Published in final edited form as:

Stat Med. 2019 May 30; 38(12): 2184–2205. doi:10.1002/sim.8100.

High-dimensional longitudinal classification with the multinomial fused lasso

Samrachana Adhikari¹, Fabrizio Lecci², James T. Becker³, Brian W. Junker², Lewis H. Kuller⁴, Oscar L. Lopez³, and Ryan J. Tibshirani²

¹Department of Population Health, New York University School of Medicine, New York, New York

²Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania

³Department of Neurology, University of Pittsburgh, Pittsburgh, Pennsylvania

⁴Department of Epidemiology, University of Pittsburgh, Pittsburgh, Pennsylvania

Abstract

We study regularized estimation in high-dimensional longitudinal classification problems, using the lasso and fused lasso regularizers. The constructed coefficient estimates are piecewise constant across the time dimension in the longitudinal problem, with adaptively selected change points (break points). We present an efficient algorithm for computing such estimates, based on proximal gradient descent. We apply our proposed technique to a longitudinal data set on Alzheimer's disease from the Cardiovascular Health Study Cognition Study. Using data analysis and a simulation study, we motivate and demonstrate several practical considerations such as the selection of tuning parameters and the assessment of model stability. While race, gender, vascular and heart disease, lack of caregivers, and deterioration of learning and memory are all important predictors of dementia, we also find that these risk factors become more relevant in the later stages of life.

Keywords

Alzheimer's disease; cardiovascular health study; cognition study; longitudinal observational data; regularization

1 | INTRODUCTION

The prevalence of Alzheimer's disease (AD) increases at an alarming rate beyond the age of 65. After 90 years of age, the incidence of AD increases dramatically, from 12.7% per year in the 90–94 age group to 21.2% per year in the 95–99 age group and to 40.7% per year for those older than 100 years.^{1–3} Possibility of clinical intervention to control the dramatic progression of AD on patients at risk hinges on identifying risk factors that can predict the

Correspondence Samrachana Adhikari, Department of Population Health, New York University School of Medicine, 180 Madison Avenue, New York, NY 10016., Samrachana.Adhikari@nyulangone.org.

DISCLAIMER

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

time of onset of the disease. Many previous studies have shown the importance of a range of risk factors in predicting the time of onset of clinical expression of AD, either as dementia or its prodromal syndrome, mild cognitive impairment (MCI). For example, the risk of dementia is associated with the presence of the APOE*4 allele, male sex, lower education, and having a family history of dementia.^{2,4,5} Medical risk factors include the presence of systemic hypertension, diabetes mellitus, and cardiovascular or cerebrovascular disease.^{6–8} Lifestyle risk factors include physical and cognitive activity, and diet.^{9–11} The complex relationships between age and the risk factors produce highly variable natural histories from normal cognition to the clinical expression of AD, thus making the prediction of the time of onset of AD challenging.^{12–16}

Our work is particularly motivated by the analysis of a large longitudinal data set provided by the Cardiovascular Health Study Cognition Study (CHS-CS). Over the past 24 years, the CHS-CS recorded multiple demographic, metabolic, cardiovascular, and neuroimaging risk factors for AD, as well as detailed cognitive assessments for people of ages 65 to 110 years old.^{12–14} A wide range of statistical approaches has been considered in the previous studies of the CHS-CS data for predicting the time of onset of AD, including exploratory statistical summaries, hypothesis tests, survival analyses, logistic regression models, and latent trajectory models. However, none of these methods can directly accommodate a large number of predictors that can potentially exceed the number of observations. A small number of variables was often chosen a priori to match the requirements of a particular model, neglecting the full potential of the CHS-CS data, which consists of thousands of variables.

In this paper, we examine data from 924 individuals in the Pittsburgh section of the CHS-CS, to identify important risk factors for the prediction of the cognitive status at $t + 10$ years of age, given predictor measurements at t years of age, for $t = 65, 66, \dots, 98$. For each age, the outcome variable assigns an individual to one of the three cognitive status: normal, dementia, or death. While we are mainly interested in the prediction of dementia, this task is complicated by the fact that risk factors for dementia are also known to be risk factors for death.¹⁷ To account for this competing risk of death, we include death as a separate category in the prediction model. We further develop a simulation study to demonstrate that the approach we introduce can accommodate an array of predictors of arbitrary dimension, using regularization to maintain a well-defined predictive model, and avoids overfitting, compared to other competing prediction models.

More generally, we study longitudinal classification problems in which the number of predictors can exceed the number of observations. The setup: we observe n individuals across discrete time points $t = 1, \dots, T$. At each time point, we record p predictor variables per individual, and an outcome that places each individual into one of K classes. The goal is to construct a model that predicts the outcome of an individual at time $t + \Delta$, given his or her predictor measurements at time t . Since we allow for the possibility that $p > n$, regularization must be employed in order for such a predictive model (eg, based on maximum likelihood) to be well defined. Borrowing from the extensive literature on high-dimensional regression, we consider two well-known regularizers, each of which also has a natural place in high-dimensional longitudinal analysis for many scientific problems of interest. The first is the

lasso regularizer, which encourages overall sparsity in the active (contributing) predictors at each time point; the second is the *fused lasso* regularizer, which encourages a notion of persistence or contiguity in the sets of active predictors across time points. Justification for this second penalty is based on the scientific intuition that predictors that are clinically important should have similar effects at successive ages.

There has been recent development in using high-dimensional data analysis tools for understanding the progression of AD.^{18–21} These statistical tools are built to utilize special features of data, such as information on brain biomarkers from magnetic resonance images (MRI), for prediction. In fact, the fused lasso has been applied to the study of AD in the work of Xin et al.²² However, Xin et al.²² consider a very different prediction problem than ours that is based on using static MRI for prediction, and they do not consider the time-varying setup. These studies make use of either cross-sectional data²² or longitudinal data collected for less than 5 years,^{18–21} both much shorter compared to the CHS-CS, and do not consider the competing risk of death in their analysis. Novelty of our analysis lies in the fact that we utilize high-dimensional longitudinal data collected over 24 years for predicting progression of AD while also accounting for competing risk of death.

The rest of this paper is organized as follows. In Section 2, we introduce the multinomial fused lasso model and discuss numerous approaches for selecting the tuning parameters λ_1 , $\lambda_2 \geq 0$ that govern the strength of the lasso and fused lasso penalties, along with the stability of estimated coefficients, and related concepts. We present the analysis of the CHS-CS data set in Section 3 and a simulation study in Section 4. In Section 5, we conclude with some final comments and lay out ideas for future work. In Appendix A, we describe a proximal gradient descent algorithm for efficiently computing a solution to the multinomial fused lasso model, while in Appendix B, we provide details about the implementation of the algorithm for our application.

2 | THE MULTINOMIAL FUSED LASSO MODEL

Given the number of parameters involved in our general longitudinal setup, it will be helpful to be clear about notation (see Table 1). The matrix Y stores future outcome values, where the element Y_{it} records the outcome of the i th individual at time $t + \tau$, and $\tau \geq 0$ determines the time lag of the prediction. In the following, we will generally use the “ $\cdot$ ” symbol to denote partial indexing; examples are $X_{i,t}$, the vector of p predictors for individual i at time t , and $\beta_{t,k}$, the vector of p multinomial coefficients at time t and for class k . While the number of individuals n is assumed to be fixed over time, in Appendix B, we introduce an extension of the basic setup in which the number of individuals can vary across time points, with n_t denoting the number of individuals at each time point $t = 1, \dots, T$.

At each time point $t = 1, \dots, T$, we use a separate multinomial logit model for the outcome at time $t + \tau$.

$$\begin{aligned}
\log \frac{\mathbb{P}(Y_{it} = 1 | X_{i \cdot t} = x)}{\mathbb{P}(Y_{it} = K | X_{i \cdot t} = x)} &= \beta_{0t1} + \beta_{\cdot t1}^T x \\
\log \frac{\mathbb{P}(Y_{it} = 2 | X_{i \cdot t} = x)}{\mathbb{P}(Y_{it} = K | X_{i \cdot t} = x)} &= \beta_{0t2} + \beta_{\cdot t2}^T x \\
&\vdots \\
\log \frac{\mathbb{P}(Y_{it} = K - 1 | X_{i \cdot t} = x)}{\mathbb{P}(Y_{it} = K | X_{i \cdot t} = x)} &= \beta_{0t(K-1)} + \beta_{\cdot t(K-1)}^T x.
\end{aligned} \tag{1}$$

The coefficients are determined by maximizing a penalized log likelihood criterion

$$(\hat{\beta}_0, \hat{\beta}) \in \operatorname{argmax}_{\beta_0, \beta} \ell(\beta_0, \beta) - \lambda_1 P_1(\beta) - \lambda_2 P_2(\beta), \tag{2}$$

where $\ell(\beta_0, \beta)$ is the multinomial log likelihood

$$\ell(\beta_0, \beta) = \sum_{t=1}^T \sum_{i=1}^n \log \mathbb{P}(Y_{it} | X_{i \cdot t}),$$

P_1 is the lasso penalty²³

$$P_1(\beta) = \sum_{j=1}^p \sum_{t=1}^T \sum_{k=1}^{K-1} |\beta_{jtk}|,$$

and P_2 is a version of the fused lasso penalty²⁴ applied across time points

$$P_2(\beta) = \sum_{j=1}^p \sum_{t=1}^{T-1} \sum_{k=1}^{K-1} |\beta_{jtk} - \beta_{j(t+1)k}|.$$

(The element notation in (2) emphasizes the fact that the maximizing coefficients $(\hat{\beta}_0, \hat{\beta})$ need not be unique, since the log likelihood $\ell(\beta_0, \beta)$ need not be strictly concave, eg, this is the case when $p > n$.)

