), |x_j| \leq 1$ for all j and $\lim_{n\to\infty} \sum_{g=1}^{G} \sqrt{\beta_g^{*T}\beta_g^*} < \infty$

where P_n is a nondecreasing sequence in n. Also, let V_ξ be such that $\widetilde{B}_n \geq 4$

Let ε_n be a sequence such that $\epsilon_n \in (0,1]$ for each n and $n\epsilon_n^2 \succ 1$ and ${\tau_g}^2 < \infty$, $g = 1,\ldots,G^*$. Then, we have,

(i) *For some $c_o > 0$*

$$\lim_{n\to\infty} P\left\{\pi\left[d(f,f^*) \leq \epsilon_n | D^n\right] \geq 1 - e^{-c_0 n\epsilon_n^2}\right\} = 1, \text{ and}$$

(ii) *For some $C_1 > 0$ and for all sufficiently large n*

$$P\left\{\pi\left[d(f,f^*) > \epsilon_n | D^n\right] \geq e^{-0.5c_1 n\epsilon_n^2}\right\} \leq e^{-0.5c_1 n\epsilon_n^2}$$

Proof of theorem 1 is part of the supplemental material.

5 Simulation

Before applying the group level Bayesian selection method on the brain image data, we run a simulation study. The simulation study has the unknown parameters in control and tests the method on controlled inputs. We work on two different scenarios where the first case is high-dimensional while the second case is large n small p scenario.

(S1) The number of observations is 60 and there are 16 predictors each with 5 levels. Thus the number of parameters here is really 80. So, we are essentially looking at a small n large p problem here. The setup here is adapted from Example 2 of Xu and Ghosh[7]'s paper. We define the jth predictor as $X_{gj} = z_g + z_{gj}$, where z_g and z_{gj} are independent standard normal variables and $g = 1,\ldots,16$, $j = 1,\ldots,5$. Thus, the predictors in a group are correlated but the predictors in different groups are independent. Assign true parameter values as follows: $\beta = ((7,6,3,4,5),\mathbf{0},(4,5,6,10,7),\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},(2,3,4,5,6))$ where $\mathbf{0}$ is a zero vector of length 5. Use the simulated X and β to generate 60 independent Bernoulli random values using the logit link. Here, 40 observations are used to train the model and the rest are used as a test dataset.

(S2) The number of observations is 100 and there are 4 predictors each with 10 levels which makes $p = 40$. The design matrix is generated exactly as in Scenario 1. Let $\beta = (\mathbf{0},\mathbf{2},\mathbf{0},\mathbf{2})$ where $\mathbf{0}$ and $\mathbf{2}$ are vectors of length 10 with all elements 0 and 2 respectively. Use the simulated X and β to generate 100 independent Bernoulli random values using the logit link. Sixty randomly selected rows were used as train dataset and the remaining as test data.

Hyperparameters, for both cases, a and b were both set to 1.5. Twenty thousand Monte-Carlo iterations were implemented. Twenty-eight bootstrapped samples were used to average out bias in estimates.

Table 1 summarizes the true and false positive rates and the negative log-likelihood of the two examples mentioned above. Both the methods are able to identify the true variables although the frequentist group lasso has a high false positive rate. This indicates that the group lasso tends to select more variables for an optimal tuning parameter. On the other hand, the model selected by median thresholding gives excellent result in terms of variable selection. We see that the method proposed in this paper gives a smaller negative log-likelihood indicating

Table 1. Mean (standard error), true/false positive rate, and negative log-likelihood in 28 simulations.

Scenario		Bayesian spike and slab group lasso	Frequentist group lasso
S1	True positive rate	1.00(0.00)	1.00(0.00)
	False positive rate	0.00(0.00)	0.78(0.50)
	Neg log-likelihood	2.65(0.53)	6.52(0.57)
S2	True positive rate	1.00(0.00)	1.00(0.00)
	False positive rate	0.00(0.00)	0.32(0.24)
	Neg log-likelihood	–2.35(2.31)	5.45(1.59)

a better model fit. To be more stringent about feature selection, we also look at the credible intervals of the estimated coefficients. Our investigation shows all the selected features are statistically significant based on the credible intervals. Since, our motivation is classification here, we omit these results. Thus, we see that in a simulated dataset, the median thresholding method is able to classify variables very well as compared to the conventional group lasso method when we have variables that have a structured correlation.