In broad terms, the lasso and fused lasso penalties encourage sparsity and persistence, respectively, in the estimated coefficients $\hat{\beta}$. A larger value of the tuning parameter $\lambda_1 = 0$ generally corresponds to fewer nonzero entries in $\hat{\beta}$; a larger value of the tuning parameter $\lambda_2 = 0$ generally corresponds to fewer change points in the piecewise constant coefficient trajectories $\hat{\beta}_{j \cdot k}$, across $t = 1, \dots, T$. We note that the form the log likelihood $\ell(\beta_0, \beta)$ specified above assumes independence between the outcomes across time points, which is a rather naive assumption given the longitudinal nature of our problem setup. However, this

naivety is partly compensated by the role of the fused lasso penalty, which ties together the multinomial models across time points, as we show later in the simulation study.

It helps to see an example. We consider a simple longitudinal problem with $n = 50$ individuals, $T = 15$ time points, and $K = 2$ classes. At each time point, we sampled $p = 30$ predictors independently from a standard normal distribution. The true (unobserved) coefficient matrix β is now 30×15 ; we set $\beta_j = 0$ for $j = 1 \dots, 27$, and set the three remaining coefficients trajectories to be piecewise constant across $t = 1, \dots, 15$, as shown in the left panel of Figure 1. In other words, the assumption here is that only 3 of the 30 variables are relevant for predicting the outcome, and these variables have piecewise constant effects over time. We generated a matrix of binary outcomes Y according to the multinomial model (1) and computed the multinomial fused lasso estimates $\hat{\beta}_0, \hat{\beta}$ in (2). The right panel of Figure 1 displays these estimates (all but the intercept $\hat{\beta}_0$) across $t = 1, \dots, 15$, for a favorable choice of tuning parameters $\lambda_1 = 2.5, \lambda_2 = 12.5$; the middle plot shows the unregularized (maximum likelihood) estimates corresponding to $\lambda_1 = \lambda_2 = 0$.

Each plot in Figure 1 has a y-axis that has been scaled to suit its own dynamic range. We can see that the multinomial fused lasso estimates, with an appropriate amount of regularization, pick up the underlying trend in the true coefficients, though the overall magnitude of coefficients is shrunk toward zero (an expected consequence of the ℓ_1 penalties). In comparison, the unregularized multinomial estimates are wild and do not convey the proper structure. From the perspective of prediction error, the multinomial fused lasso estimates offer a clear advantage, as well: over 30 repetitions from the same simulation setup, we used both the regularized coefficient estimates (with $\lambda_1 = 2.5$ and $\lambda_2 = 12.5$) and the unregularized estimates to predict the outcomes on an i.i.d. test set, with 0.5 as a prediction threshold. The average prediction error using the regularized estimates was 0.114 (with a standard error of 0.014), while the average prediction error from the unregularized estimates was 0.243 (with a standard error of 0.022).

2.1 | Implementation

We develop an algorithm based on proximal gradient descent for computing solutions of the fused lasso regularized multinomial regression problem in Equation (2). While a number of other algorithmic approaches are possible, such as implementations of the alternating direction method of multipliers,²⁵ we settle on the proximal gradient method because of its simplicity and because of the extremely efficient, direct proximal mapping associated with the fused lasso regularizer. We review proximal gradient descent in generality in Appendix A. In Appendix B, we provide details about the implementation of the algorithm for our problem and discuss a number of practical considerations like the choice of step size, stopping criteria, and a rescaling of the loss function that makes it roughly independent of the effective sample size, which might vary across time.

An efficient C++ implementation of the proximal gradient descent algorithm for the fused lasso regularized multinomial regression problem, with an easy interface to R, is available from the Github page: <https://github.com/SamAdhikari/longfused>.

2.2 | Model selection and evaluation

The selection of tuning parameters λ_1, λ_2 is clearly an important issue, as they are not known a priori. We discuss various methods for automatic tuning parameter selection in the multinomial fused lasso model (2). In particular, we consider the following methods for model selection: cross-validation, cross-validation under the one-standard-error rule, AIC, BIC, and, finally, AIC and BIC using misclassification loss (in place of the usual negative log likelihood). The cross-validation in our longitudinal setting is performed by dividing the individuals $1, \dots, n$ into folds, and, per its typical usage, selecting the tuning parameter pair λ_1, λ_2 (over, say, a grid of possible values) that minimizes the cross-validation misclassification loss. The one-standard-error rule, on the other hand, picks the simplest estimate that achieves a cross-validation misclassification loss within one standard error of the minimum. Here, “simplest” is interpreted to mean the estimate with the fewest number of nonzero component blocks. AIC and BIC scores are computed for a candidate λ_1, λ_2 pair by

$$\text{AIC}(\lambda_1, \lambda_2) = 2 \cdot \text{loss}\left((\hat{\beta}_0, \hat{\beta})_{\lambda_1, \lambda_2}\right) + 2 \cdot \text{df}\left((\hat{\beta}_0, \hat{\beta})_{\lambda_1, \lambda_2}\right),$$

$$\text{BIC}(\lambda_1, \lambda_2) = 2 \cdot \text{loss}\left((\hat{\beta}_0, \hat{\beta})_{\lambda_1, \lambda_2}\right) + \log N_{\text{tot}} \cdot \text{df}\left((\hat{\beta}_0, \hat{\beta})_{\lambda_1, \lambda_2}\right),$$

and in each case, the tuning parameter pair is chosen (again, say, over a grid of possible values) to minimize the score. In the above, $(\hat{\beta}_0, \hat{\beta})_{\lambda_1, \lambda_2}$ denotes the multinomial fused lasso

estimate (2) at the tuning parameter pair λ_1, λ_2 , and N_{tot} denotes the total number of observations in the longitudinal study, $N_{\text{tot}} = nT$ (or $N_{\text{tot}} = \sum_{t=1}^T n_t$ in the missing data setting). Also, $\text{df}((\hat{\beta}_0, \hat{\beta})_{\lambda_1, \lambda_2})$ denotes the degrees of freedom of the estimate $(\hat{\beta}_0, \hat{\beta})_{\lambda_1, \lambda_2}$, and we employ the approximation

$$\text{df}((\hat{\beta}_0, \hat{\beta})_{\lambda_1, \lambda_2}) \approx \# \text{ of nonzero blocks in } (\hat{\beta}_0, \hat{\beta})_{\lambda_1, \lambda_2},$$

borrowing from known results in the Gaussian likelihood case.^{26,27} Finally, $\text{loss}((\hat{\beta}_0, \hat{\beta})_{\lambda_1, \lambda_2})$

denotes a loss function considered for the estimate, which we take either to be the negative multinomial log likelihood $-\ell((\hat{\beta}_0, \hat{\beta})_{\lambda_1, \lambda_2})$, as is typical in AIC and BIC,²⁸ or the

misclassification loss, to put it on closer footing to cross-validation. Note that both loss functions are computed in-sample, ie, over the training samples, and hence, AIC and BIC are computationally much cheaper than cross-validation. The six model selection methods will be compared using a K-fold nested cross-validation scheme, where the best tuning parameters λ_1 and λ_2 are selected within the training data according to the model selection criteria discussed here.

For each observation in the test data, we predict outcome based on the predicted probabilities for the K different classes from the multinomial logit model in (1). An outcome class corresponding to the largest predicted probability is the predicted outcome for the observation. We compare prediction performance of different model selection methods by considering four different evaluation metrics: misclassification error rate, true positive rate, false positive rate, and positive predictive value. Misclassification error rate is the proportion of total observations incorrectly classified to a wrong outcome class. True positive rate for class k is the ratio of number of outcomes correctly predicted as class k and the number of outcomes observed in that class. False positive rate for a class k is the ratio of number of outcomes incorrectly predicted as class k and the number of observed outcomes that are not in that class. Finally, positive predictive value for class k is the ratio of number of outcomes correctly predicted as class k and the total number of outcomes predicted in that class. Ideally, we seek to find a selection method that maximizes the true positive rate and the positive predictive value while minimizing the misclassification error and the false positive rate.

2.3 | Measures of stability

Examining the stability of variables in a fitted model, subject to small perturbations of the data set, is one way to assess variable importance. Applications of stability, in this spirit, have recently gained popularity in the literature, across a variety of settings such as clustering (eg, the work of Lange et al²⁹), regression (eg, the work of Meinshausen and Bühlmann³⁰), and graphical models (eg, the work of Liu et al³¹). In this paper, we propose a very simple stability-based measure of variable importance, based on the definition of variable importance for trees and additive tree expansions.^{28,32} We fit the multinomial fused lasso estimate (2) on the data set $X_{i\cdot}, Y_i$, for $i = i_1, \dots, i_m$, a subsample of the total individuals $1, \dots, n$, and repeat this process R times. Let $\hat{\beta}^{(r)}$ denote the coefficients from the r th subsampled data set, for $r = 1, \dots, R$. Then, we define the importance of variable j for class k as

$$I_{jk} = \frac{1}{RT} \sum_{r=1}^R \sum_{t=1}^T |\hat{\beta}_{jtk}^{(r)}|, \quad (3)$$

for each $j = 1, \dots, p$ and $k = 1, \dots, K - 1$, which is the average absolute magnitude of the coefficients for the j th variable and k th class, across all time points, and subsampled data sets. Therefore, a larger value of I_{jk} indicates a higher variable importance, as measured by stability (not only across subsampled data sets r , but actually across time points t , as well). Relative importances can be computed by scaling the highest variable importance to be 100, and adjusting the other values accordingly; for simplicity, we typically consider relative variable importances in favor of absolute ones, because the original scale has no real meaning.