6 Application to ADNI MRI data

The MRI data used in this section of the paper were obtained from ADNI database. The main objective of ADNI has been to test whether serial MRI, PET, other biological biomarkers, and clinical and neuropsychological assessment can be used to detect dementia or measure its progression. Both normal aging and AD patients have brain region atrophies but it is essential to identify the abnormalities that lead to dementia. Some studies are done to study the differences of brain atrophy in these two categories of subject (Double et al.[32]). Such studies have shown that there is a significant difference in the atrophies of normal aging and AD patients, so we use this idea to classify the subjects. In this paper, we delve into classification of AD patients from normally aging control (CN) subjects at the baseline and estimation of parameters of selected volumetrics. The parameter estimates give us the log of odds of being AD at baseline for a subject with a given set of volumetric measurements. Thus, baseline volumetric values for AD and normal controls from ADNI dataset serve our purpose. ADNI data are collected from 2003 onwards by National Institute of Aging (NIA), National Institute of Biomedical Imaging and Bioengineering (NIBIB), U.S. Food and Drug Administration (FDA), and a few pharmaceutical companies as a public–private partnership. The ADNI project is a large project involving subjects across USA and Canada from more than 50 sites. This initiative was launched to develop new treatments and follow subjects through time to monitor the effectiveness of the treatments. For more information about ADNI, visit www.adni-info.org.

The volumetric segmentation and cortical reconstruction of the brain is done with the help of freely available software FreeSurfer. An early version of the longitudinal image processing framework (Reuter et al.[33]) is used to process the sequential scans. This process does motion correction and averaging of multiple volumetric TI weighted images, removes non-brain segments, automates Talairach transformation, segments subcortical white matter (WM) and deep gray matter (GM) volumetric structures. It also automates topology correction and surface deformation of the brain. For a detailed guide, please refer to the UCSF FreeSurfer Methods documented by Hartig et al.[34] Due to advancement of technology in the computing area, quantitative assessment of brain volumes, obtained through volumetric MRI is being used extensively for studies involving AD. Volumetric measurements are mainly based on brain segmentation done at reliable MR centers.

Many studies have been done to identify the ROIs associated with AD but using the entire brain segmentation to identify four different volumetric aspects of a region has not been explored. Sabuncu et al.[35] observed that baseline thickness in AD vulnerable cortical regions reduced significantly in AD patients. A particular attribute measurement of a brain subregion may be more indicative of AD atrophy than others. Thus, we are interested in not only selecting the atrophied brain subregion but also identifying which of the volume, surface area or thickness changes differ significantly in AD subjects from healthy controls. This technique includes all available subregion data in the model and identifies the subregions that are potentially associated with AD. Previous studies have isolated one or a few brain regions and used their volume measurements as a predictor of AD or MCI from CNs (Jack et al.[36,37]). We want the model to automatically select the atrophied regions rather than subsetting a brain region before start of the analysis. Classification of AD from CN has been done using FDG-PET scan (Herholz et al.[38]) using comparative statistical methods like *t*-statistics but researchers are yet to explore variable selection techniques using the entire volumetric data. Wang et al.[39] used Haar wavelets to identify ROIs using voxel level data for dimension reduction. This method identifies ROIs successfully but does not narrow down the brain hemisphere of the ROIs. Since, our data are present for each region for the left and right hemispheres, we are able to identify the exact part of the ROI that is more significantly associated with AD. For the analysis, we use AD and CN patients to distinctively understand the difference of brain regions that cause a subject to be cognitively normal or progress to AD.

We have used the longitudinal processing data for our analysis. The demographic characteristics of the 421 subjects are given in Table 2. The age and sex distribution in our dataset shows that the data are not skewed with respect to these two variables. Also, the minimum and maximum age for AD is 55.1 and 90.9, respectively, and that of CN is 59.9 and 89.6, respectively. Thus, the effect of age on the outcome has been controlled for in the ADNI dataset.