There is some subtlety in the role of the tuning parameters λ_1, λ_2 used to fit the coefficients $\hat{\beta}^{(r)}$ on each subsampled data set $r = 1, \dots, R$. Note that the importance measure (3) reflects the importance of a variable in the context of a fitting procedure that, given data samples,

produces estimated coefficients. The simplest approach would be to consider the fitting procedure defined by the multinomial fused lasso problem (2) at a fixed pair of tuning parameter values λ_1, λ_2 . But, in practice, it is seldom true that appropriate tuning parameter values are known ahead of time, and one typically employs a method like cross-validation to select parameter values. Hence, in this case, to determine variable importances in the final coefficient estimates, we would take care to define our fitting procedure in (3) to be the one that, given data samples, performs cross-validation on these data samples to determine the best choice of λ_1, λ_2 , and then uses this choice to fit coefficient estimates. In other words, for each subsampled data set $r = 1, \dots, R$ in (3), we would perform cross-validation to determine tuning parameter values and then compute $\hat{\beta}^{(r)}$ as the multinomial fused lasso solution at these chosen parameter values. This is more computationally demanding, but it is a more accurate reflection of variable importance in the final model output by the multinomial fused lasso under cross-validation for model selection.

2.4 | Alternative approach

We are mainly interested in the prediction of dementia; this task is complicated by the fact that risk factors for dementia are also known to be risk factors for death,¹⁷ and so to account for this, we include the death category in the multinomial classification model. An alternate approach would be to use a Cox proportional hazards model,³³ where the event of interest is the onset of dementia, and censorship corresponds to death.

Traditionally, the Cox model is not fit with time-varying predictors or time-varying coefficients, but it can be naturally extended to the setting considered in this work, even using the same regularization schemes. Instead of the multinomial model (1), we would model the hazard function as

$$h(t + \Delta | X_{i:t} = x) = h_0(t + \Delta) \cdot \exp(x^T \beta_{:t}), \quad (4)$$

where $\beta \in \mathbb{R}^{p \times T}$ are a set of coefficients over time, and h_0 is some baseline hazard function (that does not depend on predictor measurements). Note that the hazard model (4) relates the instantaneous rate of failure (onset of dementia) at time $t + \Delta$ to the predictor measurements at time t . This is as in the multinomial model (1), which relates the outcomes at time $t + \Delta$ (dementia or death) to predictor measurements at time t . The coefficients in (4) would be determined by maximizing the partial log likelihood with the analogous lasso and fused lasso penalties on β , as in the above multinomial setting (2).

The partial likelihood approach can be viewed as a sequence of conditional log odds models,^{34,35} and therefore, one might expect the (penalized) Cox regression model described here to perform similarly to the (penalized) multinomial regression model pursued in this paper. In fact, the computational routine described in Appendix B would apply to the Cox model with only very minor modifications (that concern the gradient computations). A rigorous comparison of the two approaches is beyond the scope of the current manuscript but is an interesting topic for future development.

3 | ALZHEIMER'S DISEASE DATA ANALYSIS

3.1 | Data

We use data from the $n = 924$ individuals in the Pittsburgh section of the CHS-CS, recorded between 1990 and 2012. Each individual underwent clinical and cognitive assessments at multiple ages, all falling in the range 65, $\dots$, 108. The matrix of (future) outcomes Y has dimension $n \times 34$; for $i = 1, \dots, 924$ and $t = 65, \dots, 98$, the outcome Y_{it} stores the cognitive status at age $t + 10$ and can assume one of the following values:

$$Y_{it} = \begin{cases} 1, & \text{if normal} \\ 2, & \text{if MCI/dementia} \\ 3, & \text{if dead,} \end{cases}$$

where 11% of the overall cases are in outcome class normal, 21% in class dementia, and 68% in class dead. MCI is included in the same class as dementia, as they are both instances of cognitive impairment. Hence, the proposed multinomial model predicts the onset of MCI/dementia, in the presence of a separate death category. This is done to implicitly adjust for the confounding effect of death, as some risk factors for dementia are also known to be risk factors for death.¹⁷

The number n_t of outcomes observed at age t varies across time, for two reasons: first, different subjects entered the study at different ages, and second, once a subject dies at time t_0 , we exclude them from consideration in the model formed at all ages $t > t_0$, to predict the outcomes of individuals at age $t + 10$. Figure 2 shows the proportion of the three outcome categories at each age, along with total number of observed outcomes at the age. The maximum number of outcomes is 604 at age 88, whereas the minimum is 7 at age 108, with 10 504 total outcomes measured in at least one time point. We resort to the strategy described in Appendix B.2.3 and use the scaled loss in (B8) to compensate for the varying sample sizes.

The array of predictors X is composed of time-varying variables that were recorded at least twice during the CHS-CS study, and time-invariant variables, such as gender and race. Predictors collected in the Pittsburgh chapter of CHS-CS are described in detail in the works of Lopez et al.^{12,14} A complication in the data set is the ample amount of missingness in the array of predictors. We impute missing values using a uniform rule for all possible causes of missingness. A missing value at age t is imputed by taking the closest past measurement from the same individual, if present. If all the past values are missing, the global median from people of age t is used. The only exception is the case of time-invariant predictors, whose missing values are imputed by either future or past values, as available.

Categorical variables with m possible outcomes are converted to $m - 1$ binary variables and all the predictors are standardized to have zero mean and unit standard deviation. This is a standard procedure in regularization, as the lasso and fused lasso penalties puts constraints on the size of the coefficients associated with each variable.³⁶ To be precise, imputation of missing values and standardization of the predictors are performed within each of the folds used in the cross-validation method for the choice of the tuning parameters λ_1 and λ_2

(discussed in the following) and then again for the full data set in the final estimation procedure that uses the selected tuning parameters. The final array of predictors X has dimension $n_t \times 1050 \times 34$, where 1050 is the number of variables recorded over the period of 34 years of age range and n_t is the number of observed outcome at age t .

Additional information about the CHS-CS data can be found in the website <https://chs-nhlbi.org/>. The data can be accessed by submitting a data request form to National Heart, Lung, and Blood Institute.

3.2 | Selection of tuning parameters λ_1 and λ_2

We compare different model selection methods on the subset of the CHS-CS data set with 140 predictors and 600 randomly selected individuals, as an illustration. The individuals are randomly split into five folds. We use 4/5 of the data set to perform model selection and subsequent model fitting with the six techniques described above: cross-validation, cross-validation with the one-standard-error rule, and AIC and BIC under negative log likelihood and misclassification losses. To be perfectly clear, the model selection techniques work entirely within this given 4/5 of the data set, so that, eg, cross-validation further divides this data set into folds. In fact, we used fourfold cross-validation to make this division simplest. The remaining 1/5 of the data set is then used for evaluation of the estimates coming from each of the six methods, and this entire process is repeated, leaving out each fold in turn as the evaluation set. We record several measures on each evaluation set: the misclassification rate, true positive rate in identifying the dementia class, true positive rate in identifying the dementia and death classes combined, and degrees of freedom (number of nonzero blocks in the estimate). Figure 3 displays the mean and standard errors of these four measures, for each of the six model selection methods.

We further investigated the out-of-sample mean true positive rate and mean positive predictive value in identifying dementia at different ages for the six different model selection methods considered, as shown in Figure 4.

Cross-validation and cross-validation with the one-standard-error rule both seem to represent a favorable balance between the different evaluation measures. The cross-validation methods provide a misclassification rate significantly better than that of the null model, which predicts according to the majority class (death), they yield two of the three highest true positive rates in identifying the dementia class and perform well in terms of identifying the dementia and death classes combined (as do all methods: note that all true positive rates here are about 0.75 or higher). The cross-validation under the usual rule also had the best performance in terms of true positive rate and positive predictive value over age, with highest values for predictors between ages 70 and 80 to predict dementia between ages 80 and 90.

We ended up settling for cross-validation under the usual rule, rather than the one-standard-error rule, because the former achieves the highest true positive rate in identifying the dementia class, which was our primary concern in the CHS-CS data analysis. By design, cross-validation with the one-standard-error rule delivers a simpler estimate in terms of degrees of freedom (196 for the one-standard-error rule versus 388 for the usual rule) though

both cross-validation models are highly regularized in absolute terms (eg, the fully saturated model would have thousands of nonzero blocks).

3.3 | Model and algorithm specification

In the AD application, the multinomial model in (1) is determined by two equations, as there are three possible outcomes (normal, MCI/dementia, death); the outcome “normal” is taken as the base class. We will refer to the two equations (and the corresponding sets of coefficients) as the “dementia vs normal” and “death vs normal” equations, respectively.

We use the proximal gradient descent algorithm described in Appendix B to estimate the coefficients that maximize the penalized log likelihood criterion in (2). The initializations $(\beta_0^{(0)}, \beta^{(0)})$ are set to be zero matrices, the maximum number of iterations is $S = 80$, the initial step size before backtracking is $\tau_0 = 20$, the backtracking shrinkage parameter is $\gamma = 0.6$ and the tolerance of the first stopping criterion (relative difference in function values) is $\epsilon = 0.001$. We select the tuning parameters by a fourfold cross-validation procedure that minimizes the misclassification error. The selected parameters are $\lambda_1 = 0.019$ and $\lambda_2 = 0.072$, which yield an average prediction error of 0.312 (standard error 0.009).

The relative variable importances for the CHS-CS data set are computed using four subsampled data sets, each one containing 75% of the total number of individuals. The tuning parameter values have been selected by cross-validation. The variable importances were defined to incorporate this selection step into the fitting procedure, as explained in Section 2.3.

3.4 | Results

Out of the 1050 coefficients associated with the predictors described above, 148 are estimated to be nonzero for at least one time point in the 34 years age range. More precisely, for at least one age, 57 coefficients are nonzero in the “dementia versus normal” equation of the predictive multinomial logit model, and 124 are nonzero in the “death versus normal” equation.

For interpreting the results, we focus on the 15 most important predictors in the 34 years age range, separately for the two equations modeling “dementia versus normal” and “death versus normal,” respectively. The meaning of these predictors and the coding used for the categorical variables are reported in Table 2. The measure of importance is described in detail in Section 2.3 and is, in fact, a measure of stability of the estimated coefficients, across four subsets of the data (the four training sets used in cross-validation).