Table 2. Demographics of patients in ADNI data used for analyses.

Category	Gender: Male (%)	Age in years: Mean (SD)
AD (n = 191)	100 (52.36%)	75.27 (7.46)
CN (n = 230)	120 (52.17%)	75.86 (5.01)

SD: standard deviation.

We have used baseline data of 421 subjects of whom 191 have AD and 230 are cognitively normal subjects. There are 72 predictors (brain regions segmented with FreeSurfer) with four levels each namely, volume, area, thickness average, and thickness standard deviation. Forty-six brain regions had volume data only. The regions marked "Unknown" and "Undetermined" were discarded beforehand because these regions were not identified in the MR scans. We used all the remaining 116 brain regions (single and four-leveled) as predictors for dementia. Our objective is to be able to select an optimal model that identifies the important brain regions for identifying the two kinds of brain cognitive functionality. The brain regions are segmentation of both GM and WM. Studies suggest that the GM is associated with cognitive disorders in elderly people. We keep both GM and WM to test the efficacy of our model i.e. if the model is efficient in selecting the correct brain regions. The analysis to identify the most significant regions from the entire brain region is done. Variable selection selects the significant brain region from a large collection of brain regions and then the model successfully classifies the subjects using the test dataset.

We have a logistic model for the two outcomes of the response variable. Around 70% of the data is used to train the model. The prior placed on the coefficients is a spike and slab type prior that encourages zero estimates for predictors which are not significant. The model is selected using median thresholding method. We run 10,000 iterations of the Monte-Carlo Markov Chain (MCMC) chain of which the first 5,000 are used as burn-ins. The usual convergence diagnostics are performed.

Two hundred and ninety-five subjects were randomly selected from the 421 subjects to train the model. The median thresholding model selects 29 out of the 116 brain subregions. These regions correspond to ROIs, namely, right bankssts, right pallidum, pars opercularis, left pars orbitals, right precuneus, putamen, right anterior cingulate cortex, superior frontal, entorhinal cortex, supramarginal gyrus, right transverse temporal, left hippocampus, left inferior lateral ventricle, middle temporal gyrus, inferior temporal gyrus, left precentral gyrus, right fusiform gyrus, left parahippocampal, paracentral, third ventricle (Feng et al.[40]), and right inferior parietal. We now want to understand if all these regions' attributes are statistically significant. Some regions have negligible amount of contribution in the model thus making the credible interval of the feature exclude the corresponding subregion. We only keep the subregions and the measurement attributes in our model which are statistically significant as given by the corresponding credible intervals. This result tells us that only the thickness average, thickness SD, surface area, volume or a combination of these attributes can be significant for a subregion while in other cases all these four attributes can be significant.

The statistically significant ROIs are given in Table 3. Previous studies have established the association of these regions in AD. Volumes of right entorhinal cortex are severely diminished in AD patients. The other regions selected are also coherent with relevant literature (Juottonen et al.,[41] Galton et al.[42]). The precentral gyrus controls motor skills, middle temporal gyrus regulates semantic memory processing, hippocampus, parahippocampal and entorhinal regulate memory and navigation. Putamen, pallidum, transverse temporal and bankssts are all known to be affected by AD (Clerx et al.[43]). Recent studies have separately analyzed all these regions and found atrophies in those areas. The proposed method identifies a subset of all atrophied regions and then selects fewer regions as discriminative features for classification.

The functions of the selected regions also intuitively implicate the precision of the model. Zhang et al.[10] classified AD and CN using support vector machine (SVM) thus leading to non-interpretability of the associated coefficients.

The method achieved fairly high accuracy of 80%. Cuingnet et al.[44] classified AD and CN based on ROIs but they restricted their analysis to a few selected ROIs namely the entorhinal thickness, supramarginal cortex thickness and hippocampal volume. Their sensitivity ranged from 69% to 70% whereas our method gives a sensitivity of 76%. The specificity in their study (90%) is, however, higher than ours (83%). These two studies are, however, not directly comparable except that they are both classification studies because the datasets used in these studies are different. The drawback of their method is that they pre-select a few ROIs and perform the classification

Table 3. Mean (standard error) of parameter estimates of selected ROIs.

ROI	Volume	Surface area	Cortical thickness avg.	Cortical thickness SD
Right entorhinal cortex	2.97 (0.03)	1.83 (0.02)	2.14 (0.01)	1.69 (0.03)
Right pallidum	–0.27 (0.12) [b]	[a]	[a]	[a]
Right pars orbitals		[b]	–0.33 (0.16)	[b]
Right precuneus	0.39 (0.05)	–0.42 (0.07)	1.0 (0.09)	[b]
Right putamen	–0.21 (0.07)	[a]	[a]	[a]
Left anterior cingulate	–0.46 (0.12)	–0.12 (0.04)	0.15 (0.05)	0.09 (0.04)
Right transverse temporal	[b]	[b]	–0.54 (0.22)	[b]
Left entorhinal cortex	[b]	0.22 (0.07)	0.21 (0.07)	–0.41 (0.13)
Left precentral gyrus	0.25 (0.08)	0.38 (0.12)	0.45 (0.15)	0.25 (0.09)
Left parahippocampal gyrus	1.97 (0.11)	1.90 (0.11)	2.0 (0.11)	2.07 (0.11)
Right bankssts	0.06 (0.02)	[b]	[b]	[b]
Left middle temporal gyrus	1.93 (0.04)	1.57 (0.04)	1.91 (0.02)	1.63 (0.03)
Left hippocampus	1.50 (0.02)	[a]	[a]	[a]

All ROI measurements in the table are statistically significant with the following exceptions.
[a]Means these region measurements were not captured in data.
[b]Means that these regions were not statistically significant although the region was selected by median thresholding.
ROI: region of interest; SD: standard deviation.

unlike our method. On the other hand, the proposed method is based on statistical foundations that is accounting and measuring uncertainties due to randomness in the data set.

Our logistic model coded CN as 1 and AD as 0 so the parameter estimates should be interpreted accordingly. Table 3 gives the mean parameter estimate and standard error (within parentheses) of the selected ROIs.

7 Discussion

In this paper, we propose a Bayesian approach of variable selection with spike and slab prior to identify cognitively healthy controls from Alzheimer's patients. This method uses whole brain parcellation data to classify dementia as well as interpret the association of each significant volumetric measure of a brain subregion. This technique captures the structured correlation in this type of data to retain all levels of the subregion that are disease related. The Bayesian approach guarantees better standard error estimates. Also, the median thresholding method for posterior model selection together with the use of spike and slab prior is a more efficient method than the frequentist group lasso method as shown in the simulation study. Liang et al.,[45] developed a Bayesian subset selection method for GLMs which can select individual variables only. In their method, they place a prior on the model to perform subset selection unlike our approach of using a spike and slab prior which directly drops out variables in the many Monte Carlo iterations. Our median thresholding, as opposed to their Maximum a posteriori (MAP) posterior probability, is able to choose the best model without comparing information criterion type quantities among several candidate models. Thus, the spike and slab prior median thresholding Bayesian group lasso has attractive properties of high-dimensional variable selection and performs efficiently with structured correlated covariates. Most of the other dimension reduction techniques are unable to tackle correlated variables in variable selection.

The significant regions selected by our model are identified from a large number of subregions thus accounting for the effect of the whole brain while performing dimension reduction. Wang et al.[39] in their paper have introduced a dimension reduction technique using Haar wavelet based on voxel level data with ADNI PET data. Also, their method builds on continuous outcomes. Our approach builds the model with the MRI brain parcellated volumetric data which are a direct indicator of dementia. This Bayesian formulation not only tackles ANOVA type dummy variables but also deals with the high dimension problem. The simulation results show that this method is effective in both low and high dimensions. The greatest advantage of this method is that it considers all the subregions while building the model and efficiently drops the ones that are not disease related and also easily interpret the risk of dementia from the parameter estimates. Thus, this method is effective in finding a needle from a stack of hay. This is a novel contribution to classification for AD to the best of our knowledge.