In Figure 5, we show the plot of relative importance and the corresponding coefficients of the important predictors for the two equations. The plots on the left show the relative importance of the 15 variables with respect to the most important one, whose importance was scaled to be 100. The plots on the right show, separately for the two equations, the longitudinal estimated coefficients for the 15 most important variables, using the data and algorithm specification described above. The nonzero coefficients that are not displayed in

Figure 5 are less important (according to our measure of stability) and, for the vast majority, their absolute values are less than 0.1.

We now proceed to interpret the results, keeping in mind that, ultimately, we are estimating the coefficients of a multinomial logit model and that the outcome variable is recorded 10 years in the future with respect to the predictors. For example, an increase in the value of a predictor with positive estimated coefficient in the top right plot of Figure 5 is associated with an increase of the (10 years future) odds of dementia with respect to a normal cognitive status. In what follows, to facilitate the exposition of results, our statements are less formal.

Inspecting the “dementia vs normal” plot we see that, in general, being Caucasian (Race white) is an important predictor of decrease in the odds of dementia, while, after the age of 85, fear (Fearful), lack of available caretakers (Find help), and deterioration of learning skills (Learn new) are important predictors of the increasing odds of dementia. Variables Diuretics (a particular diuretic) and Wake early (early wake-ups) have positive coefficients for the age ranges 65, ..., 78 and 77, ..., 91, respectively, and hence, if active, they predict the increasing risk of dementia. The “death vs normal” plot reveals the importance of several variables in the age range 65, ..., 85: longer time to rise from sitting in a chair (Chair stand), more cigarettes (Cigarettes), higher cardiac injury score (Cardiac injury) are important predictors of increasing risk of death. Other variables in the same age range, with analogous interpretations, but lower importance, are Diabetes (“new diabetes” diagnosis), Health (“fair” health status), Digitalis (use of Digitalis), Walking (difficulty in walking). By contrast, in the same age range, good performance on the digit-symbol substitution task (Digits correct) will predict decrease in the odds of death. Finally, regardless of the age, being around nonsmokers (Any smoker) or being a woman (Female) are good predictors of decrease in the odds of death. Note that caution is needed while interpreting the changes in log-odds over time, which could be due to a selection bias as discussed in the work of Hernán.³⁷ Of course, this bias will have more severe impact for a causal interpretation that is not the focus of our work.

Figure 6 shows the intercept coefficients $\hat{\beta}_{0.1}$ and $\hat{\beta}_{0.2}$, which, we recall, are not penalized in the log likelihood criterion in (2). The intercepts account for time-varying risk that is not explained by the predictors. In particular, the coefficients $\hat{\beta}_{0.2}$ increases over time, suggesting that an increasing amount of risk of death can be attributed to a subject’s age alone, independent of the predictor measurements. However, as we explain later, the risk factors of dementia become more relevant in later years, suggesting that the age alone does not explain the risk of dementia after age of 85.

The results of the proposed multinomial fused lasso methodology applied to the CHS data are broadly consistent with what is known about risk and protective factors for dementia in the elderly.³⁸ Race, gender, vascular and heart disease, lack of available caregivers, and deterioration of learning and memory are all associated with an increased risk of dementia. The results, however, provide critical new insights into the natural progression of MCI/dementia. First, the relative importance of the risk factors changes over time. As shown in Figure 5, with the exception of race, risk factors for dementia become more relevant after the

age of 85. This is critical, as there is increasing evidence³⁹ for a change in the risk profile for the expression of clinical dementia among the oldest-old. Second, the independent prediction of death and the associated risk/protection factors highlight the close connection between risk of death and risk of dementia. That is, performance on a simple, timed test of psychomotor speed (digit symbol substitution task) is a very powerful predictor of death within 10 years, as is a measure of physical strength/frailty (time to arise from a chair). Other variables, including gender, diabetes, walking, and exercise, are all predictors of death, but are known, from other analyses in the CHS and other studies, to be linked to the risk of dementia. The importance of these risk/protective factors for death is attenuated (with the exception of gender) after age 85, likely reflecting survivor bias. Taken together, these results add to the growing body of evidence of the critical importance of accounting for mortality in the analysis of risk for dementia, especially among the oldest old.³⁹

We further observe that the variables with high positive or negative coefficients for most ages in the plotted trajectories typically also have among the highest relative importances. Another interesting observation concerns categorical predictors, which (recall) have been converted into binary predictors over multiple levels: often only some levels of a categorical predictor are active in the plotted trajectories.

For our analysis, we chose a 10-year time window for risk prediction. The clinical motivation for the time window comes from the evidence in literature that the “biological” processes of neurodegeneration begins at least 10 years (if not earlier) before the “clinical” expression as MCI or dementia.^{8,40} Among individuals of age 65–75, who are cognitively normal, this is a scientifically and clinically reasonable time window to use. However, had we similar data from individuals as young as 45–50 years old, then we might wish to choose time windows of 20 years or longer. It could also be argued that a shorter time window might be more scientifically and clinically relevant among the oldest-old individuals for whom survival times of 10 years become increasingly less likely. Since we are starting at around 65 years old and using a uniform prediction window for all ages in our analysis, 10-year window is a reasonable compromise to examine prediction in a clinically meaningful way. However, the methodology introduced in this paper is in no way tied to this one prediction window.

4 | SIMULATION STUDY

We compare prediction performance of the proposed model with other competing models using simulated data. Data are simulated to emulate distributions of the observed covariates and outcome, for individuals with no missing outcome, in the CHS-CS data. We randomly chose 300 predictors, both time varying and invariant, from the observed data and simulated a matrix of predictors X_{sim}^t , at each time t for $t = 1, \dots, 34$. Continuous predictors, which are all positive, are simulated from a truncated Normal distribution, such that, for each continuous covariate j and time t ,

$$X_{\text{sim}}^{j,t} \sim \text{Truncated Normal}(\mu_{j,t}, \sigma_{j,t}^2, a_{j,t}, b_{j,t}).$$

The mean $\mu_{j,t}$ and the variance $\sigma_{j,t}^2$ are both estimated from empirical mean and variance in the observed data for covariate j at time t . The lower limit $a_{j,t}$ and the upper limit $b_{j,t}$ are estimated as minimum and maximum in the observed data. Binary predictors are simulated using Bernoulli distribution, such that, for each binary covariate l and time t ,

$$X_{\text{sim}}^{l,t} \sim \text{Bernoulli}(p_{l,t}),$$

where $p_{l,t}$ is the observed proportion of events for covariate l in time t . Categorical covariates with more than two categories are simulated from Multinomial distribution such that the probability of each category is again estimated using the proportion in the observed data. The outcome for patient i at time t , denoted as $Y_{i,\text{sim}}^t$, is then generated using a multinomial logistic model in Equation (1) with three categories. The matrix of true coefficient β is sparse and stepwise constant over time, with dimension $300 \times 34 \times 2$. The true degrees of freedom (number of nonzero blocks) β of is 94 with 20 relevant predictors, ie, these predictors had nonzero coefficient for at least one time point. In the simulated data, 24% of the patients are in class dementia, 58% are in class death, and 18% are in class normal. The proportion of different outcome categories in the simulated data is comparable to those in the observed data.

In addition to the proposed multinomial logistic regression with lasso and fused lasso penalties, we consider, multinomial logistic regression without any penalty, multinomial logistic regression with only lasso penalty, and two binomial logistic regressions, with lasso and fused lasso penalties, for comparison. In the first binomial regression model, we ignore patients who died at time $t + 1$ in training data. This logistic regression is fitted assuming that patients can either be in normal or in dementia category only. In the second binomial regression, patients in outcome categories normal and dead are pooled to form a single category, and the model is fitted to predict whether a patient will be in class dementia in time $t + 1$ versus no dementia. The two binomial regression models do not account for the competing risk of death.

Prediction performance of the five different models is compared using nested cross-validation. The internal fourfold cross-validation is used for the selection of tuning parameters that minimize the cross-validation error, whereas the external fivefold cross-validation is used to compare out-of-sample prediction for each model using the best tuning parameter. Fitting multinomial logistic regression without any penalty is equivalent to fitting the proposed model with $\lambda_1 = \lambda_2 = 0$. Similarly, for multinomial logistic regression with only lasso penalty, we fix λ_2 at zero and use cross-validation to select λ_1 .

Misclassification error, true positive rate, false positive rate, and estimated degrees of freedom are used to compare the fits from different prediction model. Since correct prediction of dementia is the primary interest of the analysis, we focus on true positive rate and false positive rate for predicting dementia only. For comparability among different models with varying number of outcome categories (two categories in binomial models and three in multinomial models), we standardize the degrees of freedom by the number of

outcome categories K in the fitted model. Finally, recall that the cases in class dead were not included in the first binomial logistic regression model for fitting normal versus dementia. However, to reflect real world scenario as closely as possible, we do not make such consideration during prediction in test data. Instead, we use the predictors for the patients in all three outcome categories to predict normal or dementia outcome and use the predicted outcome to compute evaluation metrics. This will help us identify whether and how often patients are incorrectly classified as dementia, when category dead is not used during training. In Figure 7, we show the evaluation metrics for each model computed using fivefold cross-validation.

We also compute importance measures for each covariate, following the procedure discussed in the data analysis (Section 3.2), and we compare covariates with high relative measure of importance to true important covariates. For each model, we rank the predictors using their relative importance and select the top 20 important predictors. We then compute the proportion of the true top 20 important predictors that were selected as top 20 important predictors in the fitted model. The proportions for the five models and the two outcome categories are displayed in Table 3.