Acknowledgements

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 authors would like to thank the investigators within the ADNI who contributed to the design and implementation of ADNI and/or provided data. A complete listing of ADNI investigators can be found at this link (accessed on 26 January 2017). The authors would also like to thank Dr David Zhu and Dr Andrea Bozoki, Department of Radiology and Neurology, Michigan State University, for their comments and useful discussions.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Subha Datta https://orcid.org/0000-0003-4864-3518

Supplemental material

Supplemental material for this article is available online.

References

1. Leifer BP. Early diagnosis of Alzheimer's disease: clinical and economic benefits. *J Am Geriatr Soc* 2003; **51**: S281–S288.
2. Smith CD, Andersen AH and Gold BT. Structural brain alterations before mild cognitive impairment in ADNI: validation of volume loss in a predefined antero-temporal region. *J Alzheimer's Dis* 2012; **31**: S49–S58.
3. Arlt S, Buchert R, Spies L, et al. Association between fully automated MRI-based volumetry of different brain regions and neuropsychological test performance in patients with amnestic mild cognitive impairment and Alzheimer's disease. *Eur Arch Psychiatr Clin Neurosci* 2013; **263**: 335–344.
4. Haroutunian V, Katsel P and Schmeidler J. Transcriptional vulnerability of brain regions in Alzheimer's disease and dementia. *Neurobiol Aging* 2009; **30**: 561–573.
5. Shivamurthy VK, Tahari AK, Marcus C, et al. Brain FDG PET and the diagnosis of dementia. *Am J Roentgenol* 2015; **204**: W76–W85.
6. Luo WL and Nichols TE. Diagnosis and exploration of massively univariate neuroimaging models. *NeuroImage* 2003; **19**: 1014–1032.
7. Grimmer T, Riemenschneider M, Forstl H, et al. Beta amyloid in Alzheimer's disease: increased deposition in brain is reflected in reduced concentration in cerebrospinal fluid. *Biol Psychiatry* 2009; **65**: 927–934.
8. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol* 1996; **58**: 267–288.
9. Xu X and Ghosh M. Bayesian variable selection and estimation for group lasso. *Bayesian Anal* 2015; **10**: 909–936.
10. Zhang D, Wang Y, Zhou L, et al., Alzheimer's Disease Neuroimaging Initiative. Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage* 2011; **55**: 856–867.
11. Fan J and Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 2001; **96**: 1348–1360.
12. Breiman L. Better subset regression using the nonnegative garrote. *Technometrics* 1995; **37**: 373–384.
13. Hore S, Dewanji A and Chatterjee A. Design issues related to allocation of experimental units with known covariates into two treatment groups. *J Stat Plan Inference* 2014; **155**: 117–126.
14. Hore S, Dewanji A and Chatterjee A. On optimal allocation of experimental units with known covariates into multiple treatment groups. *Calcutta Stat Assoc Bull* 2016; **68**: 69–81.
15. Hore S, Chatterjee A and Dewanji A. Improving variable neighborhood search to solve the traveling salesman problem. *Appl Soft Comput* 2018; **68**: 83–91.
16. Yuan M and Lin Y. Model selection and estimation in regression with grouped variables. *J R Stat Soc Series B Stat Methodol* 2006; **68**: 49–67.
17. Knight K and Fu W. Asymptotics for lasso-type estimators. *Ann Stat* 2000; **28**: 1356–1378.
18. Chatterjee A and Lahiri SN. Bootstrapping lasso estimators. *J Am Stat Assoc* 2011; **106**: 608–625.
19. Park T and Casella G. The Bayesian lasso. *J Am Stat Assoc* 2008; **103**: 681–686.
20. Kyung M, Gill J, Ghosh M, et al. Penalized regression, standard errors, and Bayesian lassos. *Bayesian Anal* 2010; **5**: 369–411.
21. Tibshirani R, Saunders M, Rosset S, et al. Sparsity and smoothness via the fused lasso. *J R Stat Soc Series B Stat Methodol* 2005; **67**: 91–108.