Our simulation results show that the proposed multinomial logistic regression with lasso and fused lasso penalty performs better compared to other competing models, in a longitudinal setting, when dimension of predictor space is large. The proposed model has a relatively high true positive rate for predicting dementia with very low false positive rate and relatively small degrees of freedom. Our simulation results further justify to consider death as a competing outcome, rather than either completely ignoring it (which resulted in high false positive rate) or merging it with normal category (resulting with low true positive rate). Finally, the covariates with true nonzero coefficients were also the ones with high relative importance in our proposed model, compared to other competing models.

5 | DISCUSSION AND FUTURE WORK

In this work, we proposed a multinomial model for high-dimensional longitudinal classification tasks. Our proposal operates under the assumption that a sparse number of predictors contribute more or less persistent effects across time. The multinomial model is fit under lasso and fused lasso regularizations, which address the assumptions of sparsity and persistence, respectively, and lead to piecewise constant estimated coefficient profiles. We described a highly efficient computational algorithm for this model based on proximal gradient descent, demonstrated the applicability of this model on an Alzheimer's data set taken from the CHS-CS, and discussed practically important issues such as stability measures for the estimates and tuning parameter selection. Our proposed model has relevance to many applications in medical research beyond the one considered in this paper, since use of high-dimensional data source, such as electronic health record, is gaining popularity in those applications.

A number of extensions of the basic model are well within reach. For example, placing a group lasso penalty on the coefficients associated with each level of a binary expansion for a categorical variable may be useful for encouraging sparsity in a group sense (ie, over all

levels of a categorical variable at once). As another example, more complex trends than piecewise constant ones may be fit by replacing the fused lasso penalty with a trend filtering penalty,^{41,42} which would lead to piecewise polynomial trends of any chosen order k . The appropriateness of such a penalty would depend on the scientific application; the use of a fused lasso penalty assumes that the effect of a given variable is mostly constant across time, with possible change points; the use of a quadratic trend filtering penalty (polynomial order $k = 2$) allows the effect to vary more smoothly across time.

More difficult and open-ended extensions concern statistical inference for the fitted longitudinal classification models. For example, the construction of confidence intervals (or bands) for selected coefficients (or coefficient profiles) would be an extremely useful tool for the practitioner and would offer more concrete and rigorous interpretations than the stability measures described in Section 2.3. Unfortunately, this is quite a difficult problem, even for simpler regularization schemes (such as a pure lasso penalty) and simpler observation models (such as linear regression). But, recent inferential developments for related high-dimensional estimation tasks^{43–48} shed a positive light on this future endeavor.

ACKNOWLEDGEMENTS

This research was supported through contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, and N01HC85086 and through grants U01HL080295 and U01HL130114 from the National Heart, Lung, and Blood Institute, with additional contribution from the National Institute of Neurological Disorders and Stroke. Additional support was provided through grant R01AG023629 from the National Institute on Aging. A full list of principal CHS investigators and institutions can be found at CHS-NHLBI.org.

Funding information

National Heart, Lung, and Blood Institute, Grant/Award Number: HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, U01HL080295, and U01HL130114; National Institute of Neurological Disorders and Stroke; National Institute on Aging, Grant/Award Number: R01AG023629

APPENDIX A: A PROXIMAL GRADIENT DESCENT APPROACH

Suppose that $g: \mathbb{R}^d \rightarrow \mathbb{R}$ is convex and differentiable, $h: \mathbb{R}^d \rightarrow \mathbb{R}$ is convex, and we are interested in computing a solution

$$x^* \in \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} g(x) + h(x).$$

If h were assumed differentiable, then the criterion $f(x) = g(x) + h(x)$ is convex and differentiable, and repeating the simple gradient descent steps

$$x^+ = x - \tau \nabla f(x) \quad (\text{A1})$$

suffices to minimize f , for an appropriate choice of step size τ . (In the above, we write x^+ to denote the gradient descent update from the current iterate x .) If h is not differentiable, then gradient descent obviously does not apply, but as long as h is “simple” (to be made precise

shortly), we can apply a variant of gradient descent that shares many of its properties, called *proximal gradient descent*. Proximal gradient descent is often also called composite or generalized gradient descent, and in this routine, we repeat the steps

$$x^+ = \text{prox}_{h, \tau}(x - \tau \nabla g(x)) \quad (\text{A2})$$

until convergence, where $\text{prox}_{h, \tau}: \mathbb{R}^d \rightarrow \mathbb{R}^d$ is the proximal mapping associated with h (and τ),

$$\text{prox}_{h, \tau}(x) = \underset{z \in \mathbb{R}^d}{\text{argmin}} \frac{1}{2\tau} \|x - z\|_2^2 + h(z). \quad (\text{A3})$$

(Strict convexity of the above criterion ensures that it has a unique minimizer, so that the proximal mapping is well defined.) Provided that h is simple, by which we mean that its proximal map (A3) is explicitly computable, the proximal gradient descent steps (A2) are straightforward and resemble the classical gradient descent analogues (A1); we simply take a gradient step in the direction governed by the smooth part g , and then apply the proximal map of h . A slightly more formal perspective argues that the updates (A3) are the result of minimizing h plus a quadratic expansion of g , around the current iterate x .

Proximal gradient descent has become a very popular tool for optimization problems in statistics and machine learning, where, typically, g represents a smooth loss function and h a nonsmooth regularizer. This trend is somewhat recent, even though the study of proximal mappings has a long history of in the optimization community (eg, see the work of Parikh and Boyd⁴⁹ for a nice review paper). In terms of convergence properties, proximal gradient descent enjoys essentially the same convergence rates as gradient descent under the analogous assumptions and is amenable to acceleration techniques just like gradient descent (eg, the works of Nesterov⁵⁰ and Beck and Teboulle⁵¹). Of course, for proximal gradient descent to be applicable in practice, one must be able to exactly (or even approximately) compute the proximal map of h in (A3); fortunately, this is possible for many optimization problems, ie, many common regularizers h , that are encountered in statistics. In our case, the proximal mapping reduces to solving a problem of the form

$$\hat{\theta} = \underset{\theta}{\text{argmin}} \frac{1}{2} \|x - \theta\|_2^2 + \lambda_1 \sum_{i=1}^m |\theta_i| + \lambda_2 \sum_{i=1}^{m-1} |\theta_i - \theta_{i+1}|. \quad (\text{A4})$$

This is often called the fused lasso signal approximator (FLSA) problem, and extremely fast, linear-time algorithms exist to compute its solution. In particular, we rely on an elegant dynamic programming approach proposed by Johnson.⁵²

APPENDIX B: PROXIMAL GRADIENT DESCENT ALGORITHM

We describe the implementation of proximal gradient descent algorithm for our problem and a number of practical considerations like the choice of step size and stopping criterion as follows.

B.1 | Application to the multinomial fused lasso problem

The problem in (2) fits into the desired form for proximal gradient descent, with g the multinomial regression loss (ie, negative multinomial regression log likelihood) and h the lasso plus fused lasso penalties. Formally, we can rewrite (2) as

$$(\hat{\beta}_0, \hat{\beta}) \in \underset{\beta_0, \beta}{\operatorname{argmin}} g(\beta_0, \beta) + h(\beta_0, \beta), \quad (\text{B1})$$

where g is the convex, smooth function

$$g(\beta_0, \beta) = \sum_{t=1}^T \sum_{i=1}^n \left\{ \sum_{k=1}^{K-1} -\mathbb{I}(Y_{it} = k)(\beta_{0tk} + X_{i \cdot t} \beta_{\cdot tk}) + \log \left(1 + \sum_{h=1}^{K-1} \exp(\beta_{0th} + X_{i \cdot t} \beta_{\cdot th}) \right) \right\},$$

and h is the convex, nonsmooth function

$$h(\beta_0, \beta) = \lambda_1 \sum_{j=1}^p \sum_{t=1}^T \sum_{k=1}^{K-1} |\beta_{jtk}| + \lambda_2 \sum_{j=1}^p \sum_{t=1}^{T-1} \sum_{k=1}^{K-1} |\beta_{jtk} - \beta_{j(t+1)k}|.$$

Here, we consider fixed values $\lambda_1, \lambda_2 \geq 0$. As described previously, each of these tuning parameters will have a big influence on the strength of their respective penalty terms and hence the properties of the computed estimate $(\hat{\beta}_0, \hat{\beta})$; we discuss the selection of λ_1 and λ_2 in Section 3.2. We note that the intercept coefficients β_0 are not penalized.

To compute the proximal gradient updates, as given in (A2), we must consider two quantities: the gradient of g and the proximal map of h . First, we discuss the gradient. As $\beta_0 \in \mathbb{R}^{T \times (K-1)}$, $\beta \in \mathbb{R}^{p \times T \times (K-1)}$, we may consider the gradient as having dimension $\nabla g(\beta_0, \beta) \in \mathbb{R}^{(p+1) \times T \times (K-1)}$. We will index this as $[\nabla g(\beta_0, \beta)]_{jtk}$ for $j = 0, \dots, p$, $t = 1, \dots, T$, $k = 1, \dots, K-1$; hence, note that $[\nabla g(\beta_0, \beta)]_{0tk}$ gives the partial derivative of g with respect to β_{0tk} , and $[\nabla g(\beta_0, \beta)]_{jtk}$ the partial derivative with respect to β_{jtk} , for $j = 1, \dots, p$. For generic t, k , we have

$$[\nabla g(\beta_0, \beta)]_{0tk} = \sum_{i=1}^n \left(-\mathbb{I}(Y_{it} = k) + \frac{\exp(\beta_{0tk} + X_{i \cdot t} \beta_{\cdot tk})}{1 + \sum_{h=1}^{K-1} \exp(\beta_{0th} + X_{i \cdot t} \beta_{\cdot th})} \right), \quad (\text{B2})$$

and for $j = 1, \dots, p$,

$$[\nabla g(\beta_0, \beta)]_{jtk} = \sum_{i=1}^n \left(-\mathbb{I}(Y_{it} = k)X_{ijt} + X_{ijt} \frac{\exp(\beta_{0tk} + X_{i \cdot t} \beta_{\cdot tk})}{1 + \sum_{h=1}^{K-1} \exp(\beta_{0th} + X_{i \cdot t} \beta_{\cdot th})} \right). \quad (\text{B3})$$

It is evident that computation of the gradient requires $\mathcal{O}(npTK)$.

Now, we discuss the proximal operator. Since the intercept coefficients $\beta_0 \in \mathbb{R}^{T \times (K-1)}$ are left unpenalized, the proximal map over β_0 just reduces to the identity, and the intercept terms undergo the updates

$$\beta_{0tk}^+ = \beta_{0tk} - \tau [\nabla g(\beta_0, \beta)]_{0tk} \quad \text{for } t = 1, \dots, T, k = 1, \dots, K-1.$$

Hence, we consider the proximal map over β alone. At an arbitrary input $x \in \mathbb{R}^{p \times T \times (K-1)}$, this is

$$\begin{aligned} \underset{z \in \mathbb{R}^{p \times T \times (K-1)}}{\text{argmin}} \quad & \frac{1}{2\tau} \sum_{j=1}^p \sum_{t=1}^T \sum_{k=1}^{K-1} (x_{jtk} - \beta_{jtk})^2 + \lambda_1 \sum_{j=1}^p \sum_{t=1}^T \sum_{k=1}^{K-1} |\beta_{jtk}| \\ & + \lambda_2 \sum_{j=1}^p \sum_{t=1}^{T-1} \sum_{k=1}^{K-1} |\beta_{jtk} - \beta_{j(t+1)k}|, \end{aligned}$$

which we can see decouples into $p(K-1)$ separate minimizations, one for each predictor $j = 1, \dots, p$ and class $k = 1, \dots, K-1$. In other words, the coefficients β undergo the updates

$$\begin{aligned} \beta_{j \cdot k}^+ = \underset{\theta \in \mathbb{R}^T}{\text{argmin}} \quad & \frac{1}{2} \sum_{t=1}^T \left((\beta_{j \cdot k} - \tau [\nabla g(\beta_0, \beta)]_{j \cdot k}) - \theta \right)^2 + \tau \lambda_1 \sum_{t=1}^T |\theta_t| + \tau \lambda_2 \sum_{t=1}^{T-1} |\theta_t - \theta_{t+1}|, \\ & \text{for } j = 1, \dots, p, k = 1, \dots, K-1, \end{aligned}$$

(B4)

each minimization being an FLSA problem,²⁴ ie, of the form (A4). There are many computational approaches that may be applied to such a problem structure; we employ a specialized, highly efficient algorithm by Johnson⁵² that is based on dynamic programming. This algorithm requires $\mathcal{O}(T)$ operations for each of the problems in (B4), making the total cost of the update $\mathcal{O}(pTK)$ operations. Note that this is actually dwarfed by the cost of computing the gradient $\nabla g(\beta, \beta)$ in the first place, and therefore, the total complexity of a single iteration of our proposed proximal gradient descent algorithm is $\mathcal{O}(npTK)$.

B.2 | Practical considerations

We discuss several practical issues that arise in applying the proximal gradient descent algorithm.

B.2.1 | Backtracking line search—Returning to the generic perspective for proximal gradient descent as described in Appendix A, we rewrite the proximal gradient descent update in (A2) as

$$x^+ = x - \tau G_\tau(x), \quad (\text{B5})$$

where $G(x)$ is called the *generalized gradient* and is defined as

$$G_\tau = \frac{x - \text{prox}_{h, \tau}(x - \tau \nabla g(x))}{\tau}.$$

The update is rewritten in this way so that it more closely resembles the usual gradient update in (A1). We can see that, analogous to the gradient descent case, the choice of parameter $\tau > 0$ in each iteration of proximal gradient descent determines the magnitude of the update in the direction of the generalized gradient $G_\tau(x)$. Classical analysis shows that, if ∇g is Lipschitz with constant $L > 0$, then proximal gradient descent converges with any fixed choice of step size $\tau = 1/L$ across all iterations. In most practical situations, however, the Lipschitz constant L of ∇g is not known or easily computable, and we rely on an adaptive scheme for choosing an appropriate step size at each iteration; backtracking line search is one such scheme, which is straightforward to implement in practice and guarantees convergence of the algorithm under the same Lipschitz assumption on ∇g (but importantly, without having to know its Lipschitz constant L). Given a shrinkage factor $0 < \gamma < 1$, the backtracking line search routine at a given iteration of proximal gradient descent starts with $\tau = \tau_0$ (a large initial guess for the step size), and while

$$g(x - \tau G_\tau(x)) > g(x) - \tau \nabla g(x)^T G_\tau(x) + \frac{\tau}{2} \|G_\tau(x)\|_2^2, \quad (\text{B6})$$

it shrinks the step size by letting $\tau = \gamma \tau$. Once the exit criterion is achieved (ie, the above is no longer satisfied), the proximal gradient descent algorithm then uses the current value of τ to take an update step, as in (B5) (or (A2)).

In the case of the multinomial fused lasso problem, the generalized gradient is of dimension $G_\tau(\beta_0, \beta) \in \mathbb{R}^{(p+1) \times T \times (K-1)}$, where

$$[G_\tau(\beta_0, \beta)]_{0..} = [\nabla g(\beta_0, \beta)]_{0..},$$

and

$$[G_{\tau}(\beta_0, \beta)]_{j \cdot k} = \frac{\beta_{j \cdot k} - \text{prox}_{\text{FLSA}, \tau}(\beta_{j \cdot k} - \tau[\nabla g(\beta_0, \beta)]_{j \cdot k})}{\tau} \quad \text{for } j = 1, \dots, p, k = 1, \dots, K - 1.$$

Here, $\text{prox}_{\text{FLSA}, \tau}(\beta_{j \cdot k} - [\nabla g(\beta_0, \beta)]_{j \cdot k})$ is the proximal map defined by the FLSA evaluated at $\beta_{j \cdot k} - \tau[\nabla g(\beta_0, \beta)]_{j \cdot k}$, ie, the right-hand side in (B4). Backtracking line search now applies just as described above.

B.2.2 | Stopping criteria—The simplest implementation of proximal gradient descent would run the algorithm for a fixed, large number of steps S . A more refined approach would check a stopping criterion at the end of each step and terminate if such a criterion is met. Given a tolerance level $\epsilon > 0$, two common stopping criteria are then based on the relative difference in function values, as in

$$\text{stopping criterion 1: terminate if } C_1 = \frac{|f(\beta_0^+, \beta^+) - f(\beta_0, \beta)|}{f(\beta_0, \beta)} \leq \epsilon,$$

and the relative difference in iterates, as in

$$\text{stopping criterion 2: terminate if } C_2 = \frac{\|(\beta_0^+, \beta^+) - (\beta_0, \beta)\|_2}{\|(\beta_0, \beta)\|_2} \leq \epsilon.$$

The second stopping criterion is generally more stringent and may be hard to meet in large problems, given a small tolerance ϵ .

For the sake of completeness, we outline the full proximal gradient descent procedure in the notation of the multinomial fused lasso problem, with backtracking line search and the first stopping criterion, in Algorithms 1 and 2.

Algorithm 1 Proximal gradient descent for the multinomial fused lasso

INPUT: Predictors X , outcomes Y , tuning parameter values λ_1, λ_2 , initial coefficient guesses $(\beta_0^{(0)}, \beta^{(0)})$, maximum number of iterations S , initial step size before backtracking τ_0 , backtracking shrinkage parameter γ , tolerance ε

OUTPUT: Approximate solution $(\hat{\beta}_0, \hat{\beta})$

```

1:  $s = 1, C = \infty$ 
2: while ( $s \leq S$  and  $C > \varepsilon$ ) do
3:   Find  $\tau_s$  using backtracking, Algorithm 2 (INPUT:  $\beta_0^{(s-1)}, \beta^{(s-1)}, \tau_0, \gamma$ )
4:   Update the intercept:  $\beta_0^{(s)} = \beta_0^{(s-1)} - \tau_s [\nabla g(\beta_0^{(s-1)}, \beta^{(s-1)})]_{0..}$ 
5:   for  $j = 1, \dots, p$  do
6:     for  $k = 1, \dots, (K-1)$  do
7:       Update  $\beta_{j..k}^{(s)} = \text{prox}_{\text{FLSA}, \tau_s}(\beta_{j..k}^{(s-1)} - \tau_s [\nabla g(\beta_0^{(s-1)}, \beta^{(s-1)})]_{j..k})$ 
8:     end for
9:   end for
10:  Increment  $s = s + 1$ 
11:  Compute  $C = [f(\beta_0^{(s)}, \beta^{(s)}) - f(\beta_0^{(s-1)}, \beta^{(s-1)})] / f(\beta_0^{(s-1)}, \beta^{(s-1)})$ 
12: end while
13:  $\hat{\beta}_0 = \beta_0^{(s)}, \hat{\beta} = \beta^{(s)}$ 
14: return  $(\beta_0, \hat{\beta})$ 

```

Algorithm 2 Backtracking line search for the multinomial fused lasso

INPUT: $\beta_0, \beta, \tau_0, \gamma$

OUTPUT: τ

```

1:  $\tau = \tau_0$ 
2: while (true) do
3:   Compute  $[G_\tau(\beta_0, \beta)]_{0..} = [\nabla g(\beta_0, \beta)]_{0..}$ 
4:   for  $j = 1, \dots, p$  do
5:     for  $k = 1, \dots, (K-1)$  do
6:       Compute  $[G_\tau(\beta_0, \beta)]_{j..k} = [\beta_{j..k} - \text{prox}_{\text{FLSA}, \tau}(\beta_{j..k} - \tau [\nabla g(\beta_0, \beta)]_{j..k})] / \tau$ 
7:     end for
8:   end for
9:   if  $g((\beta_0, \beta) - \tau G_\tau(\beta_0, \beta)) > g(\beta_0, \beta) - \tau [\nabla g(\beta_0, \beta)]^T G_\tau(\beta_0, \beta) + \frac{\tau}{2} \|G_\tau(\beta_0, \beta)\|_2^2$  then
10:    Break
11:   else
12:    Shrink  $\tau = \gamma \tau$ 
13:   end if
14: end while
15: return  $\tau$ 