22. Zou H and Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol* 2005; **67**: 301–320.
23. Narisetty NN and He X. Bayesian variable selection with shrinking and diffusing priors. *Ann Stat* 2014; **42**: 789–817.
24. George EI and McCulloch RE. Approaches for Bayesian variable selection. *Stat Sin* 1997; **7**: 339–373.
25. Chen Z and Dunson DB. Random effects selection in linear mixed models. *Biometrics* 2003; **59**: 762–769.
26. Zhao Z and Sarkar S. On credible intervals for selected parameters under the zero-inflated mixture prior in high dimensional inference. Unpublished manuscript, 2012.
27. Lykou A and Ntzoufras I. On Bayesian lasso variable selection and the specification of the shrinkage parameter. *Stat Comput* 2013; **23**: 361–390.
28. Zhang L, Baladandayuthapani V, Mallick BK, et al. Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer. *J R Stat Soc Series C Appl Stat* 2014; **63**: 595–620.
29. Casella G. Empirical Bayes Gibbs sampling. *Biostatistics* 2001; **2**: 485–500.
30. Meier L, van de Geer S and Buhlmann P. The group lasso for logistic regression. *J R Stat Soc Series B Stat Methodol* 2008; **70**: 53–71.
31. Jiang W. Bayesian variable selection for high dimensional generalized linear models: convergence rates of the fitted densities. *Ann Stat* 2007; **35**: 1487–1511.
32. Double KL, Halliday GM, Krill JJ, et al. Topography of brain atrophy during normal aging and Alzheimer's disease. *Neurobiol Aging* 1996; **17**: 513–521.
33. Reuter M, Schmansky NJ, Rosas HD, et al. Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage* 2012; **61**: 1402–1418.
34. Hartig M, Truran-Sacrey D, Raptentsetsang S, et al. *UCSF freesurfer methods*. ADNI Alzheimers Disease Neuroimaging Initiative: San Francisco, CA, USA, 2014.
35. Sabuncu MR, Desikan RS, Sepulcre J, et al. The dynamics of cortical and hippocampal atrophy in Alzheimer disease. *Arch Neurol* 2011; **68**: 1040–1048.
36. Jack CR, Petersen RC, Xu YC, et al. Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. *Neurology* 1999; **52**: 1397.
37. Jack CR, Petersen RC, Xu YC, et al. Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease. *Neurology* 1997; **49**: 786–794.
38. Herholz K, Salmon E, Perani D, et al. Discrimination between Alzheimer dementia and controls by automated analysis of multicenter FDG PET. *Neuroimage* 2002; **17**: 302–316.
39. Wang X, Nan B, Zhu J, et al. Regularized 3D functional regression for brain image data via Haar wavelets. *Ann Appl Stat* 2014; **8**: 1045.
40. Feng R, Wang H, Wang J, et al. Forebrain degeneration and ventricle enlargement caused by double knockout of Alzheimer's presenilin-1 and presenilin-2. *Proc Natl Acad Sci USA* 2004; **101**: 8162–8167.
41. Juottonen K, Laakso MP, Insausti R, et al. Volumes of the entorhinal and perirhinal cortices in Alzheimer's disease. *Neurobiol Aging* 1998; **19**: 15–22.
42. Galton CJ, Patterson K, Graham K, et al. Differing patterns of temporal atrophy in Alzheimer's disease and semantic dementia. *Neurology* 2001; **57**: 216–225.
43. Clerx L, Jacobs HIL, Burgmans S, et al. Sensitivity of different MRI-techniques to assess gray matter atrophy patterns in Alzheimer's disease is region-specific. *Curr Alzheimer Res* 2013; **10**: 940–951.
44. Cuingnet R, Gerardin E, Tessieras J, et al., Alzheimer's Disease Neuroimaging Initiative. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. *Neuroimage* 2011; **56**: 766–781.
45. Liang F, Song Q and Yu K. Bayesian subset modeling for high-dimensional generalized linear models. *J Am Stat Assoc* 2013; **108**: 589–606.