```

B.2.3 | Missing individuals—Often, in practice, some individuals are not present at some time points in the longitudinal study, meaning that one or both of their outcome values

and predictor measurements are missing over a subset of $t = 1, \dots, T$. Let I_t denote the set of completely observed individuals (ie, with both predictor measurements and outcomes observed) at time t , and let $n_t = |I_t|$. The simplest strategy to accommodate such missingness would be to compute the loss function g only observed individuals, so that

$$g(\beta_0, \beta) = \sum_{t=1}^T \sum_{i \in I_t} \left\{ \sum_{k=1}^{K-1} -\mathbb{I}(Y_{it} = k)(\beta_{0tk} + X_{i \cdot t} \beta_{\cdot tk}) + \log \left(1 + \sum_{h=1}^{K-1} \exp(\beta_{0th} + X_{i \cdot t} \beta_{\cdot th}) \right) \right\}.$$

An issue arises when the effective sample size n_t is quite variable across time points t in this case, the penalty terms can have quite different effects on the coefficients $\beta_{\cdot t}$ at one time t versus another. That is, the coefficients $\beta_{\cdot t}$ at a time t in which n_t is small experience a relatively small loss term

$$\sum_{i \in I_t} \left\{ \sum_{k=1}^{K-1} -\mathbb{I}(Y_{it} = k)(\beta_{0tk} + X_{i \cdot t} \beta_{\cdot tk}) + \log \left(1 + \sum_{h=1}^{K-1} \exp(\beta_{0th} + X_{i \cdot t} \beta_{\cdot th}) \right) \right\}, \quad (B7)$$

simply because there are fewer terms in the above sum compared to a time with a larger effective sample size; however, the penalty term

$$\lambda_1 \sum_{j=1}^p \sum_{k=1}^{K-1} |\beta_{jtk}| + \lambda_2 \sum_{j=1}^p \sum_{k=1}^{K-1} |\beta_{jtk} - \beta_{j(t+1)k}|$$

remains comparable across all time points, regardless of sample size. A fix would be to scale the loss term in (B7) by n_t to make it (roughly) independent of the effective sample size, so that the total loss becomes

$$g(\beta_0, \beta) = \sum_{t=1}^T \frac{1}{n_t} \sum_{i \in I_t} \left\{ \sum_{k=1}^{K-1} -\mathbb{I}(Y_{it} = k)(\beta_{0tk} + X_{i \cdot t} \beta_{\cdot tk}) + \log \left(1 + \sum_{h=1}^{K-1} \exp(\beta_{0th} + X_{i \cdot t} \beta_{\cdot th}) \right) \right\}. \quad (B8)$$

This modification indeed ends up being important for the Alzheimer's analysis that we present in Section 3, since this study has a number of individuals in the tens at some time points and in the hundreds for others. The proximal gradient descent algorithm described in this section extends to cover the loss in (B8) with only trivial modifications.

REFERENCES

1. Evans DA, Funkenstein HH, Albert MS, et al. Prevalence of Alzheimer's disease in a community population of older persons. JAMA J Am Med Assoc 1989;262(18):2551–2556.
2. Fitzpatrick AL, Kuller LH, Ives DG, et al. Incidence and prevalence of dementia in the cardiovascular health study. J Am Geriatrics Soc 2004;52(2):195–204.

3. Corrada M, Brookmeyer R, Paganini-Hill A, Berlau D, Kawas CH. Dementia incidence continues to increase with age in the oldest old: the 90+ study. *Ann Neurol* 2010;67(1):114–121. [PubMed: 20186856]
4. Tang MX, Maestre G, Tsai WY, et al. Relative risk of Alzheimer disease and age-at-onset distributions, based on APOE genotypes among elderly African Americans, Caucasians, and Hispanics in New York City. *Am J Human Genetics* 1996;58(3):574–584. [PubMed: 8644717]
5. Launer LJ, Andersen K, Dewey ME, et al. Rates and risk factors for dementia and Alzheimer's disease results from EURODEM pooled analyses. *Neurology* 1999;52(1):78–84. [PubMed: 9921852]
6. Kuller LH, Lopez OL, Newman A, et al. Risk factors for dementia in the cardiovascular health cognition study. *Neuroepidemiology* 2003;22(1):13–22. [PubMed: 12566949]
7. Irie F, Fitzpatrick AL, Lopez OL, et al. Type 2 diabetes (T2D), genetic susceptibility and the incidence of dementia in the cardiovascular health study. *Neurology* 2005;64(6):A316–A316.
8. Skoog I, Nilsson L, Persson G, et al. 15-year longitudinal study of blood pressure and dementia. *The Lancet* 1996;347(9009):1141–1145.
9. Verghese J, Lipton RB, Katz MJ, et al. Leisure activities and the risk of dementia in the elderly. *New England J Med* 2003;348(25):2508–2516. [PubMed: 12815136]
10. Erickson KI, Raji CA, Lopez OL, et al. Physical activity predicts gray matter volume in late adulthood the cardiovascular health study. *Neurology* 2010;75(16):1415–1422. [PubMed: 20944075]
11. Scarmeas N, Stern Y, Tang M-X, Mayeux R, Luchsinger JA. Mediterranean diet and risk for Alzheimer's disease. *Ann Neurol* 2006;59(6):912–921. [PubMed: 16622828]
12. Lopez OL, Jagust WJ, DeKosky ST, et al. Prevalence and classification of mild cognitive impairment in the cardiovascular health study cognition study: part 1. *Arch Neurol* 2003;60(10):1385–1389. [PubMed: 14568808]
13. Saxton J, Lopez OL, Ratcliff G, et al. Preclinical Alzheimer disease neuropsychological test performance 1.5 to 8 years prior to onset. *Neurology* 2004;63(12):2341–2347. [PubMed: 15623697]
14. Lopez OL, Kuller LH, Becker JT, et al. Incidence of dementia in mild cognitive impairment in the cardiovascular health study cognition study. *Arch Neurol* 2007;64(3):416–420. [PubMed: 17353386]
15. Sweet RA, Seltman H, Emanuel JE, et al. Effect of Alzheimer's disease risk genes on trajectories of cognitive function in the cardiovascular health study. *Am J Psychiatr* 2012;169(9):954–962. [PubMed: 22952074]
16. Lecci F An analysis of development of dementia through the extended trajectory grade of membership model. In: *Handbook of Mixed Membership Models and Their Applications* Boca Raton, FL: Chapman & Hall; 2014.
17. Rosvall L, Rizzuto D, Wang H-X, Winblad B, Graff C, Fratiglioni L. APOE-related mortality: effect of dementia, cardiovascular disease and gender. *Neurobiol Aging* 2009;30(10):1545–1551. [PubMed: 18237822]
18. Zhang D, Shen D, Initiative Alzheimer's Disease Neuroimaging. Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. *PloS One* 2012;7(3):e33182. [PubMed: 22457741]
19. Zhou J, Liu J, Narayan VA, Ye J, Initiative Alzheimer's Disease Neuroimaging. Modeling disease progression via multi-task learning. *NeuroImage* 2013;78:233–248. [PubMed: 23583359]
20. Weiner MW, Veitch DP, Aisen PS, et al. The Alzheimer's disease neuroimaging initiative: a review of papers published since its inception. *Alzheimer's Dement* 2013;9(5):e111–e194. [PubMed: 23932184]
21. Huang L, Jin Y, Gao Y, Thung K-H, Shen D, Alzheimer's Disease Neuroimaging Initiative. Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest. *Neurobiol Aging* 2016;46:180–191. <http://www.sciencedirect.com/science/article/pii/S0197458016301282> [PubMed: 27500865]

22. Bo Xin, Kawahara Y, Wang Y, Gao W. Efficient generalized fused lasso and its application to the diagnosis of Alzheimer's Disease. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence; 2014; Quebec City, Canada.
23. Tibshirani R Regression shrinkage and selection via the lasso. *J Royal Stat Soc Ser B Methodol* 1996;58(1):267–288.
24. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. *J Royal Stat Soc Ser B Stat Methodol* 2005;67(1):91–108.
25. Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternative direction method of multipliers. *Found Trends Mach Learn* 2011;3(1):1–122.
26. Tibshirani RJ, Taylor J. The solution path of the generalized lasso. *Ann Stat* 2011;39(3):1335–1371.
27. Tibshirani RJ, Taylor J. Degrees of freedom in lasso problems. *Ann Stat* 2012;40(2):1198–1232.
28. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning; Data Mining, Inference and Prediction* 2nd ed. New York, NY: Springer; 2008.
29. Lange T, Roth V, Braun M, Buhmann J. Stability-based validation of clustering solutions. *Neural Computation* 2004;16:1299–1323. [PubMed: 15130251]
30. Meinshausen N, Bühlmann P. Stability selection. *J Royal Stat Soc Ser B Stat Methodol* 2010;72(4):417–473.
31. Liu H, Roeder K, Wasserman L. Stability approach to regularization selection (StARS) for high-dimensional graphical models. Paper presented at: Neural Information Processing Systems 23 (NIPS 2010); 2010.
32. Breiman L, Friedman J, Stone C, Olshen R. *Classification and Regression Trees* Boca Raton, FL: Chapman & Hall/CRC Press; 1984.
33. Cox D Regression models and life-tables. *J Royal Stat Soc Ser B Methodol* 1972;34(2):187–220.
34. Efron B The efficiency of Cox's likelihood function for censored data. *J Am Stat Assoc* 1977;72(359):557–565.
35. Kalbflesich J, Prentice R. *The Statistical Analysis of Failure Time Data* Hoboken, NJ: John Wiley & Sons, Inc; 2002 Wiley Series in Probability and Statistics.
36. Tibshirani R The lasso method for variable selection in the Cox model. *Statist Med* 1997;16(4):385–395.
37. Hernán MA. The hazards of hazard ratios. *Epidemiol Camb Mass* 2010;21(1):13.
38. Lopez OL, Becker JT, Kuller LH. Patterns of compensation and vulnerability in normal subjects at risk of Alzheimer's disease. *J Alzheimer's Dis* 2013;33:S427–S438. [PubMed: 22669014]
39. Kuller L, Chang Y, Becker J, Lopez O. Does Alzheimer's disease over 80 years old have a different etiology? *Alzheimer's Dement* 2011;7(4):S596–S597.
40. Kivipelto M, Helkala E-L, Laakso MP, et al. Midlife vascular risk factors and Alzheimer's disease in later life: longitudinal, population based study. *BMJ* 2001;322(7300):1447–1451. [PubMed: 11408299]
41. Kim S-J, Koh K, Boyd S, Gorinevsky D. ℓ_1 trend filtering. *SIAM Review* 2009;51(2):339–360.
42. Tibshirani RJ. Adaptive piecewise polynomial estimation via trend filtering. *Annal Statist* 2014;42(1):285–323.
43. Zhang C-H, Zhang S. Confidence intervals for low-dimensional parameters with high-dimensional data 2011 arXiv: 1110.2563.
44. Javanmard A, Montanari A. Confidence intervals and hypothesis testing for high-dimensional regression 2013 arXiv: 1306.3171.
45. van de Geer S, Bühlmann P, Ritov Y. On asymptotically optimal confidence regions and tests for high-dimensional models 2013 arXiv: 1303.0518.
46. Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R. A significance test for the lasso. *Annal Statist* 2014;42(2):413–468.
47. Lee J, Sun D, Sun Y, Taylor J. Exact post-selection inference with the lasso 2013 arXiv: 1311.6238.
48. Taylor J, Lockhart R, Tibshirani RJ, Tibshirani R. Exact post-selection inference for forward stepwise and least angle regression 2014 arXiv: 1401.3889.

49. Parikh N, Boyd S. Proximal algorithms. Found Trends Optim 2013;1(3):123–231.
50. Nesterov Y Gradient Methods for Minimizing Composite Objective Function CORE Discussion Paper 2007.
51. Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J Imaging Sci 2009;2(1):183–202.
52. Johnson NA. A dynamic programming algorithm for the fused lasso and L_0 -segmentation. J Comput Graph Stat 2013;22(2):246–260.

FIGURE 1.

A simple example with $n = 50$, $T = 15$, $K = 2$, and $p = 30$. The left panel displays the true coefficient trajectories across time points $t = 1, \dots, 15$ (only 3 of the 30 are nonzero); the middle panel shows the (unregularized) maximum likelihood estimates; the right panel shows the regularized estimates from (1), with $\lambda_1 = 2.5$ and $\lambda_2 = 12.5$ [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 2.

Proportion of observed outcome in each category over age, starting at 75. The total number of observations at each age is shown in the top axis of the plot. While the proportion of dead patients increases over time, that of normal patients decreases strictly. Interestingly, the prevalence of proportion of patients with dementia is increasing until around age 90 and then starts decreasing [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 3.

Comparison of different methods for selection of tuning parameters λ_1 , λ_2 on the Cardiovascular Health Study Cognition Study data set. The x -axis in each plot parameterizes the six different methods considered, which are from left to right: AIC and BIC under negative log likelihood loss (AIC_{LD} and BIC_{LD}), AIC and BIC under misclassification loss (AIC_{MC} and BIC_{MC}), cross-validation (CV), and cross-validation with the one-standard-error rule ($CV1se$). The upper left plot shows (out-of-sample) misclassification rate associated with the estimates selected by each method, averaged over 5 iterations. The segments denote ± 1 standard errors around the mean. The red dotted line is the average misclassification rate associated with the naive estimator that predicts all individuals as dead (the majority class). The upper right and bottom left plots show different measures of evaluation (again, computed out-of-sample): the true positive rate in identifying the dementia class, respectively, the true positive rate in identifying the dementia and death classes combined. Finally, the bottom right plot shows degrees of freedom (number of nonzero blocks) of the estimates selected by each method [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 4.

Mean true positive rate (TPR) and mean positive predictive value (PPV) for correctly predicting dementia over age, for each of the methods considered for selecting the tuning parameters λ_1 and λ_2 . Cross-validation under the usual rule gives the best out of sample prediction in terms of maximizing the TPR and PPV for predicting dementia, with best prediction in terms of these evaluation metrics for predicting dementia between ages 80 and 90 using predictors observed between 70 and 80 years of age [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 5.

Cardiovascular Health Study Cognition Study data analysis. Left: relative importance plots for the 15 most important variables in the “dementia vs normal” and “death vs normal” equations of the multinomial logit model. X -axis represents measure of importance on a scale of 0–100. Right: corresponding estimated coefficients for predictors recorded between ages 65 and 100. The order of the legends follows the order of the maximum/minimum values of the estimated coefficient trajectories. Note that some coefficients are estimated to be very close to 0 and the corresponding trajectories are hidden by other coefficients [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 6. Cardiovascular Health Study Cognition Study data analysis. Estimated intercept coefficients in the two separate equations of the multinomial logit model

FIGURE 7.

Simulation: comparison of different prediction models on simulated data set. The x-axis in each plot represents the five different models considered, which are, from left to right: multinomial logistic regression without penalty (**GLM**), multinomial regression with only lasso penalty (**multi-lasso**), multinomial logistic regression with lasso and fused lasso penalties (**multi-fused**), binomial logistic regression with lasso and fused lasso penalties in which people who died are ignored from training (**bi-fused I**), and, finally, the binomial logistic regression with lasso and fused lasso penalties in which normal and dead patients are grouped into one category (**bi-fused II**). The upper left plot shows (out-of-sample) misclassification rate associated with the estimates selected by each method, averaged over 5 iterations. The segments denote ± 1 standard errors around the mean. The upper right and bottom left plots show different measures of evaluation (again, computed out-of-sample): the true positive rate (TPR) in identifying the dementia class, and the false positive rate (FPR) in identifying the dementia. Finally, the bottom right plot shows degrees of freedom (number of nonzero blocks) of the estimates selected by each method and standardized by the number of outcome categories in the fitted model, for comparability

TABLE 1

Notation used throughout this paper

Parameter	Meaning
$i = 1, \dots, n$	Index for individuals
$j = 1, \dots, p$	Index for predictors
$t = 1, \dots, T$	Index for time points
$k = 1, \dots, K$	Index for outcomes
Y	$n \times T$ matrix of (future) outcomes
X	$n \times p \times T$ array of predictors
ρ_0	$T \times (K - 1)$ matrix of intercepts
β	$p \times T \times (K - 1)$ array of coefficients

TABLE 2

The 15 most important variables in the two separate equations of the multinomial logit model

Dementia vs Normal	
Variable	Meaning (and coding for categorical variables, before scaling)
Race white	Race: "White" 1, else 0
Vitamin C	Taken vitamin C in the last 2 weeks? (number of days)
Learn new	How is the person at learning new things wrt 10 years ago? "A bit worse" 1, else 0
Estrogen	If you not currently taking estrogen, have you taken in the past? "Yes" 1, "No" 0
Fearful	How often felt fearful during last week? "Most of the time" 1, else 0
Wake early	Do you usually wake up far too early? "Yes" 1, "No" 0
Female	Gender: "Female" 1, "Male" 0
Diuretics	Medication: thiazide diuretics w/o K-sparing. "Yes" 1, "No" 0
Race other	Race: "Other (no white, no black)" 1, else 0
Orthosis	Do you use a lower extremity orthosis? "Yes" 1, "No" 0
Pulse	60 second heart rate
Gripping	What causes difficulty in gripping? "Pain in arm/hand" 1, else 0
Find help	If sick, could easily find someone to help? "Probably False" 1, else 0
Digits correct	Digit-symbol substitution task: number of symbols correctly coded
Trust	There is at least one person whose advice you really trust. "Probably true" 1, else 0
Death vs Normal	
Variable	Meaning (and coding for categorical variables, before scaling)
Digits correct	Digit-symbol substitution task: number of symbols correctly coded
Chair stand	Repeated chair stands: number of seconds
Female	Gender: "Female" 1, "Male" 0
Cardiac injury	Cardiac injury score
Chest pain	Ever had pain in chest when walking uphill/hurry? "No" 1, else 0
Cigarettes	Number of cigarettes smoked per day
Digitalis	Digitalis medicines prescribed? "Yes" 1, "No" 0
Smoker	Current smoke status: "Never smoked" 1, else 0
Health	Would you say, in general, your health is ... ? "Fair" 1, else 0
Exercise	If gained/lost weight, was exercise a major factor? "Yes" 1, "No" 0
#Medications	Number of medications taken
Diabetes	ADA diabetic status? "New diabetes" 1, else 0
Any smoker	Does anyone living with you smoke cigarettes regularly? "Yes" 1, "No" 0
Bld pressure	Blood pressure variable: left ankle-arm index
Walking	Do you have difficulty walking one-half a mile? "Yes" 1, else 0

TABLE 3

Proportion of the true top 20 important predictors that were selected as top 20 important predictors in the fitted model, for dementia versus normal and dead versus normal, for each of the five prediction models

	Method	% Dementia	% Death
Multinomial logistic regression	No regularization	65	50
	Lasso	70	75
	Lasso and fused lasso	75	75
Binomial logistic regression			
(Dementia vs normal)	Lasso and fused lasso	80	0
(Dementia vs normal or death)	Lasso and fused lasso	65	0